# *On using iterative mapping algorithms to sequence the chloroplast genome of Semele androgyna: an overview of methods and results*


**Universität für Bodenkultur Wien**
University of Natural Resources
and Applied Life Sciences, Vienna

Radu Chirovici

Universität für Bodenkultur
Department für Integrative Biologie und Biodiversitätsforschung
Institut für Integrative Naturschutzforschung

A thesis submitted for the degree of
**Master of Science**
**April 2018**

Supervisors : Univ.-Prof. Dr. Harald Meimberg, Dr. Manuel Curto

# Table of Contents

I would like to thank :

**Manuel** and **Harry** for the amazing support, guidance and often patience they provided. My limited knowledge in the field of evolutionary biology was quickly improved by their tremendous availability over the last 2 years. I will be forever grateful for this.

**Sophie**, for her continuous support, love, help and structure during often stressful times.

**My family**, for helping me out so much.

All the staff of **INF BOKU**, for making me feel at home.

# Abstract

This thesis deals with the modern challenges of genome assembly, generally. In the introductory part I will try and define these issues that these challenges raised, and where to improve upon these issues. As the scale of the project grew, I set a new focus on mitochondrias and subsequently, chloroplasts and the depth of assembled genomes. These organelles also have a very interesting and usable genome, with many applications throughout molecular ecology and biodiversity research, such as DNA barcoding, phylogenetics and population genetics.

I list the methods and software used in the second part. I have tried to focus on readily available, easy to use tools, and to implement them into a work flow that can improve and grow with each assembly or barcoding project. This work flow was engrained in **mappy**, a tool that has been developed out of the need to streamline this entire process and to provide a basis for future improvements in local, small-scale projects.

The mappy outputs have been aligned with those of two established tools of organelle assembly, NOVOplasty and MITObim. They both use different algorithms. MITObim and mappy base their calculations on MIRA, NOVOplasty uses a hash table to improve the assembly results in a timely fashion.

NOVOplasty and MITObim, due to the assembly of the inverted repeat region, do not output a contig with a correct gene order. This problem would be overcome by outputting the different regions of the cp genome (single copy regions, and inverted repeats) separately. For that the junctions among these regions need to be identified. In this thesis I tested if coverage could be used as an indicator. In order to check if depth (or coverage) in plastid genomes follows a pattern conclusive with genome structure, I extracted this data and compared spikes in coverage tables to the existing published genome of *Panax ginseng damaya*. Results show that coverage spikes in assembled bait contigs are consistent with the junctions of the quadripartite structure of chloroplast DNA.

These results are then discussed, and the work flow analyzed. I suggest some improvements, and provide a road map for the future. Using mappy, we will compare this approach of checking depth spikes in contigs and further improve the tool.

# 1. Introduction

Genome sequence assembly has come a long way and, with more and more research in the field, is becoming a mature technology, being able to produce a high number of genomes in the near future (Alkan et al. 2011). In the following chapter I will try and succinctly describe the field of genome assembly. After a brief introduction I will go into more detail and elaborate on the three main types of genomes assembled: nuclear, mitochondrial and chloroplast (or plastid, found only in plants). They are studied for different reasons and also present different analytical challenges.

## 1.1 Genome assembly – overview of sequencing methods and algorithms

The scope and use cases of assemblies increases exponentially with each year passed. A broad variety of sequencing methods are used to analyze RNA transcription and the its structure, to detect both DNA and RNA, to analyze DNA-Protein interactions, just to give a few more common examples. Throughout molecular sequencings recent history, many new technologies have been introduced(and retired) to support research purposes. I find that the main goal of these sequencing machines, expressed by scientists in both journal and layman articles is to produce **many reads (or sequenced fragments) per run at lower costs** (Hodkinson and Grice, 2015). This is, of course, true, if one's research application is to extract complete genomes, as it often is the case in human genetic medicine. Conservation biology examines genetic variation for species identification, evolutionary history of various plants, hybridization, genetic diversity and many more (Puppo et al., 2016, Fuentes-Pardo and Ruzzante, 2017). Even though WGS methods are also used in evolutionary and conservation biology, no single one is a catch-all answer (Fuentes-Pardo, Ruzzante 2017) to the uses described above. I will provide an overview of the most generally used methods then shortly describe the pros and cons of some for then to focus on the most widely used one, **Illuminas dye sequencing method**. The data sampled for this thesis has also been acquired using this method.

Second generation sequencing produces large amounts of reads in a much faster fashion and at a fraction of the cost of **Sanger sequencing** (Schatz et al, 2010), one of the first methods used to extract genetic information out of preserved samples. Frederick Sanger developed the rapid sequencing method in the 1970ies. This technology has been used to sequence the human genome (Lander et al., 1996), among others. The high associated cost and relatively limited throughput of this technology made improvements imperative.

When deciding on a way of acquiring genomic data, it is often an equation of two variables: **costs** and **scope.** Often limited material resources, paired with time constraints, a research group chooses the most cost-effective methode to produce results. The most commonly used next generation sequencing technology, **Illumina**'s sequencing by synthesis, executed by the HiSeq

2000 sequencer, produces up to $10^9$ short reads of up to 100 bp (Alkan et al. 2011, Sims et al. 2014). With other technologies, the read length can be increased to 400 bp (Schatz et al 2010).

The amount of read data constructed this way is very large. The genome of sugar beet (*Beta vulgaris)* is 714-758 megabases long (Dohm et al. 2014), of *Eucalyptus grandis* 640 megabases long (Myburg et. Al 2014) and of the pineapple (*Ananas comosus*) is a bit shorter, at 380 megabases of length (Ming et al, 2015).

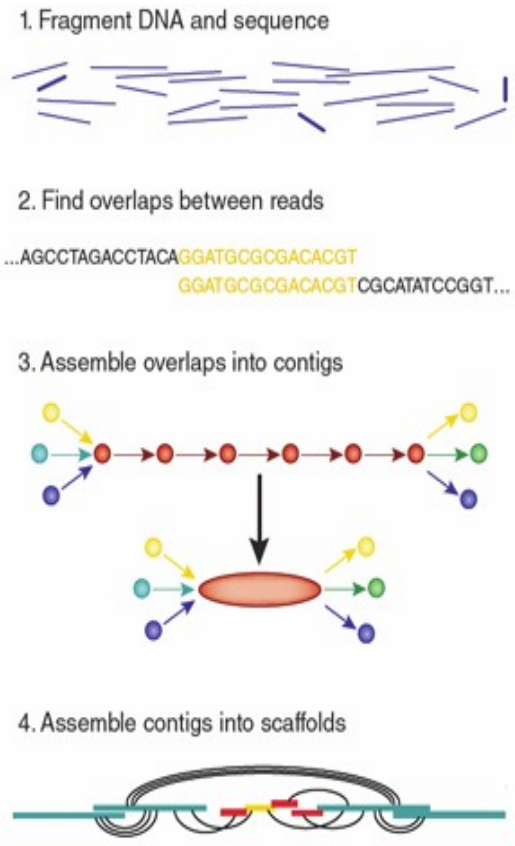



*Figure 1: A general graphical representation of the assembly process according to Michael Schatz, Cold Spring Harbor Laboratory*

The original intent of the research part of the thesis was to select the best samples from INF at that moment, namely *Micromeria varia* samples, and improve their assembly by ways of analyzing the complete nuclear genome using a map-and-extend algorithm and extracting quality data from the inputs. This project was put on hold due to questionable results. Alignments were proven to be inconclusive, the assembled contigs did not match the GenBank genome. Thus, lowering the scale of the project helped.

We found that high coverage, low error genomic scaffolds are needed only in the least amount of projects. Usually, a single region of the genome(nuclear, mitochondrial or plastid) is targeted, on different grounds: to study variation, inheritance, mutations etc. This makes lower coverage data

also usable, being free of the quality constraints of longer sequences, such as complete nuclear genomes.

This information would not be available without a way of data interpretation. This turns the analysis of reads into a computational problem. The reads need to be assembled, or put together, in order to construct the target - both nuclear and/or organelle based – genome. These sequences are joined together into a contig, or a contigous region of the genome (Baker, 2012). These contigs are then combined and added into a 'scaffold'. The assemble scaffolds can then be viewed as a consensus sequence of reads (Miller et al 2010). These gaps represent uncertain matches (Earl et al, 2011), or areas where the assembler just does not know what it assembled. Schatz et al compare this process with a jigsaw puzzle : it might have small pieces but, with enough resources, it can be assembled into a picture. A complete genome thus can also be assembled from short reads. This is the aim of assembly methods : to provide the most complete, error-free genome of a target species (Schatz et al 2011).

Miller et al categorize the assemblers based on the graphing algorithms which are being used :

- overlap/layout/consensus graph,
- de Bruijn graph and
- greedy graph.

They further expand on the algorithms. The overlap graph represents the reads, including their overlaps. In the nodes of the de Bruijn graph there are fixed-length strings (in this case, the reads), and the edges represent the overlaps. Greedy graph algorithms may use either overlap graphs and/or de Bruijn graphs – they add one more read or contig to an existing one until the operation is not possible anymore. In the original design of the thesis, we used assembly programs that used both a de Bruijn graph and a overlap graph approach.

There are a number of challenges in assembling a genome, such as repeat sequences in a certain portion of the genome (Miller et al.2010). These sequences will be discarded by all de novo assembly algorithms, which might lead to reduced or lost complexity(Alkan et. Al 2011).

Coverage, or sequence depth, can be seen as one quality measure of assembled contigs. Sims et al (2014) define theoretical or expected coverage as "the number of times that each nucleotide is expected to be sequenced given a certain number of reads of a given length and the assumption that reads are randomly distributed across an idealized genome". We can define the actual coverage as the number of times a base of a reference is covered by an aligned read during sequencing(Sims et al 2014). Thus, if there is an even, constantly high spread of coverage throughout the assembled sequence we can conclude that the assembled project is of high quality. This is often only in theory, as sequencing errors, read preparation errors and other issues

degrade the value of coverage as a quality control indicator as they lead to gaps in the assembly. Repeat complexity and read length must also be considered when assessing the quality of a project.

### 1.2 Organelle sequencing – mitochondria and chloroplast genome assembly

Cellular organelles are small, intracellular compartments of eukaryotic cells, surrounded by membranes. They are surprisingly important in assuring the correct functioning of organisms (Alberts et al., 2002). They were also the first sequencing markers, as they are highly represented in the cell, and their genomes are single copied(or haploid, in comparison to the nuclear diploid genome).

This thesis objectives lie in the field of chloroplast genome assembly methods.

The chloroplast organelle is proprietary to plants. It is specialized in photosynthesis and has a interesting evolutionary history: all chloroplasts were derived from a single primary endosymbiotic event involving the capture of a cyanobacterium into an ancient eukaryotic cell. (Yagi and Shiina 2014, Turmel et al. 1999). The constant size and staticity (mutations in length of above 800 bp are quite rare) of its circular genome make it an ideal candidate for phylogenetic and evolutionary studies of plants(Palmer et al., 1983). The circular structure includes two inverted repeat regions. This presents an assembly challenge: the genome cannot reliably be assembled using one baiting sequence, as the algorithms used commonly have difficulties identifying the edges of these IR regions. Specifically, during the assembly process, the assembler fails to convolute the edges of the IR with the SSC and LSC parts of the genome, mostly because of direction confusion due to the IR size and nucleotide sequence. Most organelle assembly methods from low coverage data extend existing sequences by recursively mapping the reads to a chloroplast or a known mitochondria sequence (Hahn et al.). Because some of the reads just partially partially to the reference, when creating a consensus sequence out of this assembly they will extend the reference. This new consensus sequence can be used for the next step over and over again until the complete organelle genome is recovered, This process may create problem in the chloroplast while assembling the inverted repeat region. While assembly this region reads from both inverted repeats are mapped at the same time which can lead to the extension of the wrong part of the chloroplast when the IR regions are finished assembly.

In contrast to chloroplasts, mitochondrias are ubiquitous: they can be found in plants, animals and fungi (Castro et al, 2008). As the organelles role is to produce ATP from glucose, it can be described as the ‘powerhouse of the cell’ (McBride et al., 2006). The chromosome is highly variable in size and contain many repeated elements, making the study of mitochondrial DNA

challenging (Galtier, 2011). Yet, the maternal inheritance and its mutation characteristics due to variability, coupled with the high number of mtDNA copies in the cell make the genome a prime marker candidate for taxonomy studies (using mtDNA markers as barcodes for species identification), biodiversity, evolution and phylogeny (Anmarkrud et al, 2017, Gibbs et al, 2007). Even if the DNA is degraded or the sample number is limited, a high number of genetic data can be extracted from the mitochondrial DNA.

Chloroplast and mitochondria genomes are used as the main source of molecular markers for DNA barcoding of animals and plants, respectively (Valentini 2009, Hebert et al. 2003). A good barcoding marker should be variable enough to differentiate between species and at the same time have primer binding sites conserved enough to be used in a wide taxonomic range. This is not easily achieved, and many times more than one marker is need to increase the power of species identification or different primers for different taxonomic groups need to be design (Kress and Erickson 2009). A response to these limitations would be to sequence the complete mitochondria or chloroplast genome.

To improve the quality of the these subsequently assembled mtDNA and/or cpDNA genomes, a large number of runs, or sequencing projects, have to be executed in order to acquire one with satisfactory quality. As explained in Chapter 1.1, low coverage assembled contigs are often the only available output datasets of nuclear genome assembly projects. That is why I will try and detail on why lower coverage per assembled contig/scaffold should not always be an impediment to genomics projects.

Lower depth in resulting contigs can be a result of various factors, including the algorithm used and number or quality of the reads used. One of the main hypotheses that will be tested in this thesis will be the presumption that, if the coverage resulting from the assembly process is ‘spiking’ across parts of the chloroplast DNA, that those spikes coincide with the junction points of the general cpDNA genome. This makes chloroplast genome assembly an even more relevant research focus.

### 1.3 Thesis objectives

This thesis has the objective of establishing bioinformatic resources that can be used to recover whole chloroplast genomes for DNA barcoding using low coverage shot gun sequencing data. This will be done by testing the existing assembly approaches and developing a new algorithm that can overcome some of the challenges specific to chloroplast assembly. We expect that during the assembly of the IR regions the coverage spikes. This characteristics could be used to define junctions of the different chloroplast regions (SSC, LSC, and IR) by stopping the assembly every time that there is a significant change in coverage. The whole chloroplast would be then recovered

by using multiple initial references from the different regions of the chloroplast genome and combining the obtained contigs in the end.

The plant that has been chosen for this study is *Semele androgyna* (in German: Klettermäusedorn) a member of the *Asparagaceae* family. The genus is palaeoendemic to the Atlantic island chains of Madeira and the Canary Islands (Carvalho et al, 2004). Semele androgyna is a species endemic to Madeira, Deserta Grande and Porto Santo, and is specialized (as a climbing shrub) on thriving in humid, low lying forests known as laurel forest, growing all throughout the islands (Capelo, 2005). In some of these islands, the habitats inhabited by S. androgyna are highly fragmented due to human activities. This makes this plant a good system to study the impacts of habitat fragmentation on laurel forest plants.

The main points of the thesis evolved based on availability and quality of *S. androgyna* samples. We discarded some reads to attain an large number of satisfactory contigs – or better yet, an assembly – and to test the main hypothesis of this work : depth in cpDNA assemblies spikes when IRs are starting to be assembled. To conclude this chapter and to refine the hypothesis, I want to define four mainspecific objectives, which I will try to attain.

→ **Developing a lab-internal workflow to acquire data from low coverage reads,**
→ **To provide an overview of commonly used software and cpDNA extraction methods to use in the work flow and investigate possible limitations.**
→ **Inspecting if the coverage spikes in assembled contigs correspond to the junctions of the IR,**
→ **Creat a script to improve the iterative mapping method of chloroplast assembly**

# 2. Methods

This chapter will provide an overview of existing methods and software used in the assembly of organelle DNA as well as how I tested their accuracy in recovering a complete chloroplast genome. Moreover, this section provides a detailed description of the new assembly algorithm developed in the scope of this thesis. This section is structured in the following way: first, a quick introduction into data collection and processing is given. Second, algorithm description of the existing assembly programs and strategy for used for quality assessment is detailed. Third, I provide a complete description of the of the new work flow and methodology used to test if the concept behind it works.

## 2.1 Data collection and read quality processing

The DNA samples were prepared in the lab at the INF (following the method described by Curto et al., 2018). The 500µl lysis buffer (2% SDS, 2%PVP 40, 250 mM NaCl, 200 M Tris HCl, 5mM EDTA, pH8) and 16,67 µl of proteinase K (10mg/mL) was incubated for 2.5 hours at 56°C. Then it was taken out with cleaned tweezers and put in NucleoSpin Filters and centrifuged for 1 min at 2300 rpm. For DNA binding, the 400µl of the supernatant were mixed with 15µl of Mag-Si:DNA beads (size 300nm, MagSi-DNA beads from MagnaMedics) and 600µl binding buffer (COMPOSITION) and incubated at room temperature for 5 min. To separate the supernatant from the beads samples were laid on a magnet separator SL-MagSep96 (Steinbrenner, Germany) for 1 minute. Beads were washed two times by mixing 600µl of 80% ethanol. To exclude excess of ethanol beads were air-dried at room temperature for 10 minutes. Two elutions were done with 20µl and 25µl by mixing preheated (65°C) elution buffer (10 nM Tris with a pH of 8) and letting samples incubate for 5 min at room temperature. DNA was sent for library preparation and sequencing in an Illumina MiSeq at the Genomics Service Unit from the Ludwig Maximilian University of Munich.

Libraries were then prepared with an insert length between 400 and 500bp. Sequencing was done in a pairwise manner with a read length of 300 bp.Quality control has been performed using FastQC (Andrews, 2010). Low quality regions were removed using Cutadapt (Martin, 2011). The paired end reads were then subsequently combined using PEAR (Zhang et al, 2014).

## 2.2 Overview of existing scripts and algorithms

The first objective of the thesis deals with testing the existing approaches used for small-scale organelle assembly projects, where sometimes data is of lower quality and assembly coverage is lower than expected. For testing purposes, assembly has been conducted and result-compared in **MITObim, mappy(both MIRA-based) and NOVOplasty**.

**Mitobim**(Hahn et al., 2013) has been developed to assemble mitochondrial DNA genomes using baits and a reference mtDNA genome of a (closely or even more distant) related individual.

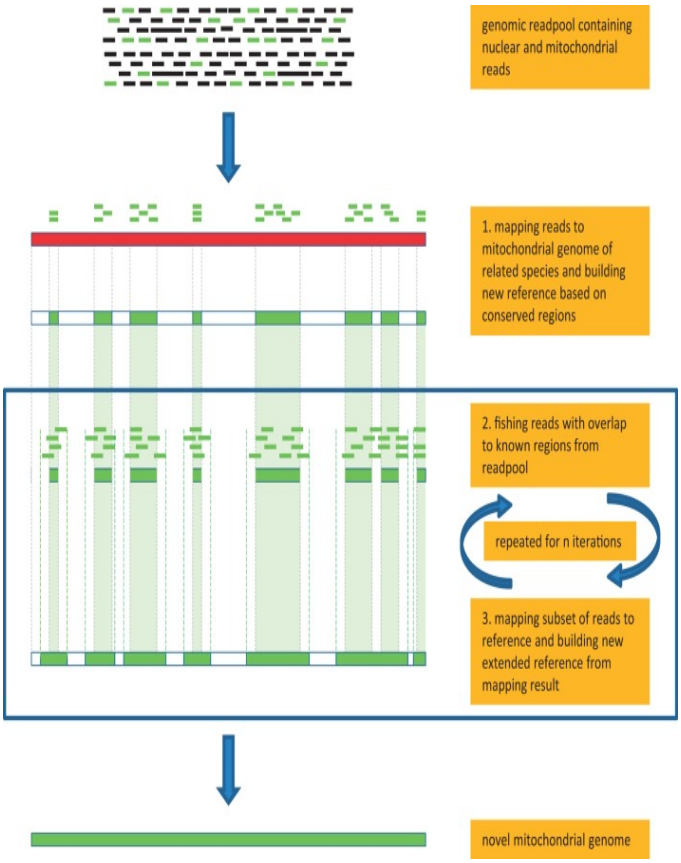



*Figure 2: The graphical overview of the mitobim assembly process. Hahn et al., 2013*

The script has been successfully used to assemble and quality compare a large number of mtDNA genomes. It is essentially a wrapper script based on the MIRA assembler by Bastien Chevreux (Chevreux et al., 1999), but adds functionality and straightforwardness, to improve mtDNA detection(as it is highly similar to nuclear DNA. The reference genome which provides the basis of the organelle DNA assembly can be so far upstream as in the same family (due to high conservation rate of cpDNA), which provides a breath of possibilities for assembly. For smaller projects this can be crucial, as in some cases, research efforts are concentrated unto lesser known plants from a genus or family.

MITObim is a reference based can be used using assembler that uses either a complete or a portion of the organelle genome from other species or genus. In case the complete genome is used, MITObim first maps the reads using MIRA to the reference and creates a consensus sequence based on the mapped reads (Figure 2). Then, it uses the mirabait(which is included in MIRA) to fish the reads with some overlap with the reference (bait). These reads are then mapped

and used to extend the reference. Mapped reads are excluded from the read pool. This process is iterated until all reads are mapped. When using a small portion of the mitochondrial genome as reference (ex: an amplified gene), MITObim goes directly to the second step (fishing reads with MIRA baits) and runs until all reads are mapped.

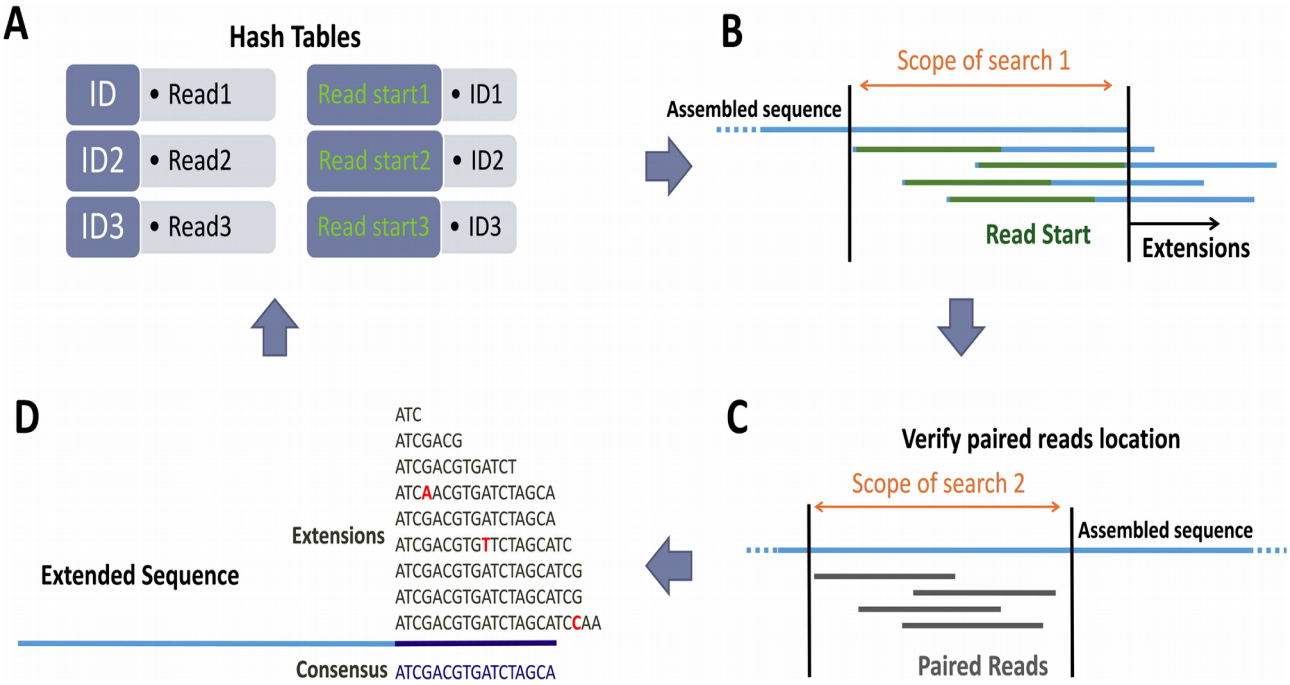



*Figure 3: The graphical overview of the NOVOplasty assembly process. Dierckxsens et al, 2016*

**NOVOplasty**(Dierckxsens et al, 2016) uses a similar approach, but with some defining differences. Sequences are stored into a hash table (Figure 3). A seed is then used to start the assembly, which extends the sequence bidirectionally. The end and start of the seed are then scanned for overlapping reads in the hash table. Related, similar reads are then grouped and then extended. It is to be noted that the script does not extend every read, it extends the seed until the genome is formed, which, in this case, is the formation of a circular molecule. The circularization is detected when the both ends of the seed overlap by at least 200bp. Dierckxsens also states that the assembler detects common errors, such as a high error rate after SNRs when using HiSeq and MiSeq SGS technologies. These problematic regions, when identified, do not cause gaps in the contig, the assembly point simply omits them.

### 2.3 Testing existing aproaches

To test both approaches (MITObim and NOVOplasty) we ran them with the merged Semele reads for MITObim and quality controlled paired reads for NOVOplasty. Both programs were run using five amplicon sequences taken from Genbank as baits (see section 2.4). MITObim was run using a

Kmer size of 20 for baiting and a minimum identity of 85. For NOVOplasty, k-mer sizes ranging from 20 to 40 were tested and the one resulting in larger contig size was considered to be the most suitable for analysis. Values such as insert size and read length were chosen so that they closely resemble our read pool, such at it was described in chapter 1 (insert library size, average read length).

In order to better understand and test applicability of the genomes generated by the scripts, a number of statistics and figures have been used to check if the results are comparable to existing published results of other studies. More specifically, quality of the assemblies was evaluated by testing homology of the obtained contigs with the chloroplast genomes of *Panax ginseng* (same family). The *P. ginseng damaya* cp genome was divided into the different chloroplast regions (SSC, LSC, IR) prior to blasting. This way, it was possible to check which parts of the contigs belonged to the different chloroplast regions and if they were assembled in the correct order. Additionally, gene order in the obtained contigs was evaluated by annotating them using the program GeSeq with the standard parameters(also adding ARAGORN from the same suite for tRNA discovery).

### 2.4 Description of used baits

The five baits used corresponded to some of the nucleotide sequences available for *S. androgyna* chloroplast in genebank. These corresponded to portions of the following regions (Figure 4):

The **ndhF** gene (https://www.ncbi.nlm.nih.gov/nuccore/499068597) is located in the SSC region of the cp genome. It is commonly located next to the starting point of a IR and is highly conserved(Neyland et al.,1996). It is making it useful for a wide range of phylogenetic analyses (Patterson et al., 2014, Dong et al., 2012, Kress et al., 2005)

**matK** (https://www.ncbi.nlm.nih.gov/nuccore/313756533) is also very conserved (Selvaraj et al., 2008) and also chosen due to its proximity to another start of a inverted repeat. Additionally, it is used as a barcoding marker.

**AtpB** (https://www.ncbi.nlm.nih.gov/nuccore/499069331) is short for ATPase beta subunit (Poessner et al., 1986). We selected it due to its location in the middle of LSC.

The **rbcl** gene (https://www.ncbi.nlm.nih.gov/nuccore/HM640442.1 is used as a phylogenetic marker at higher taxonomic levels, as it does not contain enough information to analyze relations between closely related individuals. Nevertheless is commonly used as a barcoding marker. Just as atpB, it has also been chosen due to its location in the cpDNA.

A portion of the **internal spacer (https://www.ncbi.nlm.nih.gov/nuccore/1204021**) has also been selected. At lower taxonomic levels it is also used for a variety of comparisons, barcoding and phylogenetic studies. (Degtjareva et al., 2012)

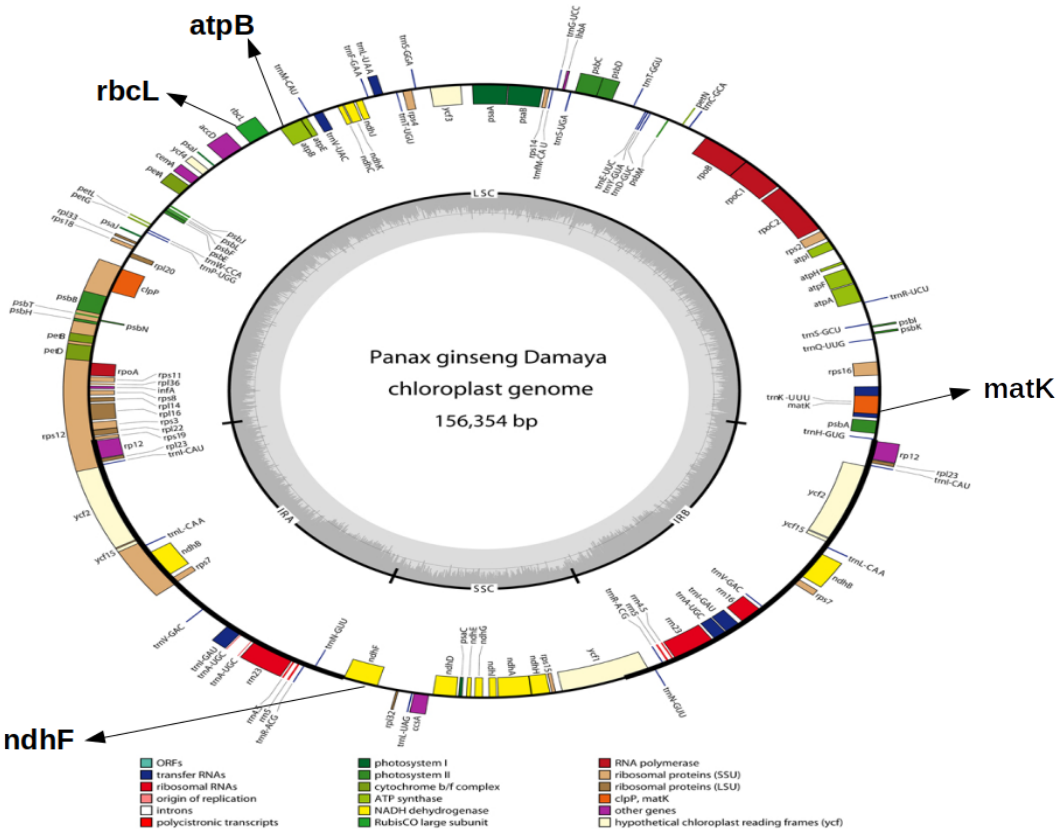



*Figure 4: annotated genome of P. ginseng Damaya with selected bait locations(Zhao et al, 2015)*

The figure above shows the P. ginseng Damaya annotated cpDNA with the chosen baits and their locations in the genome.

### 2.5 New approach for chloroplast assembly

The script developed in this thesis (mappy) corresponds to an approach that potentially can overcome the limitations of MITObim and NOVOplasty, which as it will be shown in the results section are not able to assemble the different chloroplast region in the corrector order. The algorithm is similar to MITObim with the innovation of detecting the junctions of the inverted repeat by looking into variations of coverage during the assembly process. consists in the following steps:

1. Mapping of merged reads to the bait and create a consensus sequence with MIRA

2. Reading of the average coverage of the mapped reads and the coverage a the coverage per first and last 30 base pairs of a contig. Average coverage is outputted by MIRA while for the coverage at the contigs ends further processing was necessary:

2.1 The MIRA alignment output in the maf format was transformed into the .sam format using miraconvert.

2.4 The .sam file is converted into the .bam format with samtools (Heng et al., 2009)

2.2 The .bam file is then transformed into a bed format using bedtools(Quinlan et al., 2010)

2.3 The coverage per position was obtained using bedtools.

2.4 The average coverage in the first and last 30 bp of each contig was calculated using the Numpy python library.

3. The consensus sequence is saved in a specific folder. The coverage results are outputted in a text tab-delimited file.

4. The steps 1 to 3 are repeated for 120 iterations using the contig from step 3 as the new bait.

In the scope of this thesis it was not possible to produce the final version of this script. Instead it just focuses on the proof of concept. In its final version, the script will be able to detect if there is a change in coverage and it will stop the iteration process. The resulting contig should correspond to one of the single copy regions or inverted repeat depending on the position of the initial bait used. This approach was implemented in python and the code can be found in the supplementary material portion of the thesis.

The same blast comparisons used to evaluate the MITObim and NOVOplasty approaches were used for the last contig obtained. Additionally, to test suitability and to review if coverage changes actually occurred around the edges of the IR's, a comparison of coverage graphs throughout iterations and visual inspection of the annotated picture of the resulting cpDNA strands is performed. To do so, the contigs obatined in the points where these coverage changes were annotated with **GeSeq** and **DOGMA**, both of which are freely available online.

### 2.6 Software tools

The following subsection will provide a short overview of the application and tools used for improving the cpDNA assembly results.

The thesis focuses on free-to-use, readily available tools, with no licensing costs incurred. It is of note that Geneious presents a solid alternative to much of the stack, especially in annotation

performance and post-assembly quality control. The program was also not used due to its modest performance under Linux.

**Cutadapt** was used to remove Illumina adapter sequences from short reads in order to reduce upstream contig falsity and error rate.

**PEAR** by Zhang et al, 2014 is a paired end read merger for Illumina reads.

**samtools** by Li et al., 2009 are being used to post-process alignment/reference files in the *.sam* format.

**bedtools** (Quinlan et al, 2010) has been developed at the University of Utah. It is a self-described 'swiss army knife' of genome arithmetic procedures. Specifically, the genomecov option has been used,

**MIRA** (Chevreux et al., 1999) is a powerful genome assembler, used in many assembly projects. As described above, it has been chosen due to its powerful algorithm which reliably assembles organelle genomes.

**MITObim** by Hahn et al, 2013 is an organelle assembly tool based on the MIRA software suite. It uses the bait-and-map approach, reconstructing mainly mtDNA genomes(but can also be used for plastids) by mapping reads to a reference of a varied degree of closeness to the target species. It is most commonly used with Illumina libraries, as MIRA is ill-equipped to handle uncorrected data produced by other technologies.

**blast** by Altschul et. Al, 1997 is a commonly used software suite for biological analysis.

To further compare results, I have used **NOVOplasty** by Dierckxsens et al on the *Semele* dataset. The script uses genome sequencing data in a seed-and-extend manner to assemble circular cpDNA genomes.

The cpDNA strands were annotated with **Chlorobox GeSeq (https://chlorobox.mpimp-golm.mpg.de/geseq.html)** by Tillich et al, 2017 and/or **DOGMA** by Wyman et al., 2004.

# 3. Results

### Shotgun sequencing results

A total of 2,667,658 pared reads were obtained from the MiSeq run. From these 1,706,680 were retained after the quality control. A total of 1,576,198 paired reads overlapped and were merged with PEAR.

### 3.1 MITObim Results

The results of the assembly tests using MITObim are detailed below. The script has run in de-novo mode, using the same five baits, **MATK, ATPB, NDHF, RBCL** and a general internal **SPACER,** an increasing number of times, until it concluded into a large scaffold. According to the program, the number of iterations required to recover the complete chloroplast varied between 174 for the rbcl and 242 for the spacer (Table 1). The contig length also varied between 96480 and 122397 depending on the bait. The assembly with the highest coverage (62,5) was obtained when the **ndhF** marker was used as bait. The other baits resulted in a coverage ranging between 59,55 and 62,5. The GC content ranged from 36,81% and 36,92% for the **ndhF** bait.

The assembly results were aligned using **blastn** to the *P. ginseng damaya* chloroplast genome. The results were very good, with the whole range of the contigs matching with the chloroplast genome reference. This is expected due to the high conservation rate of the plastid genome. The identities spun a range between a low of 72% and a max of 100% across all 5 baits (Figure 5). The alignment gap percentage was also at a low level, always under 10%. This is also expected and consistent with other test attempts and journal research. All contigs returned similar identities when matched to the assembled reference. When all contigs are considered all the chloroplast sequence information was recovered.

**MATK – 118 iterations**

| Name | Length | Avg. qual | No. reads | Max. cov | Avg. cov | GC% |
|---|---|---|---|---|---|---|
| HM640556.1 | 96568 | 88 | 18111 | 172 | 62.44 | 36.91 |

**ATPB – 179 iterations**

| Name | Length | Avg. qual | No. reads | Max. cov | Avg. cov | GC% |
|---|---|---|---|---|---|---|
| JX903682.1 | 122309 | 88 | 21800 | 172 | 59.58 | 36.81 |

**NDHF – 184 iterations**

| Name | Length | Avg. qual | No. reads | Max. cov | Avg. cov | GC% |
|---|---|---|---|---|---|---|
| JX903263.1 | 96480 | 88 | 18112 | 172 | 62.5 | 36.92 |

**SPACER – 242 iterations**

| Name | Length | Avg. qual | No. reads | Max. cov | Avg. cov | GC% |
|---|---|---|---|---|---|---|
| L41571.1 | 122397 | 88 | 21803 | 172 | 59.55 | 36.81 |

**RBCL – 174 iterations**

| Name | Length | Avg. qual | No. reads | Max. cov | Avg. cov | GC% |
|---|---|---|---|---|---|---|
| HM640442.1 | 122397 | 88 | 21803 | 172 | 59.55 | 36.81 |

*Table 1: The results of the S. androgyna assembly by MITObim.*

In order to obtain more quality scores and to further inspect the assembled results, alignment has been pursued using the on-line blastn suite, with parameters described in chapter 2 (default parameters). The results are visible in the charts below, with identity/gap percentage drawn on the y-axis and the number of assembled contigs on the y-axis.

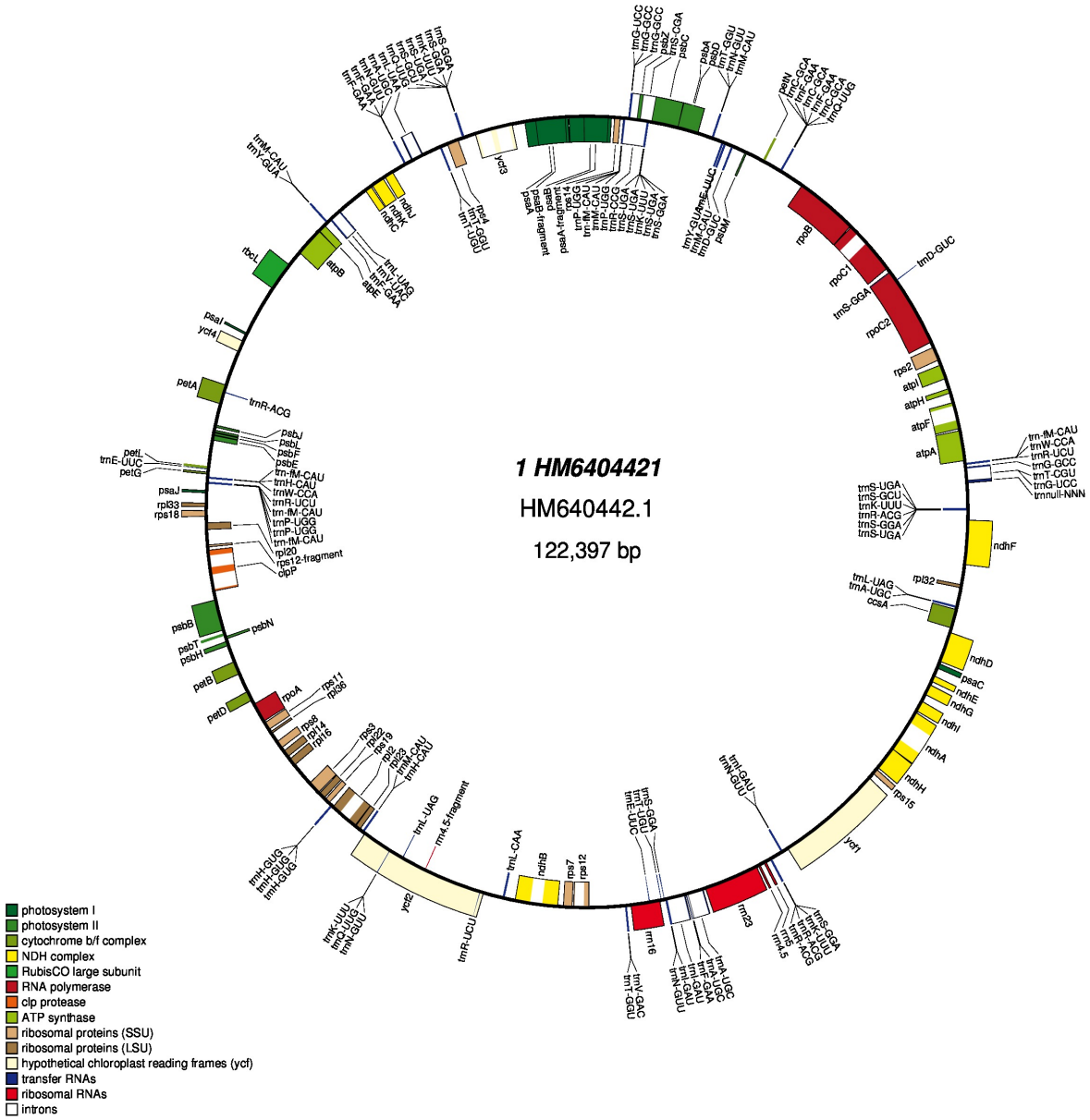



*Figure 5: The annotated cpDNA genomic contig prediction of Semele androgyna*

Based on the blast results of *Semele androgyna,* a contig was manually constructed with its annotation depicted in Figure 5. The bait choice has fallen on the **rbcL** bait: it produced the longest contig, together with the **SPACER** bait. The location of **rbcL** on the model cpDNA strand tipped the scale in it's favour: it is located in the middle of the LSC, which gives the assembly leeway and iteration time until it reaches the point in which it starts to assemble the edge of the IR (A or B).

As the aim of this thesis is not to provide a finite, complete genome, assumptions on completeness or correctness will not be made. The hurdles towards that result are thoroughly described in both

chapters 1 and 2, thus they will not be detailed here. It is sufficient to only mention them as a theoretical guidance for the next two sub-chapters.

- A single set of reads from a single individual sequenced by a single technology means that availability of good quality data is limited.

- Thus, it is difficult to ascertain whether an assembled contig is a ‘real’ (on both a biological and computational level) cpDNA strand.

### 3.2 NOVOplasty Results

**MATK**

| k value | No. of contigs | Largest contig | Smallest contig | Ins. Size | Total contig length |
|---|---|---|---|---|---|
| 20 | 4 | 45470 | 2894 | 365 | 122500 |
| **25** | **8** | **38771** | **319** | **369** | **166600** |
| 30 | 6 | 73244 | 319 | 373 | 204900 |
| 35 | 6 | 56468 | 386 | 374 | 186670 |
| 40 | 6 | 56468 | 389 | 377 | 186676 |

**ATPB**

| k value | No. of contigs | Largest contig | Smallest contig | Ins. Size | Total contig length |
|---|---|---|---|---|---|
| 20 | 1 | 64921 | 64921 | 394 | 65400 |
| 25 | 1 | 64921 | 64921 | 408 | 64921 |
| **30** | **1** | **64921** | **64921** | **401** | **64921** |
| 35 | 1 | 301 | 301 | 500 | 301 |
| 40 | 1 | 301 | 301 | 500 | 301 |

**NDHF**

| k value | No. of contigs | Largest contig | Smallest contig | Ins. Size | Total contig length |
|---|---|---|---|---|---|
| 20 | 1 | 354 | 354 | 500 | 375 |
| 25 | 1 | 349 | 349 | 500 | 349 |
| 30 | 9 | 71489 | 315 | 368 | 181520 |
| 35 | 7 | 71489 | 319 | 370 | 158839 |
| **40** | **5** | **71489** | **319** | **373** | **158094** |

**SPACER**

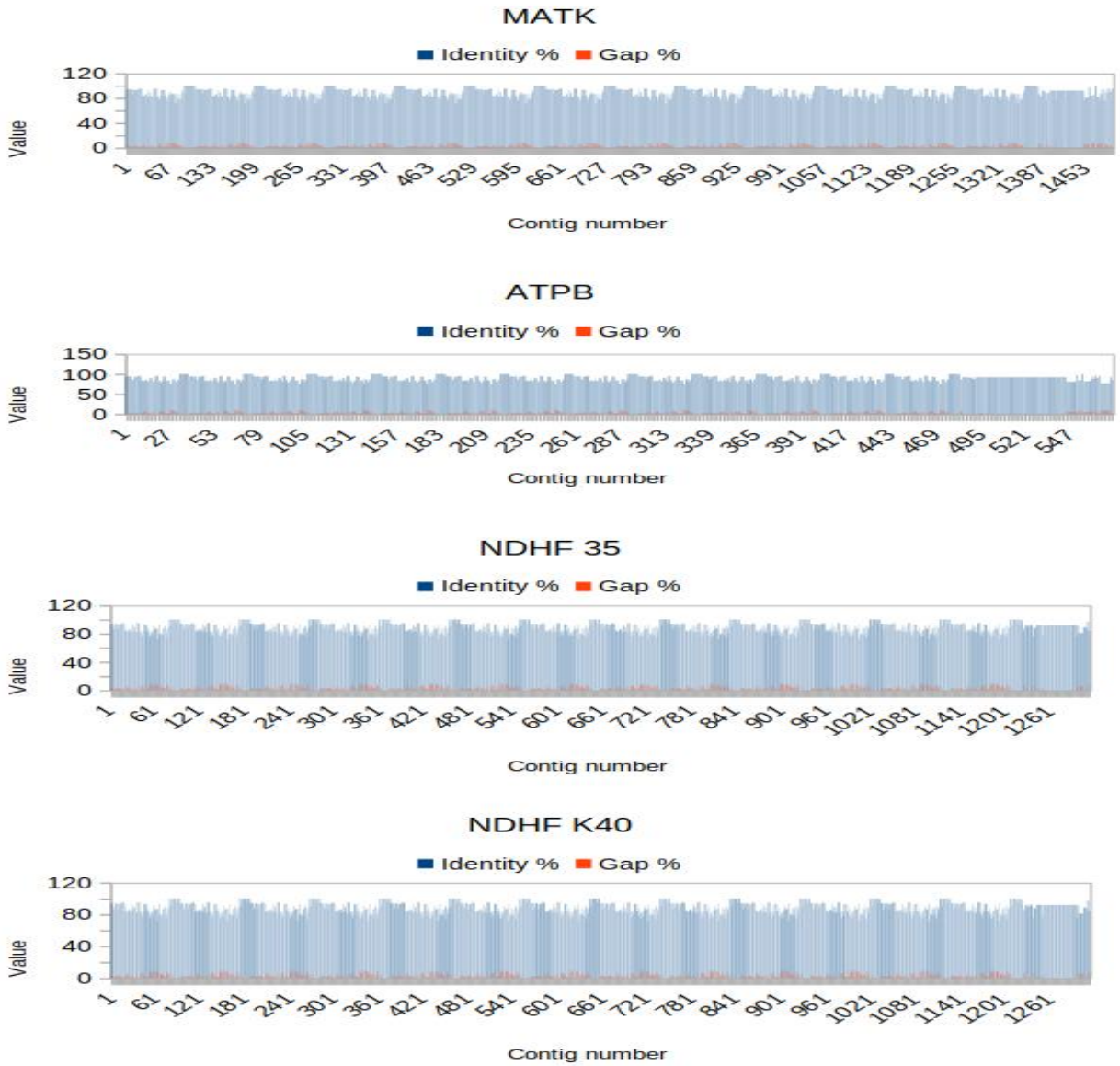
| k value | No. of contigs | Largest contig | Smallest contig | Ins. Size | Total contig length |
|---|---|---|---|---|---|
| 20 | 2 | 26450 | 2894 | 396 | 29344 |
| **25** | **5** | **26450** | **319** | **396** | **59400** |
| 30 | 3 | 26450 | 319 | 397 | 42119 |
| 35 | 4 | 26450 | 419 | 402 | 58465 |
| 40 | 4 | 26450 | 418 | 402 | 58465 |

**RBCL**

| k value | No. of contigs | Largest contig | Smallest contig | Ins. Size | Total contig length |
|---|---|---|---|---|---|
| 20 | 4 | 64921 | 2894 | 376 | 122500 |
| **25** | **8** | **64921** | **319** | **378** | **166700** |
| 30 | 12 | 64963 | 315 | 378 | 230000 |
| 35 | 10 | 64921 | 319 | 384 | 174866 |
| 40 | 10 | 64921 | 319 | 387 | 205560 |

*Table 2: Results of the de novo NOVOplasty assembly, with various statistics described*

**NOVOplasty** uses a *de novo*, k-mer table approach, similar to other string-overlap assemblers such as SSAKE and VCAKE (Dierckxsens et al., 2017). After initial analysis, the most suitable k-values to regard in assembly was **ndhF** (Table 2). Therefore, an as-close-as-possible approach was found to be the most suitable. A reference contig length of 150.000bp was used. Then, included in the analysis were the number of assembled contigs and the total contig length. As it was the case with MITObim, the alignment was done using blastn with the megablast algorithm for highly similar sequences on the KC686331.1 sequence as a reference(the *P. ginseng damaya* genome).



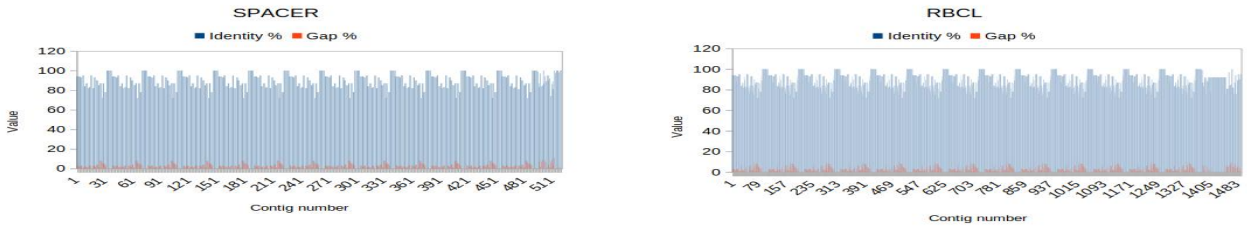



*Figure 6: Alignment statistics of the NOVOplasty-assembled results to the P. ginseng damaya cp genome*

The alignments are found above. On the x-axis, the number of contigs is illustrated, and on the y-axis, the identity percentage in blue and gap percentage in orange are shown.

There is a **strong discrepancy** between the number of contigs assembled by MITObim and the number of contigs aligned by **blastn.** This expected, as the number of gaps might force the blastn aligner to 'split' the assembled scaffold into smaller ones.

| Bait name | NOVOplasty | Blastn |
|---|---|---|
| MATK | 8 | 1494 |
| ATPB | 1 | 570 |
| NDHF K35 | 7 | 1316 |
| NDHF K40 | 5 | 1315 |
| SPACER | 5 | 521 |
| RBCL | 8 | 1495 |

*Table 3: Number of contigs after assembly and after alignment*

Also, after both blasting the contigs with A. thaliana and graphically inspecting the annotation the aligned contigs (Figure 7) , there is a clear trend of the repetition of DNA motifs in tandem, which is the trend that repeats itself over all 6 chosen bait regions. This might indicate repetitive regions or, due to another algorithm used in NOVOplasty against the one used by MITObim(MIRA), another potential sign to **repetitive regions** or **overmapping**. These repetition of motifs only happen after the extention of the IR region.

Both the high number of identity regions and length contigs, paired with the low average gap number across all contigs, point to the **RBCL** assembled contig as a suitable candidate for

annotation. This step is performed in GeSeq, with the default parameters, such as ARAGORN and tRNAscan(cutoff score of 15) for gene prediction.

The results(in both a linear and circular fashion) of the annotation is shown below.

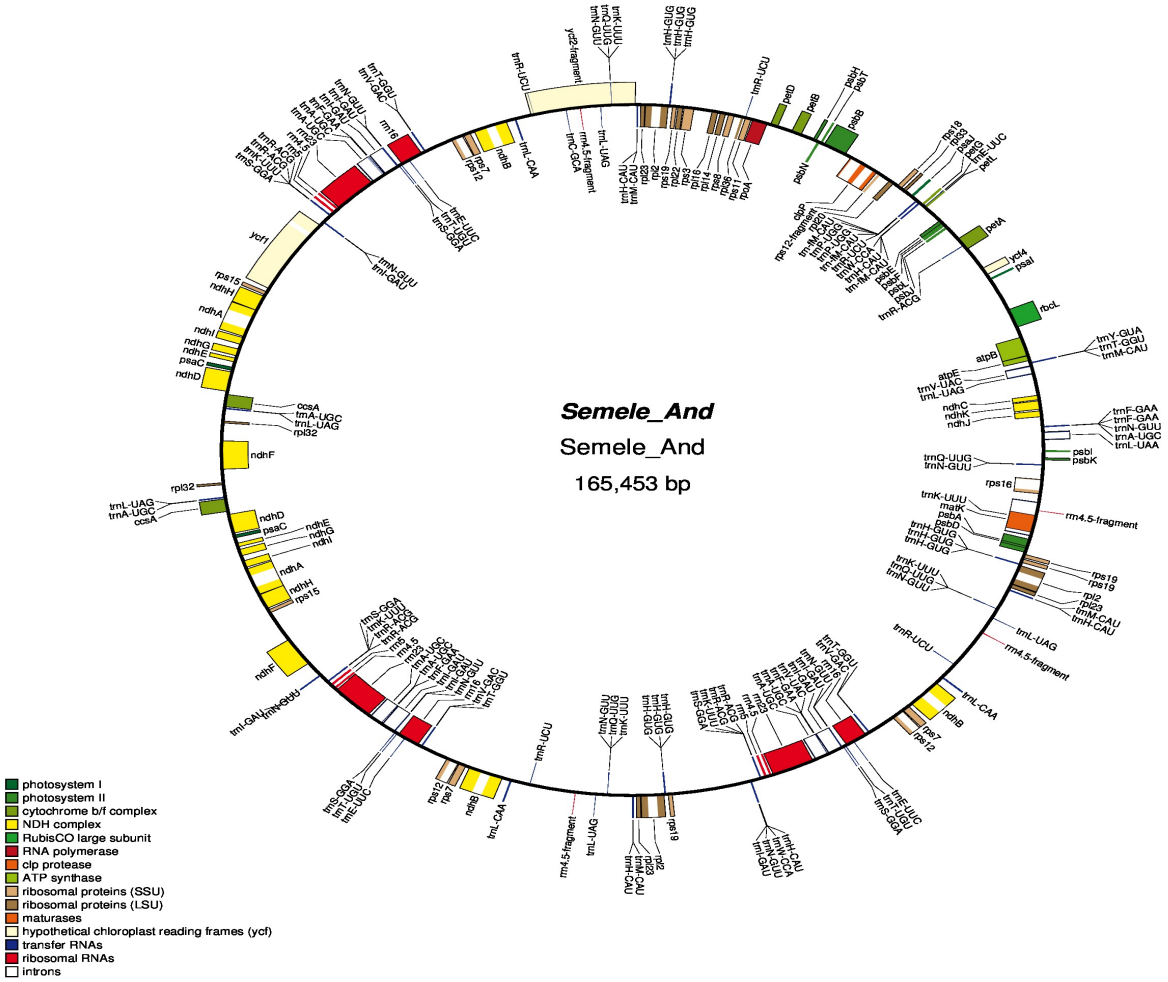



*Figure 7: Circular annotation of the RBCL bait-assembled genome of S. androgyna*

### 3.3 mappy Results

After running the mappy script for 120 iterations the contig lengths varied between 1337 bp for the **spacer** bait and 111151 for the **matk** bait (Table 4). Average coverage varied between 26,88 for **matk** and 101,53 for **spacer**.

As depicted in Figure 9 the average coverage in the beginning and end of the contigs varies depending on the iteration. As it is marked in the Figure 8 there were several points where there was change in coverage. This change in coverage corresponded to an increase for the baits located in the single copy regions and a decrease for the bait located in the inverted repeat. The average coverage considering the complete contig length showed a similar but not so pronounced pattern (Figure 11). The contig length increased gradually tending to reach a plateau.

| SPACER | | | RBCL | | | NDHF | | |
|---|---|---|---|---|---|---|---|---|
| Iterations | Contig length | Av. coverage/contig | Iterations | Contig length | Av. coverage/contig | Iterations | Contig length | Av. coverage/contig |
| 0 | 1337 | 68.09 | 0 | 2417 | 38.51 | 0 | 2973 | 35.36 |
| 1 | 2338 | 75.87 | 1 | 3131 | 41.99 | 1 | 3576 | 46.67 |
| 2 | 3242 | 92.14 | 2 | 3941 | 40.54 | 2 | 4325 | 47.8 |
| 3 | 4193 | 98.27 | 3 | 4862 | 39.99 | 3 | 5161 | 50.97 |
| 4 | 5187 | 101.53 | 4 | 5716 | 42.58 | 4 | 5983 | 52.14 |
| 5 | 6182 | 98.02 | 5 | 6599 | 42.88 | 5 | 6706 | 53.82 |
| 6 | 7043 | 97.6 | 6 | 7455 | 42.6 | 6 | 7569 | 54.96 |
| 7 | 7798 | 96.14 | 7 | 8262 | 43.31 | 7 | 8466 | 58.96 |
| 8 | 8480 | 94.37 | 8 | 9134 | 44.39 | 8 | 9313 | 60.53 |
| 9 | 9382 | 91.18 | 9 | 9739 | 44.46 | 9 | 10226 | 62.89 |
| 10 | 10219 | 90.36 | 10 | 10462 | 44.17 | 10 | 11142 | 64.77 |
| 11 | 11061 | 88.9 | 11 | 10951 | 43.61 | 11 | 12131 | 66.51 |
| 12 | 11962 | 87.96 | 12 | 11607 | 42.56 | 12 | 13042 | 69.58 |
| 13 | 12750 | 88.24 | 13 | 12630 | 42 | 13 | 13967 | 70.09 |
| 14 | 13505 | 86.69 | 14 | 13547 | 42.74 | 14 | 14778 | 70.65 |
| 15 | 14427 | 84.77 | 15 | 14451 | 43.06 | 15 | 15600 | 70.55 |
| 16 | 15407 | 83.85 | 16 | 15326 | 43.31 | 16 | 16557 | 70.3 |
| 17 | 16375 | 82.37 | 17 | 15968 | 44.06 | 17 | 17485 | 71.16 |
| 18 | 17204 | 82.73 | 18 | 16871 | 43.35 | 18 | 18437 | 71.5 |
| 19 | 18179 | 81.78 | 19 | 17724 | 43.83 | 19 | 19203 | 73.12 |
| 20 | 19110 | 81.57 | 20 | 18508 | 44.2 | 20 | 20021 | 72.79 |
| 21 | 20012 | 82.13 | 21 | 19452 | 44.08 | 21 | 20820 | 72.79 |
| 22 | 20808 | 82.64 | 22 | 20280 | 43.82 | 22 | 21538 | 73.3 |
| 23 | 21722 | 82.2 | 23 | 20942 | 43.05 | 23 | 22512 | 72.17 |
| 24 | 22564 | 82.13 | 24 | 22023 | 42.62 | 24 | 23489 | 72.85 |
| 25 | 23509 | 82.27 | 25 | 22703 | 42.11 | 25 | 24390 | 73.03 |
| 26 | 24347 | 82.67 | 26 | 23621 | 42 | 26 | 25277 | 72.95 |
| 27 | 25287 | 82.46 | 27 | 24617 | 41.61 | 27 | 26160 | 73.5 |
| 28 | 26224 | 82.18 | 28 | 25445 | 41.1 | 28 | 27118 | 73.75 |
| 29 | 26947 | 82.78 | 29 | 26238 | 40.81 | 29 | 28198 | 73.29 |
| 30 | 27851 | 82.55 | 30 | 27110 | 40.73 | 30 | 29074 | 73.58 |
| 31 | 28697 | 82.29 | 31 | 27865 | 41.17 | 31 | 29802 | 74.15 |
| 32 | 29647 | 81.93 | 32 | 28542 | 41.32 | 32 | 30569 | 74.78 |
| 33 | 30489 | 81.83 | 33 | 29448 | 41.2 | 33 | 31496 | 74.09 |
| 34 | 31426 | 81.75 | 34 | 30273 | 41.23 | 34 | 32446 | 74.27 |
| 35 | 32209 | 82.22 | 35 | 31196 | 41.29 | 35 | 33212 | 74.66 |
| 36 | 33193 | 81.53 | 36 | 32161 | 41.24 | 36 | 34004 | 74.68 |
| 37 | 34113 | 81.49 | 37 | 33154 | 41.42 | 37 | 34875 | 74.49 |
| 38 | 35113 | 81.15 | 38 | 34063 | 42.19 | 38 | 35750 | 74.55 |
| 39 | 35923 | 81.06 | 39 | 34953 | 42.8 | 39 | 36578 | 74.56 |
| 40 | 36503 | 81.03 | 40 | 35924 | 43.07 | 40 | 37232 | 74.93 |
| 41 | 37202 | 80.58 | 41 | 36666 | 43.39 | 41 | 38053 | 74.74 |
| 42 | 38051 | 80.03 | 42 | 37476 | 43.58 | 42 | 38480 | 75.09 |
| 43 | 38811 | 79.34 | 43 | 38347 | 43.6 | 43 | 38999 | 75.32 |
| 44 | 39781 | 78.46 | 44 | 39144 | 43.76 | 44 | 39482 | 75.84 |
| 45 | 40684 | 78.15 | 45 | 39984 | 44.06 | 45 | 39862 | 76.3 |
| 46 | 41688 | 77.82 | 46 | 40742 | 43.66 | 46 | 40310 | 76.38 |
| 47 | 42709 | 78.3 | 47 | 41714 | 43.54 | 47 | 40740 | 76.42 |
| 48 | 43556 | 78.79 | 48 | 42622 | 43.72 | 48 | 41184 | 76.6 |
| 49 | 44450 | 79.2 | 49 | 43401 | 44.2 | 49 | 41684 | 76.59 |
| 50 | 45137 | 79.31 | 50 | 44262 | 44.21 | 50 | 42220 | 76.59 |
| 51 | 45618 | 79.32 | 51 | 45199 | 44.04 | 51 | 42714 | 76.69 |
| 52 | 46119 | 78.76 | 52 | 46124 | 43.94 | 52 | 43204 | 76.94 |
| 53 | 46554 | 78.33 | 53 | 46961 | 44.21 | 53 | 43733 | 77.62 |
| 54 | 47053 | 77.99 | 54 | 47861 | 44.49 | 54 | 44203 | 78.36 |
| 55 | 47500 | 77.57 | 55 | 48742 | 44.64 | 55 | 44677 | 79.02 |
| 56 | 47923 | 77.25 | 56 | 49665 | 44.62 | 56 | 45200 | 79.29 |
| 57 | 48439 | 76.85 | 57 | 50515 | 44.31 | 57 | 45640 | 79.3 |
| 58 | 48941 | 76.5 | 58 | 51348 | 44.52 | 58 | 46034 | 78.92 |
| 59 | 49371 | 76.14 | 59 | 52306 | 44.41 | 59 | 46461 | 78.48 |
| 60 | 49829 | 75.8 | 60 | 53223 | 44.54 | 60 | 46967 | 78.13 |
| 61 | 50247 | 75.62 | 61 | 53950 | 44.36 | 61 | 47411 | 77.71 |
| 62 | 50703 | 75.45 | 62 | 54974 | 43.95 | 62 | 47834 | 77.4 |
| 63 | 51176 | 75.23 | 63 | 55958 | 44.05 | 63 | 48349 | 76.99 |
| 64 | 51610 | 74.98 | 64 | 56847 | 44.07 | 64 | 48856 | 76.63 |
| 65 | 52039 | 74.75 | 65 | 57793 | 44.31 | 65 | 49279 | 76.28 |
| 66 | 52550 | 74.5 | 66 | 58625 | 44.66 | 66 | 49741 | 75.94 |
| 67 | 53000 | 74.31 | 67 | 59552 | 45.46 | 67 | 50158 | 75.75 |
| 68 | 53455 | 74.13 | 68 | 60420 | 45.8 | 68 | 50614 | 75.58 |
| 69 | 53911 | 74.05 | 69 | 61336 | 46.68 | 69 | 51086 | 75.36 |
| 70 | 54301 | 73.84 | 70 | 62247 | 47 | 70 | 51525 | 75.11 |
| 71 | 54767 | 73.6 | 71 | 63254 | 47.16 | 71 | 51946 | 74.88 |
| 72 | 55190 | 73.42 | 72 | 64144 | 47.55 | 72 | 52464 | 74.63 |
| 73 | 55596 | 73.16 | 73 | 65011 | 47.91 | 73 | 52911 | 74.44 |
| 74 | 56024 | 72.96 | 74 | 65787 | 48.24 | 74 | 53375 | 74.25 |
| 75 | 56459 | 72.8 | 75 | 66677 | 48.47 | 75 | 53825 | 74.17 |
| 76 | 56911 | 72.67 | 76 | 67561 | 48.66 | 76 | 54213 | 73.96 |
| 77 | 57432 | 72.46 | 77 | 68372 | 49.05 | 77 | 54676 | 73.73 |
| 78 | 57814 | 72.37 | 78 | 69244 | 49.54 | 78 | 55101 | 73.53 |
| 79 | 58248 | 72.06 | 79 | 70303 | 49.73 | 79 | 55507 | 73.28 |
| 80 | 58690 | 71.69 | 80 | 71333 | 50.15 | 80 | 55935 | 73.08 |
| 81 | 59104 | 71.53 | 81 | 72132 | 50.86 | 81 | 56367 | 72.93 |
| 82 | 59522 | 71.38 | 82 | 73044 | 50.94 | 82 | 56826 | 72.78 |
| 83 | 59972 | 71.16 | 83 | 73940 | 51.29 | 83 | 57343 | 72.57 |
| 84 | 60430 | 71 | 84 | 74969 | 51.55 | 84 | 57734 | 72.48 |
| 85 | 60724 | 70.96 | 85 | 75910 | 52.14 | 85 | 58162 | 72.17 |
| 86 | 61148 | 70.6 | 86 | 76707 | 52.63 | 86 | 58603 | 71.8 |
| 87 | 61501 | 70.38 | 87 | 77617 | 52.91 | 87 | 59015 | 71.64 |
| 88 | 61879 | 70.13 | 88 | 78437 | 53.31 | 88 | 59434 | 71.49 |
| 89 | 62399 | 69.81 | 89 | 79193 | 53.52 | 89 | 59882 | 71.27 |
| 90 | 62813 | 69.46 | 90 | 80150 | 53.76 | 90 | 60345 | 71.1 |
| 91 | 63216 | 69.21 | 91 | 81116 | 54.17 | 91 | 60631 | 71.07 |
| 92 | 63760 | 68.87 | 92 | 82082 | 54.52 | 92 | 61059 | 70.7 |
| 93 | 64223 | 68.59 | 93 | 82727 | 54.77 | 93 | 61410 | 70.48 |
| 94 | 64713 | 68.38 | 94 | 83662 | 55.11 | 94 | 61791 | 70.24 |
| 95 | 65250 | 68.36 | 95 | 84662 | 55.22 | 95 | 62310 | 69.91 |
| 96 | 65705 | 68.39 | 96 | 85618 | 55.48 | 96 | 62726 | 69.56 |
| 97 | 66023 | 68.29 | 97 | 86400 | 55.69 | 97 | 63134 | 69.32 |
| 98 | 66451 | 68.02 | 98 | 87322 | 55.91 | 98 | 63680 | 68.96 |
| 99 | 66914 | 67.86 | 99 | 88225 | 55.96 | 99 | 64137 | 68.7 |
| 100 | 67272 | 67.67 | 100 | 89083 | 56.22 | 100 | 64625 | 68.48 |
| 101 | 67783 | 67.45 | 101 | 89969 | 56.26 | 101 | 65153 | 68.45 |
| 102 | 68214 | 67.44 | 102 | 90878 | 56.38 | 102 | 65592 | 68.37 |
| 103 | 68592 | 67.24 | 103 | 91801 | 56.73 | 103 | 65932 | 68.41 |
| 104 | 69011 | 67.05 | 104 | 92575 | 57.05 | 104 | 66376 | 68.18 |
| 105 | 69482 | 66.74 | 105 | 93306 | 57.39 | 105 | 66829 | 67.98 |
| 106 | 69864 | 66.56 | 106 | 94305 | 57.53 | 106 | 67190 | 67.76 |
| 107 | 70385 | 66.41 | 107 | 95138 | 57.78 | 107 | 67709 | 67.61 |
| 108 | 70686 | 66.43 | 108 | 95974 | 57.97 | 108 | 68118 | 67.49 |
| 109 | 71074 | 66.06 | 109 | 96828 | 58.08 | 109 | 68504 | 67.34 |
| 110 | 71574 | 65.9 | 110 | 97631 | 58.47 | 110 | 68938 | 67.23 |
| 111 | 72100 | 65.85 | 111 | 98429 | 58.63 | 111 | 69386 | 66.8 |
| 112 | 72444 | 65.65 | 112 | 99311 | 58.86 | 112 | 69779 | 66.66 |
| 113 | 72762 | 65.53 | 113 | 100222 | 59.1 | 113 | 70317 | 66.57 |
| 114 | 73281 | 65.27 | 114 | 100660 | 59.36 | 114 | 70587 | 66.43 |
| 115 | 73733 | 65.36 | 115 | 101097 | 59.71 | 115 | 70990 | 66.16 |
| 116 | 74214 | 65.35 | 116 | 101565 | 59.97 | 116 | 71492 | 65.99 |
| 117 | 74651 | 65.14 | 117 | 102057 | 59.96 | 117 | 72010 | 65.94 |
| 118 | 75091 | 65.02 | 118 | 102520 | 60.1 | 118 | 72356 | 65.73 |
| 119 | 75480 | 64.92 | 119 | 102905 | 60.38 | 119 | 72676 | 65.63 |

*Table 4: Average contig and contig length of all 5 baits assembled by mappy*

| MATK | | | ATPB | | |
|---|---|---|---|---|---|
| Iterations | Contig length | Av. coverage/contig | Iterations | Contig length | Av. coverage/contig |
| 0 | 2510 | 26.88 | 0 | 2365 | 35.92 |
| 1 | 3374 | 30.15 | 1 | 3282 | 38.76 |
| 2 | 4071 | 35.11 | 2 | 4107 | 43.38 |
| 3 | 4841 | 36 | 3 | 5097 | 44.47 |
| 4 | 5681 | 37.55 | 4 | 5916 | 46.68 |
| 5 | 6613 | 39.73 | 5 | 6836 | 47.1 |
| 6 | 7395 | 47.33 | 6 | 7258 | 47.4 |
| 7 | 8338 | 50.46 | 7 | 7830 | 45.4 |
| 8 | 9279 | 55.35 | 8 | 8751 | 43.61 |
| 9 | 10255 | 57.38 | 9 | 9628 | 44.92 |
| 10 | 11251 | 57.06 | 10 | 10454 | 45.43 |
| 11 | 12140 | 57.4 | 11 | 11227 | 44.88 |
| 12 | 12841 | 57.71 | 12 | 12071 | 44.56 |
| 13 | 13687 | 57.4 | 13 | 12917 | 44.96 |
| 14 | 14598 | 57.05 | 14 | 13547 | 44.96 |
| 15 | 15521 | 57.2 | 15 | 14144 | 44.42 |
| 16 | 16441 | 57.83 | 16 | 14961 | 43.51 |
| 17 | 17369 | 59.16 | 17 | 15911 | 42.62 |
| 18 | 18306 | 59.75 | 18 | 16722 | 42.22 |
| 19 | 19247 | 59.99 | 19 | 17745 | 41.69 |
| 20 | 19936 | 62.97 | 20 | 18503 | 41.95 |
| 21 | 20938 | 62.56 | 21 | 19445 | 41.76 |
| 22 | 21782 | 63.11 | 22 | 20199 | 41.75 |
| 23 | 22697 | 63.74 | 23 | 21127 | 41.59 |
| 24 | 23524 | 64.6 | 24 | 21947 | 42.06 |
| 25 | 24387 | 65.12 | 25 | 22807 | 42.15 |
| 26 | 25430 | 65.07 | 26 | 23638 | 42.69 |
| 27 | 26275 | 65.78 | 27 | 24348 | 42.86 |
| 28 | 27012 | 66.67 | 28 | 25223 | 42.32 |
| 29 | 27978 | 66.59 | 29 | 26168 | 41.84 |
| 30 | 28918 | 67.57 | 30 | 27056 | 41.99 |
| 31 | 29865 | 67.73 | 31 | 28003 | 41.91 |
| 32 | 30823 | 68.25 | 32 | 28945 | 41.99 |
| 33 | 31778 | 68.79 | 33 | 29847 | 42.72 |
| 34 | 32788 | 68.8 | 34 | 30714 | 42.75 |
| 35 | 33844 | 68.61 | 35 | 31595 | 42.62 |
| 36 | 34826 | 69.03 | 36 | 32350 | 43.1 |
| 37 | 35659 | 69.27 | 37 | 33082 | 43.5 |
| 38 | 36736 | 69.09 | 38 | 34010 | 43.36 |
| 39 | 37733 | 69.32 | 39 | 34664 | 43.73 |
| 40 | 38699 | 68.85 | 40 | 35527 | 43.62 |
| 41 | 39787 | 69.21 | 41 | 36473 | 43.34 |
| 42 | 40915 | 70.03 | 42 | 37516 | 42.78 |
| 43 | 41774 | 70.45 | 43 | 38281 | 43.14 |
| 44 | 42757 | 70.6 | 44 | 39021 | 43.4 |
| 45 | 43896 | 70.32 | 45 | 39918 | 43.5 |
| 46 | 44898 | 70.54 | 46 | 40837 | 43.49 |
| 47 | 45852 | 71.23 | 47 | 41694 | 43.55 |
| 48 | 46964 | 70.73 | 48 | 42436 | 43.66 |
| 49 | 48066 | 71.37 | 49 | 43405 | 43.71 |
| 50 | 49078 | 71.08 | 50 | 44255 | 44.06 |
| 51 | 50032 | 71.79 | 51 | 45036 | 43.88 |
| 52 | 51133 | 72.02 | 52 | 45843 | 43.97 |
| 53 | 52106 | 72.15 | 53 | 46806 | 44.05 |
| 54 | 53174 | 71.93 | 54 | 47791 | 44.22 |
| 55 | 54172 | 72.02 | 55 | 48726 | 44.33 |
| 56 | 55225 | 71.61 | 56 | 49437 | 44.54 |
| 57 | 56374 | 71.44 | 57 | 50414 | 44.15 |
| 58 | 57385 | 71.18 | 58 | 51333 | 44.22 |
| 59 | 58484 | 71.25 | 59 | 52242 | 44.49 |
| 60 | 59380 | 71.01 | 60 | 53194 | 44.55 |
| 61 | 60541 | 70.06 | 61 | 53927 | 44.28 |
| 62 | 61481 | 69.79 | 62 | 54899 | 43.96 |
| 63 | 62507 | 69.4 | 63 | 55697 | 43.96 |
| 64 | 63497 | 68.88 | 64 | 56621 | 43.9 |
| 65 | 64501 | 68.28 | 65 | 57481 | 43.85 |
| 66 | 65542 | 67.62 | 66 | 58461 | 43.62 |
| 67 | 66572 | 67.03 | 67 | 59372 | 43.73 |
| 68 | 67517 | 66.55 | 68 | 60298 | 43.85 |
| 69 | 68478 | 66.13 | 69 | 61076 | 43.76 |
| 70 | 69357 | 65.76 | 70 | 61931 | 44.05 |
| 71 | 70374 | 65.62 | 71 | 62858 | 44.51 |
| 72 | 71429 | 65.09 | 72 | 63605 | 45.37 |
| 73 | 72445 | 64.74 | 73 | 64527 | 45.88 |
| 74 | 73438 | 64.42 | 74 | 65449 | 46.6 |
| 75 | 74537 | 64.13 | 75 | 66482 | 46.88 |
| 76 | 75619 | 63.85 | 76 | 67405 | 47.27 |
| 77 | 76704 | 63.75 | 77 | 68352 | 47.59 |
| 78 | 77785 | 63.08 | 78 | 69235 | 47.79 |
| 79 | 78652 | 63 | 79 | 70204 | 48.07 |
| 80 | 79594 | 62.51 | 80 | 71204 | 48.36 |
| 81 | 80493 | 62.23 | 81 | 72052 | 48.79 |
| 82 | 81449 | 62.08 | 82 | 72937 | 49.08 |
| 83 | 82438 | 61.63 | 83 | 73765 | 49.51 |
| 84 | 83428 | 61.75 | 84 | 74655 | 49.75 |
| 85 | 84326 | 61.17 | 85 | 75642 | 50.05 |
| 86 | 85224 | 61.59 | 86 | 76526 | 50.73 |
| 87 | 86282 | 60.64 | 87 | 77424 | 50.83 |
| 88 | 87347 | 60.71 | 88 | 78092 | 51.1 |
| 89 | 88333 | 61.27 | 89 | 79021 | 51.24 |
| 90 | 89334 | 60.73 | 90 | 79945 | 51.71 |
| 91 | 90319 | 61.25 | 91 | 80849 | 52 |
| 92 | 91258 | 60.38 | 92 | 81685 | 52.33 |
| 93 | 92399 | 59.96 | 93 | 82560 | 52.74 |
| 94 | 93454 | 59.75 | 94 | 83402 | 52.98 |
| 95 | 94249 | 59.92 | 95 | 84279 | 53.2 |
| 96 | 95164 | 59.68 | 96 | 85267 | 53.41 |
| 97 | 95998 | 59.7 | 97 | 86177 | 53.69 |
| 98 | 96842 | 59.32 | 98 | 87021 | 54.01 |
| 99 | 97715 | 58.92 | 99 | 87889 | 54.33 |
| 100 | 98766 | 58.43 | 100 | 88719 | 54.72 |
| 101 | 99635 | 58.33 | 101 | 89705 | 54.89 |
| 102 | 100557 | 57.86 | 102 | 90581 | 55.11 |
| 103 | 101421 | 57.8 | 103 | 91423 | 55.25 |
| 104 | 102059 | 57.89 | 104 | 92300 | 55.4 |
| 105 | 103049 | 57.23 | 105 | 93107 | 55.67 |
| 106 | 103714 | 57.39 | 106 | 93856 | 55.68 |
| 107 | 104180 | 56.86 | 107 | 94791 | 55.68 |
| 108 | 104732 | 57.09 | 108 | 95752 | 55.97 |
| 109 | 105234 | 56.78 | 109 | 96155 | 56.45 |
| 110 | 105708 | 57.02 | 110 | 96561 | 56.65 |
| 111 | 106151 | 56.57 | 111 | 97036 | 56.83 |
| 112 | 106667 | 56.48 | 112 | 97473 | 57.09 |
| 113 | 107215 | 56.47 | 113 | 97934 | 57.43 |
| 114 | 107499 | 56.54 | 114 | 98408 | 57.58 |
| 115 | 108091 | 56.11 | 115 | 98834 | 58.07 |
| 116 | 108844 | 58.91 | 116 | 99285 | 58.34 |
| 117 | 109717 | 57.26 | 117 | 99754 | 58.74 |
| 118 | 110491 | 62.83 | 118 | 100199 | 59.07 |
| 119 | 111151 | 55.7 | 119 | 100634 | 59.34 |

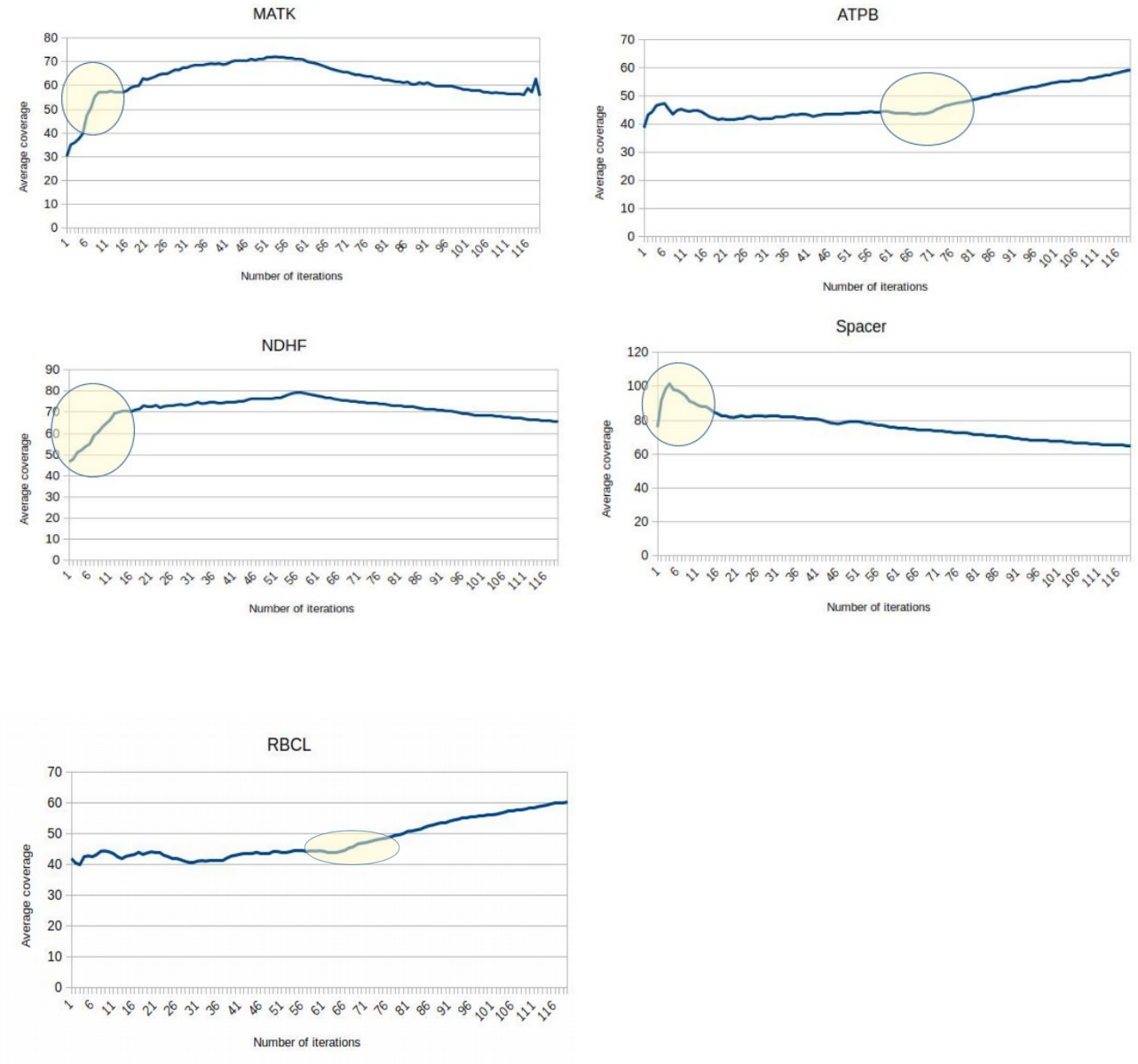



*Figure 8: Average coverage per iteration, with coverage points detailed*

Table 4 provides a numerical overview of the coverage per base pair, and , the coverage changes are depicted in Figure 9 above.

→ **MATK :** the **12th iteration** was chosen, the point that is clearly a part of the spike, and consequently the point(or iteration) in which coverage stabilizes throughout the assembly.

→ **ATPB :** the **71st iteration** was the point that represented a clear disruption. The coverage increases steadily throughout the assembly.

→ **NDHF :** in the **15th iteration,** the coverage reaches a plateau, in which the value stabilizes for a longer amount of increase in contig length.

→ **SPACER :** The **4th iteration** is the clear winner here: it is the highest coverage achieved throughout the assembly process.

→ **RBCL :** The **67th iteration** has been chosen here, as it is the point that marks a continuous increase in assembly coverage.

The existence of these junction regions, characterized by changes, points towards a hypothesis of this thesis: **the assembly of border regions of IRs in cpDNA genomes causes an increase in coverage.** On the other hand, for the bait in the inverted repeat there is a decrease of coverage, which corresponds to the starting of the assembly of the single copy region.

To ensure that these points corresponded to the junctions of the inverted repeat region, the contigs from the above mentioned iterations were blasted and annotated using the the *P. ginseng damaya* cpDNA as reference(Figure 11).

When aligning the selected contigs at the iterations 12, 71, 15, 4 und 67 it is to note that hey are not part of a completed cpDNA scaffold, thus it is expected to find a lower identity percentage and number of contigs throughout the baits.

In Figure 10 below, a side-by-side comparison of both length-based and coverage-based variation throughout 120 iterations is presented.

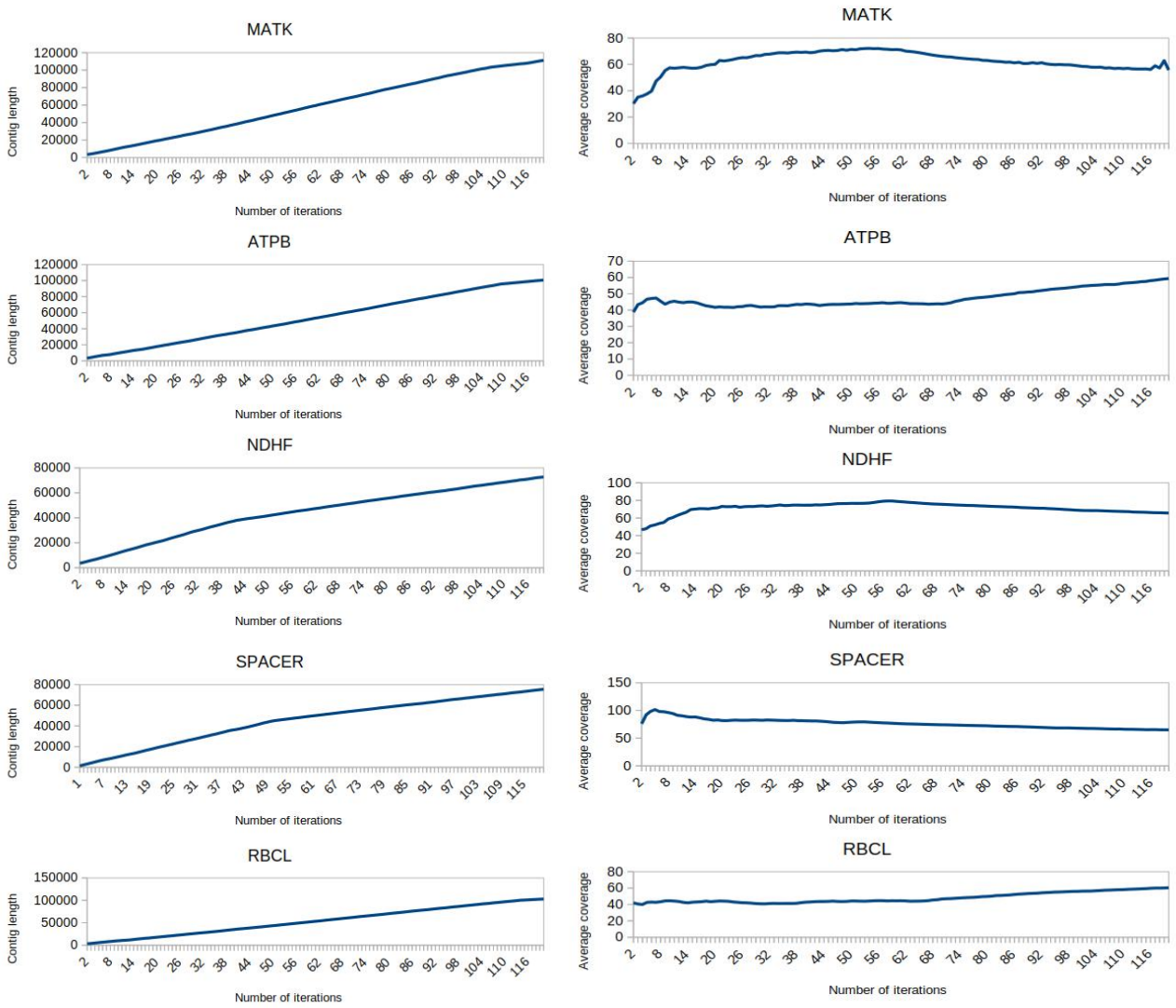



*Figure 9: A side-by-side comparison between the contig length, number of iterations and the average coverage considering the complete contig*

| Bait name | Iteration number | Coverage | GC% |
|---|---|---|---|
| MATK | 12 | 57,71 | 35,33 |
| ATPB | 71 | 44,51 | 36,07 |
| NDHF | 15 | 70,55 | 39,21 |
| SPACER | 4 | 101.53 | 47,97 |
| RBCL | 67 | 45,46 | 35,96 |

*Table 5: The coverage and GC% of the selected baits*

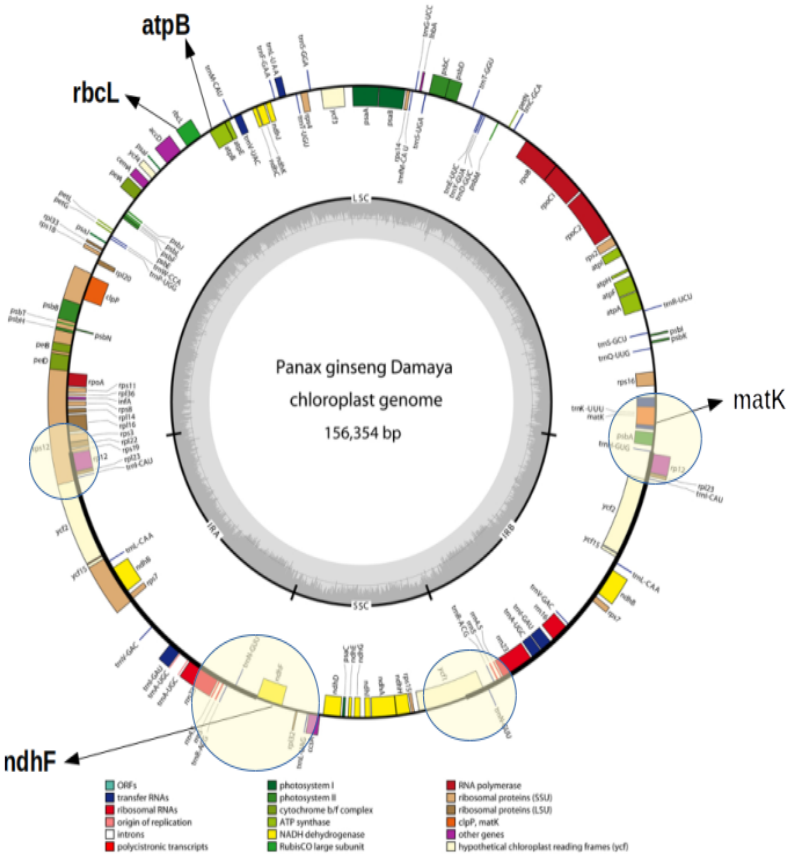



*Figure 10: The annotated chloroplast of P. ginseng Damaya, showing the regions where there was a change of coverage.*

The spots defined as depth 'jumping points' (and highlighted in Figure 10) **roughly coincide with the depth increase slope derived from the mappy results.** These junctions were obtained based on the annotation shown in Figures 15 to 16 the genes present in the beginning and the end

of the contigs from these iterations where the coverage changes corresponds to the genes in the junctions of the IR in the P. geniseng genome

The alignment of the contigs to the *P. ginseng damaya* cpDNA strand is performed with blastn (megablast, for high similarity) with the default settings chosen. The results of this alignment are shown below.

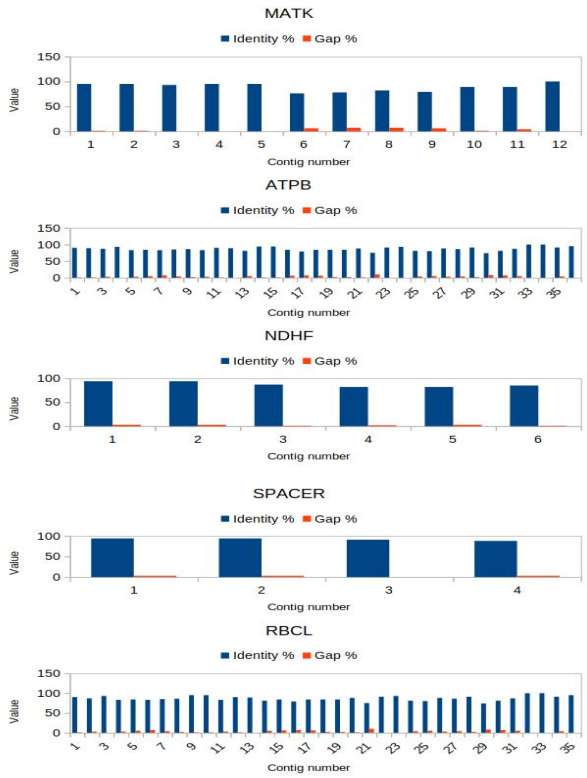


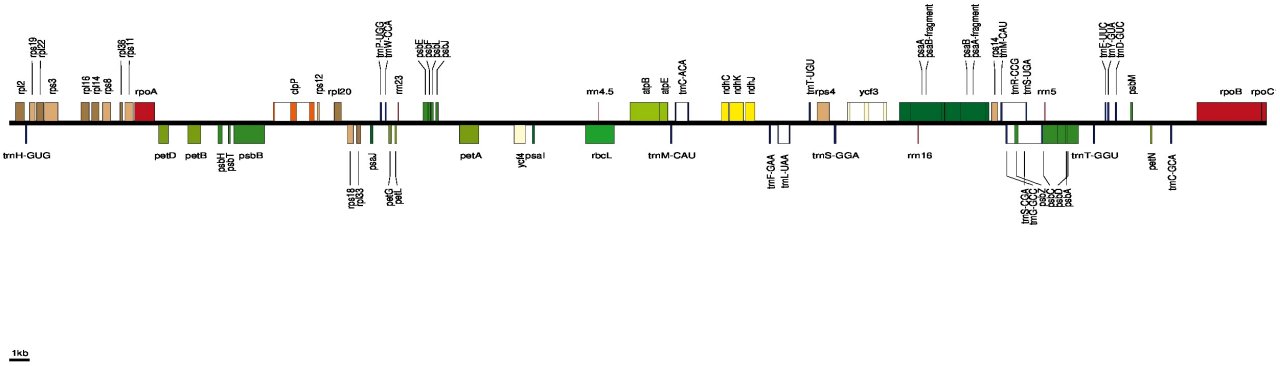

*Figure 11: blastn alignment graphs of the five chosen contig baits*

After alignment, the contigs are annotated using the GeSeq suite. Annotation was performed using both the 'circular' and the 'non-circular' option, for easier identification of coding sequences.

***ATPB---cov***

ATPB---cov

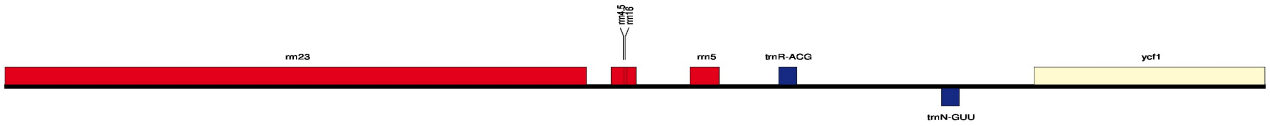
62,858 bp




***SPACER---c***

SPACER---c

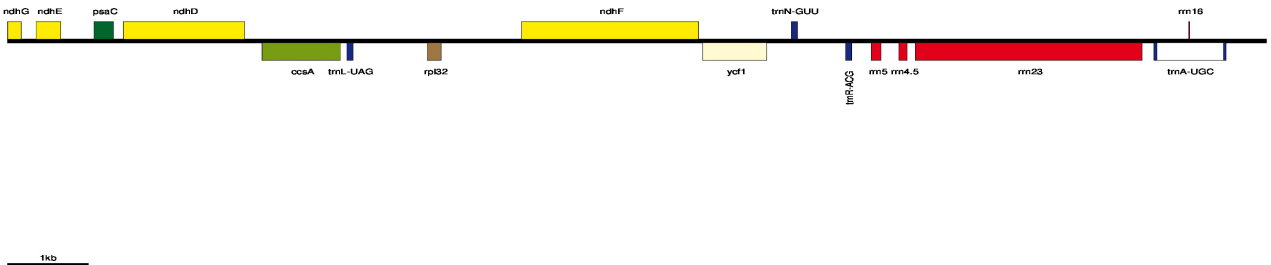
5,187 bp




***NDHF----co***

NDHF----co

15,600 bp



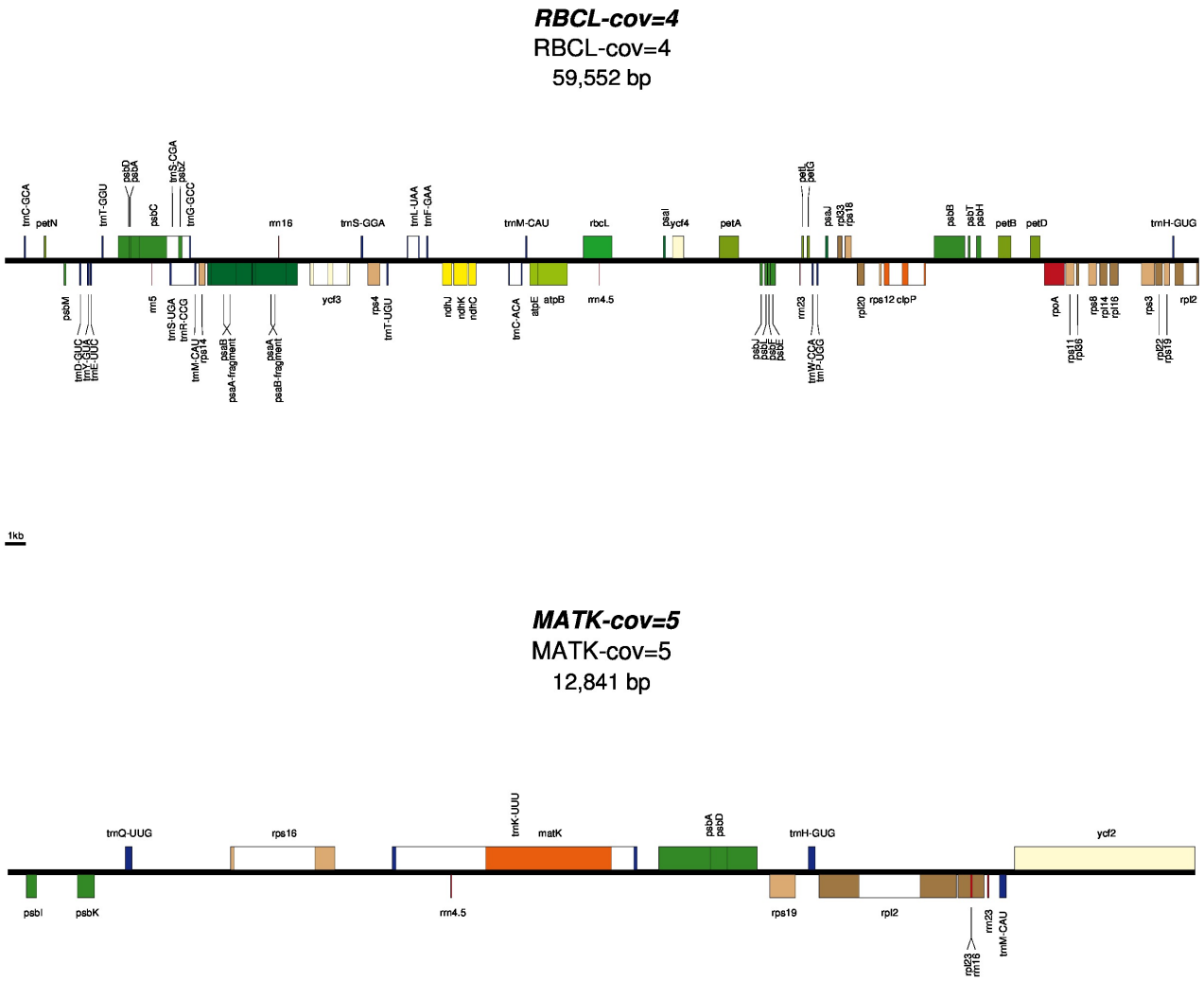



*Figure 12: Linear annotation with GeSeq of mappy contig baits*

# 4. Discussion

## Applicability of complete chloroplast genome for DNA barcoding

The stated purpose of the workflow which concludes with mappy is not being a universal solution to most plastid assembly issues. It can provides an additional analysis blueprint from which further steps can be undertaken to increase robustness of assembled contigs and to add to their applicability, especially for DNA barcoding. The main limitations of the DNA barcoding process is the use of a set of markers that are able to determine a sample at the species level for a wide taxonomic range (Taberlet et al 2009). This is often impossible because either the primers are not specific enough or the loci used are not powerful enough to differentiate between species. Another limitation would be the fact that different loci may show different mutation rates depending on the taxonomic group and thus a marker that works well in one group may fail in the other. These limitations may be completely overcome by using the complete chloroplast genome. The need for a primer preparation process decreases, as species identification can be done when barcoding the whole genome and mapping it to an existing reference. Availability of a complete chloroplast genome makes primer choice obsolete: one can not have more information than what is assembled properly.

Taking this into consideration the approaches described in this thesis can be good solution for the recovery of the complete chloroplast sequence. They do not require any prior information of the cpDNA or primer design. Moreover they do not require a high read coverage being possible to sequence a relative high amount of samples, especially when enrichment approaches are used. Nevertheless, to ensure that the method can be used for DNA barcoding, the quality of the results should be high.

## Mappy-based improvements to assembled contig

When using the existing tools for Semele chloroplast genome assembly, although they were able to recover the complete chloroplast genome, the resulting contigs were composed by regions in the wrong order. This is most likely a consequence of the contig extension method applied in combination with the characteristics of the chloroplast genome. The cpDNA is composed by two inverted repeats regions that are identical among them but just in a different order. While assembling the inverted repeat part reads originating from both regions are assembled simultaneously. Thus, when the IR finishes assembling the reads mapping from the junctions can be from the correct IR or from the wrong one. If, by chance, the wrong IR junction read is assembled the following single copy region will represented in the contig reverse complemented. In case of MITObim, because the mapped reads are eliminated, the assembly stops after the second

single copy region. For this reason, the resulting contigs always had two regions assembled in one direction and the remaining in the other direction. In case of NOVOplasty, this does not happen. Instead the assembly only stops when there is a significant overlap between the beginning and end of the contig. For this reason, the assembly can go on including several copies of inverted repeat and the second single copy region, which explains the results obtained.

The proposed script proposes a novel concept to plastid assembly. It is based on the idea that one just needs the assembly within each one of the chloroplast DNA regions to be correct and these can be analyzed separately. By doing so the limitations of NOVOplasty and MITObim do not apply any more. To make this possible it is necessary to detect the junctions of these regions. The results confirmed that the coverage 'jumping points' at the edge of the contig summarily coincide with the edges of the IR regions. Thus, this statistic can be used to define the junctions of the different chloroplast region. Nevertheless, this approach provides an additional challenge. A complete chloroplast genome using a single is an impossible goal to attain. A set of initial baits from each one of the regions of the genome need to be used. For the context of DNA barcoding, sequences from a close relative species or genera may not be available thus the characteristic of the baits used still need to be tested. These would be how small or dissimilar they can be. In the meanwhile, using Illumina technology together with Sanger sequencing baits can be a good alternative.

**Literature comparison**

Improvements to current methods and software currently lead to significant increases in both quality and quantity of available cpDNA genomes. The following schema (Figure 13) provides an overview of the current workflow found in literature no older than 2016.

Furthermore, the journal articles have been chosen based on method proximity and structure similarity, thus making a comparison viable.

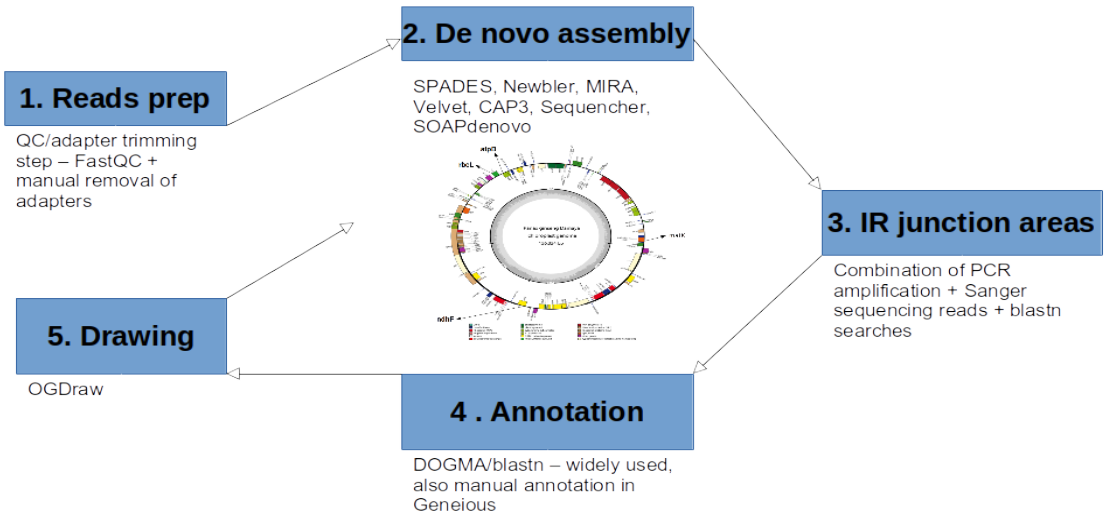



*Figure 13: A graphical synthesis of the assembly workflow found throughout the inspected journal articles*

As seen above, the workflow starts with the **preparation of the reads** for the subsequent assembly. This step is used in all reviewed papers. The quality control step is performed with FastQC or internal tools, to remove Illumina adapters, low quality reads with a phred score of under 20 thus improving upstream quality of assembled contigs.

**Assembly** is mostly done with freely available, open-source assembly programs. Out of the eight programs deployed for this task, SPADES (Bankevich et al, 2012) was used in four examples, making it the most widely used cpDNA assembly software at the moment.

The main objective of the thesis was to investigate the four **IR junction points** (LSC to IRA, LSC to IRB, SSC to IRA and SSC to IRB) and to extrapolate on their location and existence based on coverage changes in those areas. This main novelty was not found in any of the articles. Instead, researchers extracted the points by PCR amplification and overlapping (longer) Sanger reads and then performing a blastn search on those points.

Depending on scope, this step goes hand in hand with **tRNA and intron discovery**. ARACHNE and tRNAscan-se were used in one example. The two apps are used both as stand-alone, or as a part of a online framework such as Geseq.

After an outline of the complete genome has been completed, **annotation** can begin. In this step, the gene content of the assembled contig is identified and corroborated with intron and tRNA content information from the previous step. In all but few cases, DOGMA by Wyman et al., 2004 has been used to perform annotation. The script, hosted on the University of Texas webpage(free to use), can be considered the *de facto* leader in the field. Its wide usage is complemented by

blastn use and Geneious tests, where needed. The output is a text table which includes the name of the gene, its start and end position, ending with the direction of the gene strand.
The **graphical Figure** of the annotated list was conducted in **OGDraw**, which is used to serve an example to the quadripartite structure of the plastid strand(LSC, SSC and two IRs).

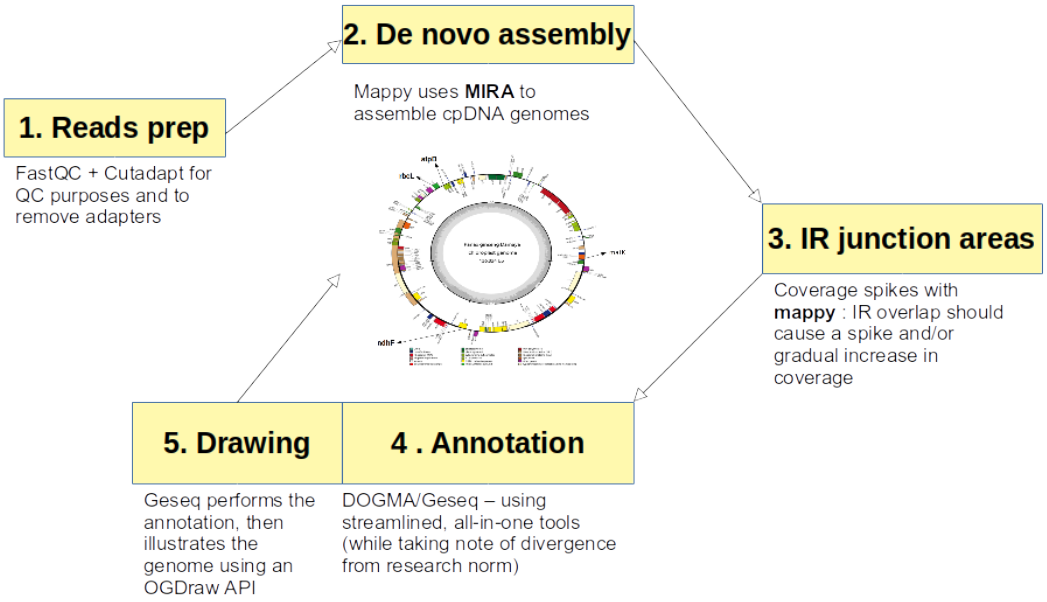



*Figure 14: The mappy workflow*

The **mappy** workflow is similar but not identical to the industry standards, so that the differences have been highlighted in the graphical depiction above.
* Reads are usually cleaned using scripts developed in the corresponding institutes, and seldom by a open source(or paid) tool.
* After many considerations, and the propagation of MITObim as a state-of-the-art tool for assembly using a map-and-extend algorithm, it has been decided that mappy, in it's first versions, will use **MIRA** as it's primary assembly program. This fact adds to the complexity of the process, as a manifest file, which has strict formatting rules, needs to be generated. As mappy had to run for a total of 600 times, a time consideration was also added into the mix. It was fast and efficient, and ran on a local workstation in the INF, therefore confirming it's relative ease of use and resource efficiency.
* **mappy 1.0** produces IR junction information by analyzing coverage changes in the assembly product. Preliminary analysis has shown that the extension at the edges of the IRs triggers a spike in assembly depth. Nevertheless, it is still missing the definition of the coverage difference necessary to accurately define these junctions. The IRs are mirrored, thus the start of IRA, for

example, can be identical to the end point of IRB. As explained in chapters 2 and 3, mappy has been designed based on this condition.

* The annotation/genetic map drawing part was performed in GeSeq, an online tool hosted by the Max Planck Institute. It's use has not been documented in recent journal entries and articles. In order to get a full list of genes(including directions), **DOGMA** was also employed.

## 5. Conclusion

The thesis provides key insights to a relatively new and exciting branch of bioinformatic analysis: de-novo circular genome assembly. The assembly of such sequence information is of particular importance due to its applications in fields such as DNA barcoding and description of existing biodiversity. The three scripts used are considered 'new', or 'cutting edge'. The oldest one is MITObim, published in 2013. Since then, many attempts at an unified workflow have been made, and many projects have been published (Nagy et al, 2017, Ni et al, 2017, Liu et al., 2017 just to name a few). **MITObim** and **NOVOplasty** seem to over- or mis-assemble. When aligned to the reference genome, the output presents repetitive spikes in both identity and gap percentage, which leads us to believe that it also has difficulties with piecing together IR junctions. It would be interesting to add different assembly **baits** and **coverage information** to the results. A more diversified result pool would surely improve this assembly. Hence the importance of developing alternative approaches.

As both sequencing and assembly performance increase, it is to be expected that many new tools are developed and existing ones improved. One of the more prestigious works published are the Assemblathons 1 and 2, by Earl et al, 2011 and Bradnam et al, 2013. One consideration becomes evident after analyzing the results of those papers, and reinforces one of the ideas of this thesis: it is still difficult to present an unifying solution, workflow or app which drastically and directly improves assembly results.

**Mappy** has identified the existence of these coverage points by accurately extracting coverage information from the MIRA output information folder. By visually inspecting the annotated results, it can be seen that the location of these IR 'spikes' is conclusive with an approximate genomic distance to the bait location. This shows that coverage at the edge of the contigs can be used as a measurement to define the junctions among the IR and SSC or LSC regions without any a priori knowledge of the DNA sequence. In all approaches used, the assembly within these parts is correct. In genetic diversity studies and in DNA-barcoding approaches the different parts of chloroplast can be among different individuals compared/aligned separately and then merged retaining this way the whole cpDNA genome information. The use of coverage (or depth) has a

greater applicability because it is not data-specific. Nevertheless, the applicability of mappy to other systems, as well as, the characteristics of the baits and the input sequence data need to be further tested. As described throughout this thesis, mappy was a proof of concept, and is definitely not finished.

Most projects seem to have focused on improving *de novo* assembly results by means of constantly increasing the number of reads, adding longer Sanger reads to accurately assemble a long sequence, or to target a particular one. However, the practicality of such an proposition is sometimes in question. Availability of such technology, research budgets, time constraints and many more factors affect many projects in more than one way. Approaches similar to mappy are good answers to such constrains.

**Mappy** has identified the existence of these coverage points by accurately extracting coverage information from the MIRA output information folder. By visually inspecting the annotated results, it can be seen that the location of these IR 'spikes' is conclusive with an approximate genomic distance to the bait location. As described throughout this thesis, mappy was a proof of concept, and is definitely not finished.

**Improvements and outlook**

There is opportunity for optimization in these workflows and programs. One consideration would be to implement both tools in such a way as them to complement each other: **mappy** provides IR junction points provided by quality measures extracted directly from MIRA outputs and log files. **MITObim** and **NOVOplasty** did not completely satisfy our quality requirements, though this might be just one scenario in which it does not deliver perfect results. A look on the project page (https://github.com/ndierckx/NOVOplasty) shows the latest improvement updates, more recently heteroplasmy calling, functionality that is not available in **MITObim.** This support for bi dimensional heteroplasmy(both for mt and cp) will further refine the results.

In order to improve mappy, a features roadmap is detailed below. I hope that these will be implemented with time.

**→ automatically detect coverage changes and output a contig corresponding to the region of the cpDNA genome that is currently being assembled.**

**→ add a machine learning engine to identify mt or cp reads and proceed with a corresponding set of settings.**

**→ request access to ncbi API, compare local assembled contigs to online sequences automatically.**

**→ add a visualization platform, to see coverage and contig length in real time.**

→ **add fine tuning options such as base quality and automatic read clean-up**

Most importantly, **mappy,** as a complete tool, needs to have its results spread throughout a large number of species. It needs to be proven that the good results are not specific of *S. androgyna* and that it can reliably detect junction points automatically. Moreover, it will allow to specify some of the parameter such as a threshold for detecting significant coverage variation. The minimum required data quality also needs to be tested. And statistics such as the minimum number of reads necessary to recover the complete chloroplast genome can be obtained. Finally, different baits with different degrees of similarity to the target species and different lengths should also be tested.

# 6. Addendum

## 6.1 List of figures



## 6.2 List of tables



## 6.3 Abbreviations

cpDNA – Chloroplast DNA

mtDNA – mitochondrial DNA

nDNA – nuclear DNA

NGS – next generation sequencing

WGS – Whole genome sequencing

GC% - GC content, or percentage of bases G and C in an assembled contig

LSC – Large single copy (of a cpDNA strand)

SSC – short single copy (of a cpDNA strand)

IRA – Inverted repeat A

IRB – Inverted repeat B

### 6.3 mappy – a MIRA-based wrapper for cpDNA workflow procedure, developed in python

```
import os  # libraries used
import sys
import shutil
import subprocess
import numpy as np


name = sys.argv[1] + '-0' # project name
ref_loc = sys.argv[2] # reference location
read_loc = sys.argv[3] # read location
manifest = open('manifest-0.conf', 'w')
```

***→ MIRA uses a manifest file with a specific structure in order to run. Here we create the original file, the ones for further iterations are created further below.***

```
manifest.write('project = {}\njob = genome,mapping,accurate\nparameters =
-GE:not=4 -NW:mrnl=0 -AS:nop=1 SOLEXA_SETTINGS -CO:msr=no COMMON_SETTINGS
-SB:tor=no\nreadgroup\nis_reference\ndata = {}\nstrain = {}\nreadgroup =
reads\ndata = {}\ntechnology=solexa'.format(name,ref_loc,name,read_loc))
manifest.close()
```

***→ Initialization of variables.***

```
avg_cov = 0
count = 0
new_avg_cov = 0
contig_length = 0
beg_cov = 0
beg_end = 0


statistics = open('general_cov_stats_{}.txt'.format(sys.argv[1]), 'w')
statistics_per_bp = open('coverage_per_bp_{}.txt'.format(sys.argv[1]), 'w')
destination = os.path.join('home','user','Desktop','new_test','test_results' +
name)
```

***→ general results folder, containing the outputs per base pair and general depth stats of assembled contigs.***

```
os.mkdir('/{}/'.format(destination))
```

**→ mappy runs a total of 120 times for each bait sequence**

```
for count in range(0,120):
  command1 = 'mira manifest-{}.conf'.format(str(count))
  command2 = 'miraconvert
/home/user/Desktop/new_test/{}_assembly/{}_d_results/{}_out.maf
/home/user/Desktop/new_test/output_{}_converted.sam'.format(name,name,name,name)
  command3 = 'samtools view -bS
/home/user/Desktop/new_test/output_{}_converted.sam -o
/home/user/Desktop/new_test/output_{}_converted.bam'.format(name,name)
  command4 = 'samtools sort
/home/user/Desktop/new_test/output_{}_converted.bam{}/home/user/Desktop/new_test
/output_{}_sorted'.format(name,' ', name)
  command5 = 'samtools index
/home/user/Desktop/new_test/output_{}_sorted.bam'.format(name)
  command6 = 'bedtools genomecov -d -ibam output_{}_sorted.bam >
output_{}_coverage.txt'.format(name,name)

  avg_cov = new_avg_cov
```

***→ After the commands are defined, mappy executes them according to proposed workflow.***

```
  subprocess.call(command1, shell=True)
  print('MIRA Assembly - completed')
  subprocess.call(command2, shell=True)
  print('Converted .maf file to .bam file')
  subprocess.call(command3, shell=True)
  print('Converted .sam output to .bam output')
  subprocess.call(command4, shell=True)
  print('Sorted the .bam file')
  subprocess.call(command5, shell=True)
  print('Indexed the .bam file')
  subprocess.call(command6, shell=True)
  print('Depth per base pair extracted')

  with open('{}_assembly/{}_d_info/{}_info_contigstats.txt'.format(name, name,
name)) as f:
    print('Fishing for coverage:')
    for line in f:
```

```
      if not line.startswith('#'):
        new_avg_cov = float(line.split('\t')[5])
        contig_length = int(line.split('\t')[1])
```

***→ Write the average coverage of the assembled contig into a statistics file.***

```
  statistics = open('general_cov_stats_{}.txt'.format(sys.argv[1]), 'a')
  statistics.write(str(count) + '\t' + str(contig_length) + '\t' +
str(new_avg_cov) + '\n')
  statistics.close()
```

***→ mappy extracts coverage from first and last 100 bp of the output file***

```
 statistics_per_bp = open('coverage_per_bp_{}.txt'.format(sys.argv[1]), 'a')
  data = np.genfromtxt('output_{}_coverage.txt'.format(name), delimiter='\t',
usecols=(2), dtype=int)
  beg_cov = data[:30].mean()
  end_cov = data[-30:].mean()
  statistics_per_bp.write("{} {} {}\n".format(count, beg_cov, end_cov))
  statistics_per_bp.close()
```

***→ remove intermediary files.***

```
  os.remove('./output_{}_coverage.txt'.format(name))
  os.remove('./output_{}_converted.sam'.format(name))
  os.remove('./output_{}_converted.bam'.format(name))
  os.remove('./output_{}_sorted.bam'.format(name))
  os.remove('./output_{}_sorted.bam.bai'.format(name))
```

***→ a folder is created, where all contigs are put in.***

```
result_source=os.path.join('/home/user/Desktop/new_test/{}_assembly/
{}_d_results/'.format(name,name))
  for files in os.listdir(result_source):
    if files.endswith("AllStrains.unpadded.fasta"):
      shutil.copy(result_source + files,'/' + destination + '/')
```

***→ The current assembled contig is chosen as the new bait file and a new manifest file is generated.***

```
  new_bait = name + '_assembly/' + name +'_d_results/'+ name
+'_out_AllStrains.unpadded.fasta'
  print('The new bait is ' + new_bait)
```

***→ Update the name and content of the manifest file and use it as a new manifest file for the next iteration.***

```
  name = name.split('-')[0] + '-' + str(count)

   manifest = open('manifest-' + str(count) +'.conf', 'w')

   print('new manifest is ' + 'manifest-' + str(count) +'.conf')

   manifest.write('project = ' + name +'\njob =
genome,mapping,accurate\nparameters = -GE:not=4 -NW:mrnl=0 -AS:nop=1
SOLEXA_SETTINGS -CO:msr=no COMMON_SETTINGS
-SB:tor=no\nreadgroup\nis_reference\ndata = /home/user/Desktop/new_test/' +
new_bait + '\nstrain = ' + name + '\nreadgroup = reads\ndata = ' + read_loc +
'\ntechnology=solexa')

   manifest.close()
```

→ ***Delete the previous folders.***

```
   if os.path.exists('./manifest-{}.conf'.format(str(count-2))):

     os.remove('./manifest-' + str(count-2) + '.conf')

     shutil.rmtree({}{}{}{}{}{}.format('./',name.split('-')[0],'-',str(count-
2),'_assembly/'))
```

→ ***If the result file exists, it is moved to a destination folder and the other assembly workflow files are deleted (for disk usage reasons).***

```
print('MIRA has run {} times!'.format(str(count)))
```

# 7. Bibliography

Alkan, C., Sajjadian, S. & Eichler E. E. (2011) Limitations of next-generation sequence assembly. Nature Methods 8/1, 61-64

Andrews, Simon. "FastQC: a quality control tool for high throughput sequence data." (2010).Altschul, Stephen F., et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic acids research* 25.17 (1997): 3389-3402.

Anmarkrud, Jarl A., and Jan T. Lifjeld. "Complete mitochondrial genomes of eleven extinct or possibly extinct bird species." *Molecular ecology resources* 17.2 (2017): 334-341.

Baker, Monya. "De novo genome assembly: what every biologist should know." *Nature Methods* (2012): 333.

Bankevich, Anton, et al. "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." *Journal of computational biology* 19.5 (2012): 455-477.

Bradnam, Keith R., et al. "Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species." *GigaScience* 2.1 (2013): 10.

Capelo, Jorge, et al. "The vegetation of Madeira Island (Portugal). A brief overview and excursion guide." *Quercetea* 7 (2005): 95-122.

Castro, José A., Antònia Picornell, and Misericòrdia Ramon. "Mitochondrial DNA: a tool for populational genetics studies." *International Microbiology* 1.4 (1998): 327-332.

Chevreux, Bastien, Thomas Wetter, and Sándor Suhai. "Genome sequence assembly using trace signals and additional sequence information." *German conference on bioinformatics*. Vol. 99. 1999.D

Degtjareva, G. V., et al. "Organization of chloroplast psbA-trnH intergenic spacer in dicotyledonous angiosperms of the family Umbelliferae." *Biochemistry (Moscow)* 77.9 (2012): 1056-1064.

Dierckxsens, Nicolas, Patrick Mardulyn, and Guillaume Smits. "NOVOplasty: de novo assembly of organelle genomes from whole genome data." *Nucleic acids research* 45.4 (2016): e18-e18.

Dohm, Juliane C., et al. "The genome of the recently domesticated crop plant sugar beet (Beta vulgaris)." *Nature* 505.7484 (2014): 546.Earl, Dent, et al. "Assemblathon 1: a competitive assessment of de novo short read assembly methods." *Genome research* 21.12 (2011): 2224-2241.

Dong, Wenpan, et al. "Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding." *PloS one* 7.4 (2012): e35071.

Fuentes-Pardo, Angela P., and Daniel E. Ruzzante. "Whole-genome sequencing approaches for conservation biology: advantages, limitations, and practical recommendations." *Molecular ecology* (2017).

Galtier, Nicolas. "The intriguing evolutionary dynamics of plant mitochondrial DNA." *BMC biology* 9.1 (2011): 61.Gibbs, Richard A., et al. "Evolutionary and biomedical insights from the rhesus macaque genome." *science* 316.5822 (2007): 222-234.

Gillham, Nicholas W., and John E. Boynton. "The sequence of the chloroplast atpB gene and its flanking regions in Chlamydomonas reinhardtii." *Gene* 44.1 (1986): 17-28.

Hebert, P. D., Ratnasingham, S., & de Waard, J. R. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences*, *270*(Suppl 1), S96-S99.

Hodkinson, Brendan P., and Elizabeth A. Grice. "Next-generation sequencing: a review of technologies and tools for wound microbiome research." *Advances in wound care* 4.1 (2015): 50-58.

Kress, W. John, et al. "Use of DNA barcodes to identify flowering plants." *Proceedings of the National Academy of Sciences of the United States of America* 102.23 (2005): 8369-8374.

Kress, W. J., & Erickson, D. L. (2007). A two-locus global DNA barcode for land plants: the coding rbcL gene complements the non-coding trnH-psbA spacer region. *PLoS one*, *2*(6), e508.

Kress, W. John, et al. "Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama." *Proceedings of the National Academy of Sciences* 106.44 (2009): 18621-18626.

Li, Heng, et al. "The sequence alignment/map format and samtools." *Bioinformatics* 25.16 (2009): 2078-2079.

Martin, Marcel. "Cutadapt removes adapter sequences from high-throughput sequencing reads." *EMBnet. journal* 17.1 (2011): pp-10.

McBride, Heidi M., Margaret Neuspiel, and Sylwia Wasiak. "Mitochondria: more than just a powerhouse." *Current biology* 16.14 (2006): R551-R560.

Miller, Jason R., Sergey Koren, and Granger Sutton. "Assembly algorithms for next-generation sequencing data." *Genomics* 95.6 (2010): 315-327.

Ming, Ray, et al. "The pineapple genome and the evolution of CAM photosynthesis." *Nature genetics* 47.12 (2015): 1435.

Myburg, Alexander A., et al. "The genome of Eucalyptus grandis." *Nature* 510.7505 (2014): 356.

Neyland, Ray, and Lowell E. Urbatsch. "The ndhF chloroplast gene detected in all vascular plant divisions." *Planta* 200.2 (1996): 273-277.

Palmer, Jeffrey D. "Chloroplast DNA and molecular phylogeny." *Bioessays* 2.6 (1985): 263-267.

Patterson, Thomas B., and Thomas J. Givnish. "Phylogeny, concerted convergence, and phylogenetic niche conservatism in the core Liliales: insights from rbc L and ndh F sequence data." *Evolution* 56.2 (2002): 233-252.

Pinheiro De Carvalho, Miguel Ângelo Almeida, et al. "A review of the genus Semele (Ruscaceae) systematics in Madeira." *Botanical Journal of the Linnean Society* 146.4 (2004): 483-497.

Puppo, Pamela, Manuel Curto, and Harald Meimberg. "Genetic structure of Micromeria (Lamiaceae) in Tenerife, the imprint of geological history and hybridization on within-island diversification." *Ecology and evolution* 6.11 (2016): 3443-3460.

Quinlan, Aaron R., and Ira M. Hall. "bedtools: a flexible suite of utilities for comparing genomic features." *Bioinformatics* 26.6 (2010): 841-842.

Schatz, Michael C., Arthur L. Delcher, and Steven L. Salzberg. "Assembly of large genomes using second-generation sequencing." *Genome research* 20.9 (2010): 1165-1173.

Schuler, G. D., et al. "A gene map of the human genome." *Science* 274.5287 (1996): 540-546.

Selvaraj D, Sarma RK, Sathishkumar R. Phylogenetic analysis of chloroplast *matK* gene from Zingiberaceae for plant DNA barcoding. *Bioinformation*. 2008;3(1):24-27.

Sims, David, et al. "Sequencing depth and coverage: key considerations in genomic analyses." *Nature Reviews Genetics* 15.2 (2014): 121.

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., et al. (2017). GeSeq - versatile and accurate annotation of organelle genomes. *NUCLEIC ACIDS RESEARCH, 45*(W1), W6-W11. doi:10.1093/nar/gkx391.

Turmel, Monique, Christian Otis, and Claude Lemieux. "The complete chloroplast DNA sequence of the green alga Nephroselmis olivacea: insights into the architecture of ancestral chloroplast genomes." *Proceedings of the National Academy of Sciences* 96.18 (1999): 10248-10253.

Yagi, Yusuke, and Takashi Shiina. "Recent advances in the study of chloroplast gene expression and its evolution." *Frontiers in plant science* 5 (2014): 61.

Zhang, Jiajie, et al. "PEAR: a fast and accurate Illumina Paired-End reAd mergeR." *Bioinformatics* 30.5 (2013): 614-620.

Zhao, Yongbing, et al. "The complete chloroplast genome provides insight into the evolution and polymorphism of Panax ginseng." *Frontiers in plant science* 5 (2015): 696.