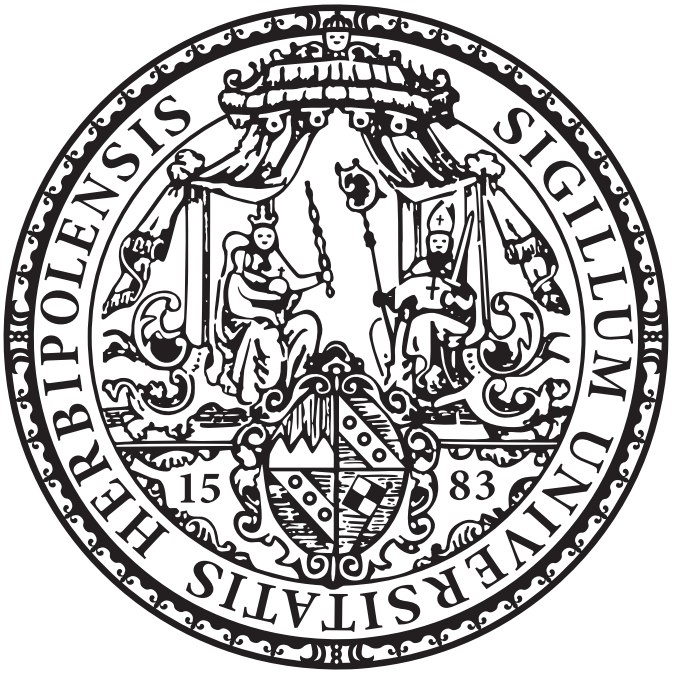
Genomics of carnivorous Droseraceae and Transcriptomics of Tobacco pollination as case studies for neofunctionalisation of plant defence mechanisms

–

Genomik karnivorer Droseraceae und Transkriptomik der Befruchtung von Tabak als Fallstudien zur Umfunktionierung pflanzlicher Verteidigungsmechanismen



Dissertation zur Erlangung des naturwissenschaftlichen Doktorgrades
der Graduate School of Life Sciences, Julius-Maximilians-Universität Würzburg,
Section Integrative Biology

vorgelegt von

Niklas Terhoeven

aus
Kempen

Würzburg, 2019

Eingereicht am _____

(Bürostempel)

Mitglieder des Promotionskomitees:

Vorsitzender:	Prof. Dr. Christian Janzen
1. Betreuer:	Prof. Dr. Jörg Schultz
2. Betreuer:	Prof. Dr. Rainer Hedrich
3. Betreuer:	Prof. Dr. Dirk Becker

Tag des Promotionskolloquiums: _____

Doktorurkunden ausgehändigt am: _____

Contents

Summary v

Zusammenfassung vi

I. Introduction 1

1. Big Data in Biology 2
1.1. From Biology to Data Science . . . 2
1.2. Genome Sequencing and Comparative Genomics . . . 2
1.3. RNA-Sequencing and Transcriptomics . . . 3
1.4. Repetitive Elements . . . 4

2. Plants used in this Study 8
2.1. Introduction to Carnivory . . . 8
2.2. The Venus Flytrap . . . 9
2.3. The waterwheel Plant . . . 11
2.4. The Sundew . . . 12
2.5. Pollination in Tobacco . . . 13
2.6. Plant Defence Mechanisms . . . 14

3. Objectives of this Study 18

II. Genomics of Carnivory in Plants 19

4. Materials and Methods 20
4.1. Software and Data . . . 20
4.2. Genome Assemblies . . . 20
4.3. Annotation . . . 20
4.4. Functional Annotation . . . 21
4.5. Repeat Annotation . . . 21
4.6. Detection of Centromeres . . . 22
4.7. Heterozygosity . . . 22

5. Results 24
5.1. The *Aldrovanda vesiculosa* Genome Assembly . . . 24
5.2. Annotations . . . 25

5.3. The Centromeres . . . 26
5.4. Repetitive Elements . . . 26
5.5. Heterozygosity . . . 27
5.6. GO-enrichments of carnivore specific-genes . . . 27
5.7. Improving the *Dionaea muscipula* Transcriptome . . . 28

6. Discussion **33**
6.1. An Overview of the Genomes . . . 33
6.2. Repetitive Elements . . . 33
6.3. GO enrichment . . . 34
6.4. Heterozygosity . . . 35
6.5. Differences in Genome Size . . . 35

III. Transcriptomics of Pollen Tube Guidance **37**

7. Methods **38**
7.1. estimating sequencing needs . . . 38
7.2. RNASeq analysis . . . 38

8. Results and Discussion **41**
8.1. Estimating sequencing needs . . . 41
8.2. First Sequencing . . . 41
8.3. Second Sequencing . . . 42
8.4. Downstream Analyses . . . 42

IV. Reper: Genome-wide identification, classification and quantification of repetitive elements without an assembled genome **49**

9. Implementing a kmer based repeat analysis workflow **50**

V. Conclusion **53**

10. Biological Insights and Lessons learned **54**

VI. Appendix **58**

Supplemental Materials **59**
1. Evolution of Carnivory in Plants 59
   1.1. Sequencing Libraries 59
   1.2. Centromere Candidates 62
   1.3. GO enrichment Plots 63
2. Transcriptomics of Pollen Tube Guidance 65
   2.1. Sequencing Libraries 65

List of Figures **66**

List of Tables **67**

List of Source Code **68**

References **69**

Acknowledgement **82**

Publication List **83**

Curriculum Vitæ **84**

Affidavit / Eidesstattliche Erklärung **85**

Summary

Plants have evolved many mechanisms to defend against herbivores and pathogens. In many cases, these mechanisms took other duties. One example of such a neofunctionalisation would be carnivory. Carnivory evolved from the defence against herbivores. Instead of repelling the predator with a bitter taste, the plant kills it and absorbs its nutrients. A second example can be found in the pollination process. Many of the genes involved here were originally part of defence mechanisms against pathogens.

In this thesis, I study these two examples on a genomic and transcriptomic level. The first project, Genomics of carnivorous Droseraceae, aims at obtaining annotated genome sequences of three carnivorous plants. I assembled the genome of *Aldrovanda vesiculosa*, annotated those of *A. vesiculosa*, *Drosera spatulata* and *Dionaea muscipula* and compared their genomic contents. Because of the high repetitiveness of the *D. muscipula* genome, I also developed reper, an assembly free method for detection, classification and quantification of repeats. With that method, we were able to study the repeats without the need of incorporating them into a genome assembly.

The second large project investigates the role of DEFL (defensin-like) genes in pollen tube guidance in tobacco flowers. We sequenced the transcriptome of the SR1 strain in different stages of the pollination process. I assembled and annotated the transcriptome and searched for differentially expressed genes. We also used a method based on Hidden-Markov-Models (HMM) to find DEFLs, which I then analysed regarding their expression during the different stages of fertilisation.

In total, this thesis results in annotated genome assemblies of three carnivorous Droseraceae, which are used as a foundation for various analyses investigating the roots of carnivory, insights into the role of DEFLs on a transcriptomic level in tobacco pollination and a new method for repeat identification in complex genomes.

Zusammenfassung

Im laufe der Evolution haben Pflanzen viele Methoden entwickelt, um sich gegen Fressfeinde und Pathogene zu verteidgen. Viele dieser Methoden wurden im Laufe der Zeit umfunktioniert. Ein Beispiel hierfür ist die Karnivorie, welche aus der Verteidigung gegen Fressfeinde entstanden ist. Anstelle einen Angreifer durch bitteren Geschmack zu vertreiben, tötet die Pflanze das Tier und nimmt seine Nährstoffe auf. Ein weiteres Beispiel ist der Bestäubungs- und Befruchtungsprozess. Viele der Gene, die hier involviert sind, stammen ursprünglich aus Mechanismen zur Verteidigung gegen Pathogene.

In dieser Arbeit untersuche ich diese beiden Beispiele auf genomischer und transkriptomischer Ebene. Die Zielsetzung des ersten Projekts, Genomik von karnivoren Droseraceaen, ist es, assemblierte und annotierte Genome von drei karnivoren Pflanzen zu generieren. Ich habe dazu das Genom von *Aldrovanda vesiculosa* assembliert und dieses, sowie die Genome von *Drosera spatulata* und *Dionaea muscipula* annotiert und miteinander verglichen. Aufgrund des hohen Anteils repetitiver Elemente im *D. muscipula* Genom habe ich reper, eine Methode zum Detektieren, Klassifizieren und Quantifizieren von Repeats, entwickelt. Mit dieser Methode ist es nun möglich, repetitive Elemente zu untersuchen, ohne diese in einem Genomassembly integrieren zu müssen.

Das zweite große Projekt untersucht die Rolle von DEFL (defensin-like) Genen im Pollenschlauchwachstum in Tabakblüten. Dazu haben wir das Transkriptom der SR1 Variante zu verschiedenen Zeitpunkten im Befruchtungsprozess sequenziert. Ich habe dieses Transkriptom assembliert und annotiert und darin nach differentiell exprimierten Genen gesucht. Zudem haben wir mit einer auf Hidden Markov Modellen (HMM) basierten Methode nach DEFL Genen gesucht und ich habe die Expression dieser in den verschiedenen Stadien untersucht.

Zusammenfassend beinhalten die Ergebnisse dieser Thesis annotierte Genomassemblies von drei karnivoren Droseraceaen, Erkenntnisse über die Rolle von DEFL Genen bei der Befruchtung auf einer transkriptomischen Ebene und eine neue Software zur Analyse von repetitiven Elementen in komplexen Genomen.

Part I.

Introduction

1. Big Data in Biology

1.1. From Biology to Data Science

For a long time, calculations have been a big part of biology. For example, Gregor Mendel calculated hybridisation patterns of peas in 1865 (Mendel 1866). With the invention of computers, scientists started to use them to answer biological questions. One prominent example is Margaret Oakley Dayhoff. During the 1960s she developed computational methods to compare protein sequences and infer their phylogeny (Dayhoff and Eck 1966). Together with her colleagues Richard Eck and Robert Ledley, Dayhoff is now seen as one of the founders of bioinformatics.

As computers became much cheaper and much more powerful during the last decades, their importance and use cases in biology increased as well. While Dayhoff was able to analyse the sequences of a few proteins in the 1960s, the 1001 Genomes Consortium could analyse the genomes of 1135 *Arabidopsis thaliana* plants in 2016 (Alonso-Blanco et al. 2016). High throughput analyses like this show the possibilities that arise from combining biology with data science. Not only large genomic studies are becoming feasible, but also other biological disciplines are relying on computational methods. A good example of this is the newly initiated Center for Computational and Theoretical Biology at the University of Würzburg. Here, researchers from very different biological fields, from molecular biology to ecology, are working together with similar computational methods to answer their own specific questions. Since my work is about genomics and transcriptomics, the following sections focus on the opportunities and challenges that arise in these particular fields.

1.2. Genome Sequencing and Comparative Genomics

Similar to the development of computers, the cost and methodology of DNA sequencing improved rapidly during the last decade (NHGRI 2016). For decades, Sanger chain-termination sequencing was the most widely used method to identify the bases in a DNA string. This method was developed by Frederick Sanger and colleagues in 1977 and provided an easy way to sequence DNA fragments of a few hundred basepairs length (Sanger et al. 1977). Such a sequenced fragment is called read. By today's standards, Sanger sequencing is still one of the most accurate methods and therefore still used as the gold standard, but it is time and cost intensive compared to other methods. Next-Generation-Methods, such as Illumina dye sequencing, are much wider used today.

Illumina sequencing machines can sequence a DNA fragment from both sides. This process is called paired-end sequencing and allows to sequence longer fragments, as the

length of the gap between the two reads is known. Also, a negative gap size is possible which is used in the so-called overlap sequencing. Newer Illumina machines can sequence paired reads up to 300 bp long with an accuracy of 99.9 % in 56 hours resulting in a total of up to 15 Gb of data (Manufactures specification for the MiSeq system with the Reagent Kit v3).

However, assembling a complete genome from 300 bp long reads requires complex algorithm and large computing resources. To solve this, third-generation sequencing technologies were invented. These technologies aim for longer reads but have to sacrifice some accuracy. To compensate the high error rates, a given fragment is sequenced multiple times and a consensus is built for each base pair. With this method, the so-called consensus accuracy is in the same range as the Illumina technology. The Single Molecule Real Time (SMRT) technology by Pacific Bioscience (PacBio), for example, yields reads up to 300 Kbp with a consensus accuracy of 99.999 % (according to the manufactures specification for the Sequel System 6.0). Another technology known for extremely long reads is Oxford Nanopore. Here, the maximum read length is not limited by the sequencing technology. That leads to reads as long as the fragments in the sample. Recently the first sequencing of a 200 Mbp read was reported (Payne et al. 2018).

As modern sequencing technologies are becoming better and cheaper, more complex studies are feasible. There are studies involving extremely large genomes, like the 20 Gbp Norway Spruce (Nystedt et al. 2013) or the 32 Gbp axolotl (Nowoshilow et al. 2018). Other studies involve very complex genomes, for example the allohexaploid wheat (Montenegro et al. 2017, Clavijo et al. 2017) or many smaller genomes like the 1001 *A. thaliana* genome project (Alonso-Blanco et al. 2016).

Sequencing a single genome can help to understand a lot about the organism. The presence or absence of certain genes, for example, can give hints about the evolutionary origin and how phenotypic features are manifested on a molecular level. More value can be drawn from a genome if it is compared with others. Depending on the phylogenetic distance between the compared organisms, different results can be achieved. If very closely related organisms are compared, for example different strains of a single species, the comparison can show differences on a single basepair scale. This allows to pinpoint phenotypic differences to single mutations and, for example, explain the different flowering times in *A. thaliana* plants (Alonso-Blanco et al. 2016). On the other hand, genomes of a wide range of plants can be compared to understand the "Evolution of Flowering Plants" (Amborella Genome Project et al. 2013).

1.3. RNA-Sequencing and Transcriptomics

One question that cannot be answered by genome sequencing is whether a gene is actually expressed. To answer this, we need a transcriptomic approach called RNA-Sequencing (RNASeq). RNASeq is a technology similar to genome sequencing, but using RNA instead of DNA. In RNASeq experiments, the whole RNA from a cell or a tissue is extracted, reverse transcribed to cDNA and then sequenced using the same techniques, mostly Illumina, as used in genome sequencing. In contrast to DNA sequencing, RNASeq

shows only those genes, that are currently expressed. This allows analysing expression profiles of different tissues, different treatments or different developmental stages. For example, Bemm et al. (2016) compared the gene expressions in different tissues of the Venus Flytrap, Fracasso et al. (2016) analysed drought stress in Sorghum, and Sierro et al. (2014) studied differences between young, adult and senescent tobacco flowers.

Another difference to genome sequencing is the concept of coverage, or sequencing depth. While the coverage is used to ensure that all parts of the genome have been sequenced in genome sequencing, it is used to quantify the actual expression levels in transcriptomic analyses. Here, the reads corresponding to a certain gene are counted and then compared between genes. This allows an estimate of how many copies of a gene are present in the sample and from this infer significant differences between samples. This so-called a differential expression analysis is the main result of many transcriptomic studies. Together with the knowledge of how the samples relate to each other, these results can be interpreted to gain insights into the genes involved in the studied process.

Regarding the computational methods, two main strategies can be used: A *de-novo* and a reference-based approach. For the reference based approach, an annotated genome assembly is needed. The RNASeq reads can be mapped to this assembly and the reads mapping to regions of annotated genes can be counted and compared. A popular software package used for this approach is cufflinks (Trapnell et al. 2012). The *de-novo* approach starts with a creating a transcriptome assembly. Here, the RNASeq reads are assembled into transcript sequences, the so-called isoforms. Often multiple isoforms represent one gene. This can have biological reasons like multiple copies of a gene or splice variants or it can be caused by assembly artefacts. Such a group of isoforms is called unigene. In the second step, the RNASeq reads are mapped to the isoforms and the reads mapping to each isoform are counted and analysed in a similar way as in the reference-based approach. The high number of isoforms obtained during the *de-novo* assembly represents a disadvantage compared to the reference-based approach. Usually, a few hundred thousand isoforms are assembled in this process, which is about ten-fold higher than a usual gene count of 20 to 30 thousand genes. This not only increases the need for computational resources but also influences the analysis as many of these isoforms are artefacts with no biological importance. To tackle this, a combination of both approaches can be used. The transcriptome assembler Trinity (Grabherr et al. 2011) can be run in a so-called genome-guided mode. In this mode, a reference genome of a different strain or closely related organism can be used in the assembly which leads to a smaller number of isoforms and with that, fewer artefacts in the assembly.

1.4. Repetitive Elements

The Need for a new software

While working on the genome assembly of *D. muscipula*, Thomas Hackl suspected repeats to be one of the main causes of fragmentation. However, there are genomes published of other repeat-rich plants, for example, tobacco (70 % repeats, Sierro et al. 2014), wheat (80 % repeats, Clavijo et al. 2017) and *Capsicum annuum* (81 % repeats,

Qin et al. 2014). Therefore we decided to identify the repeats in *D. muscipula* to see if they are different from the repeats of other plants. I evaluated different methods for this task in my Master Thesis (Terhoeven 2014) and reached the conclusion that a method using kmer counting combined with a Trinity based repeat assembly should yield the best results. This approach is now implemented in reper. A similar approach is implemented in a software called dnaPipeTE (Goubert et al. 2015). However, it is difficult to install and depends on some non-free software like trf (Benson 1999) and the giri repbase libraries (Bao et al. 2015). In contrast, the reper package does not have any non-free dependencies and can be installed easily via docker.

Structure and Groups of Repeats

First, we have to get an idea of what repeats actually are. Repetitive elements, or repeats, are DNA sequences that have multiple occurrences in a genome. This is not entirely true, as some sequences occur only once and are still classified as repeat because of their evolutionary origin. In general, repeats are divided into two groups, tandem and dispersed repeats. Tandem repeats consist of a motif that is repeated several times consecutively (see fig. 1.1a). An example of tandem repeats are centromeres.

The second group, the dispersed repeats, is also called transposons (TE - transposable element). This name indicates their behaviour of "jumping around in the genome", which was discovered by Barbara McClintock in the 1940s and awarded with the Nobel Prize in 1983. The transposition can be achieved via DNA or RNA. RNA mediated TEs (also called class I) are copied in two stages. First, the DNA is transcribed to RNA, then the RNA is reverse transcribed to DNA and inserted in a new position in. The class II elements (DNA mediated) do not need the RNA transcription step. Instead, they are cut out and inserted directly using an enzyme called transposase. This process relies on the structural features of TEs. DNA mediated TEs have an inverted repeat and a so-called target-site duplication (see fig. 1.1b). A target-site duplication is also present in the class I TEs. Within this group, one can distinguish between transposons that have a long-terminal-repeat (LTR) and those without (non-LTR transposons). The most prominent examples of the latter group are long interspersed nuclear elements (LINE) and short interspersed nuclear elements (SINE). An overview of the structures is given in fig. 1.1.

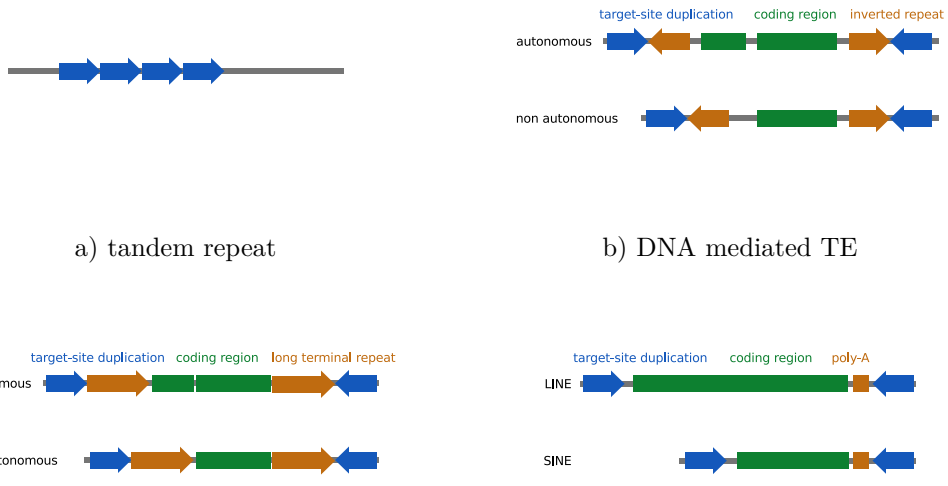
Both, class I and class II TEs can encode for the proteins they rely on during transposition. In contrast to those autonomous elements, the protein-coding regions can be mutated or degenerated, which makes them non-functional and the TE has to rely on an autonomous element for transposing (see fig. 1.1b and fig. 1.1c)

Identification of Repeats

The classical approach to identify repeats relies on their structure and homology. An example workflow is given by Campbell et al. (2014) (`http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction-Advanced`). In this example, tools like ltrharvest (Ellinghaus et al. 2008) are used to identify structures typical

for LTR-retrotransposons. Also the RepeatMasker (Smit et al. 2013) and RepeatModeler (Smit and Hubley 2008) software packages are used here. They search for repeats by screening a library of known repeats for similar sequences. The approach outlined in the tutorial works well on an assembled genome of a model organism. When working with non-model organisms, repeat detection gets more complicated mainly because of two reasons. There may be no high-quality genome assembly available. This makes it difficult to search for structures as the expected lengths may exceed the average contig length. The second limitation of this approach is also based on the used of a genome assembly. Most assembly algorithms cannot correctly resolve repetitive sequences. This leads to collapsing repeats in the assembly, which makes it difficult to determine how many instances of a repeat are present.

The identification approach I developed here, uses a different feature of repeats, namely their repetitiveness. Using the raw sequencing data, a kmer analysis can find overrepresented sequences. These sequences can then be assembled to form a set of repeats. Additionally, this approach allows for a quantification of the repeats by mapping and counting the reads on the repeats. To classify the identified repeats, a reference library of known repeats is used. A detailed description of the process is given in chapter 9.



a) tandem repeat

b) DNA mediated TE

c) LTR Retrotransposon

d) non-LTR retrotransposon

Figure 1.1.: **Structures of different repeat types.** Fig. a) shows the typical structure of a tandem repeat. A motive that is repeated several times consecutively (blue). b) shows the organization of a DNA-mediated class II transposable element. It consists of a target-site duplication (blue), an inverted repeat (brown) and a coding region (green). The non-autonomous element (lower) has a shortened coding region. The LTR retrotransposon (c)) looks very similar. It contains a long terminal repeat (brown) and also occurs as autonomous and non-autonomous elements. Fig. d) shows examples for LINE and SINE, the two main groups of non-LTR transposons. They consist of a target-site duplication, a coding region and a poly-A sequence.

2. Plants used in this Study

2.1. Introduction to Carnivory

It is a concept that seems unusual and counter-intuitive on the first glance. However, it is known and subject to scientific studies for centuries. In 1875, Charles Darwin published his book "Insectivorous Plants", in which he described the insect catching and digesting behaviour of several plants.

To catch their prey, carnivorous plants developed different trapping strategies. Popular examples are the vacuum driven bladder traps found in the *Utricularia* genus, pitcher traps, examples of which can be found in the families Nepenthaceae, Cephalotaceae or Sarraceniaceae, the sticky traps of *Drosera* and snap traps used by *Dionaea muscipula* and *Aldrovanda vesiculosa* (see fig. 2.5). During the last years, several studies gained insights into these unusual plant organs. For example, the pitcher traps of different *Nepenthes* species show a great variation in morphology and composition of digestive fluids to accommodate specific niches and prey preferences (Gaume et al. 2016). These fluids also serve as habitat for several insects, bacteria and algae (Adlassnig et al. 2011). *Cephalotus follicularis* switches from flat leaves to traps depending on the temperature (Fukushima et al. 2017). And some of the sticky flypaper traps found in *Drosera* show movement towards the prey. This enclosure of the prey allows catching larger animals and thus led to the evolution of the snap traps found in *D. muscipula* and *A. vesiculosa* (Gibson and Waller 2009, Poppinga et al. 2013).

While the trapping of the prey is achieved very differently, the digestion is similar. A cocktail of enzymes is secreted to dissolve the chitin shell and make the nutrients available to the plant (Schmeil 1911).

As the taxa mentioned suggest, carnivorous plants can be found all over the tree of life (Albert et al. 1992) and the carnivorous lifestyle has evolved independently (Darwin 1875). The reason for this may be found in the natural habitat of these plants. Most carnivores live in swamps, muddy, sandy shores or in the water (Heubl et al. 2006). These habitats cannot offer many nutrients. Especially Nitrogen is rare. However, carnivores can overcome this issue by taking up additional nutrients from their prey.

The aquatic *Utricularia gibba* was the first carnivorous plant to have its genome sequenced. Despite *U. gibba* being octaploid and having undergone three whole genome duplications, the genome is very small (77 Mbp). It contains about 3 % repeats and 28500 genes. *U. gibba* does not have roots and many genes associated with root development are missing from the genome (Ibarra-Laclette et al. 2013). A recent resequencing, using PacBio technology, showed that the last WGD event may have been an allopolyploidization and that many tandem-duplicated genes are the main factor of the adaption to carnivory in *U. gibba* (Lan et al. 2017).

In the same year, the genome of *Genlisea aurea* was published by Leushkin et al. (2013). With 63.6 Mbp the genome of *G. aurea* is even smaller than the genome of *U. gibba*. Due to the sub-optimal quality of the Illumina assembly, covering only 68 % of the genome with over 10000 contigs and an N50 value of 5800 bp, no reliable conclusions about the presence and absence of certain genes can be drawn.

The first Droseraceae genome sequenced was the one of *Drosera capensis*. In the 293 Mbp genome 8120 genes were identified containing several new proteases (Butts et al. 2016).

In 2017, the first genome of a pitcher plant, *C. follicularis* was published. The Illumina/PacBio hybrid assembly spans 76 % of the 2.11 Gbp genome and contains 36500 genes. The genome showed a lineage-specific expansion of digestion related genes, which produce digestive fluids similar to those found in Droseraceae (Fukushima et al. 2017).

In this study, the genomes of the Droseraceae *Aldrovanda vesiculosa*, *Dionaea muscipula* and *Drosera spatulata* are analysed and compared to gain insights into the evolution of these unusual plants.

2.2. The Venus Flytrap

The Venus Flytrap, *Dionaea muscipula* (SOL) J.Ellis, is one of the most famous carnivorous plants. It was first mentioned in 1760 by Arthur Dobbs and has been described in 1769 by John Ellis (Ellis 1770). While Ellis found hints of carnivory, Carl von Linné dismissed this theory, because it contradicts the Bible. About 100 years later, in 1875, Charles Darwin published his book “Insectivorous Plants” in which he proved the carnivorous lifestyle of the Venus Flytrap (Darwin 1875). Since then, *D. muscipula* has been subject to various scientific studies and breeding programs.

The wild habitat of *D. muscipula* consists of swamps in North and South Carolina (within a 100 Km radius of Wilmington, North Carolina). Because of its very small natural habitat, the Venus Flytrap is considered as "vulnerable" on the IUCN Red List of Threatened Species (IUCN 2000).

The morphology of the Venus Flytrap is described by Ellis (1770). The plant is rather small with a high stem and a white flower on the top (see fig. 2.1). The high stem and the colour difference between trap and flower are probably the reason for the fact, that *D. muscipula* does not prey on its pollinators (Youngsteadt et al. 2018). The most unusual morphological feature is the snap trap.

Figure 2.1.: **Drawing of the Venus Flytrap** (*D. muscipula*) by William Curtis (1746-1799). Public Domain

These traps consist of two parts and can close rapidly to trap prey. The closure takes only 10 ms and is one of the fastest movements observed in plants (Forterre et al. 2005). The inner parts of the trap are covered in red glands.

In 1769, John Ellis suggested in a letter to Linnaeus that these excrete a sweet liquor to bait insects into the trap which are then squeezed to death (Ellis 1770). The outer rim is green and builds spikes, which further keep the insect inside of the trap. On the surface of the trap, three to six trigger hairs can be found. These hairs react to contact and can trigger the trap closure and the digestion process (Darwin 1875).

The previously described morphology is not universally true. While there is a great variety in natural shapes and colours, commercial breeding of the Venus Flytrap created more than 30 registered cultivars (see examples in fig. 2.6).

When a Venus Flytrap needs nutrients, its traps turn red and excretes olfactory substances to lure prey. This process is similar to the attraction of pollinators in flowers. When an insect lands on the trap and touches the trigger hairs, action potentials (AP) are produced. If two APs are triggered within 30 s, the trap closes rapidly and locking the prey inside. Additionally, the jasmonate signalling pathway is activated. When the animal now tries to escape, further APs are generated leading the plant to continue the digestion process. In this phase, the trap seals tightly and enzymes are excreted to digest the captured prey. The released nutrients are then taken up via transporters in the glands. This whole process takes up to five days and after the digestion, the trap is opened again for further hunting. The two-step AP triggering, together with and additional chemosensing, is necessary to detect false positive initial triggers (Darwin 1875, Hodick and Sievers 1989, Bemm et al. 2016, Böhm et al. 2016, Hedrich and Neher 2018).

While scientists acquired a lot of knowledge about the Venus Flytrap and its insectivorous lifestyle, there is not much known about its genome. *D. muscipula* is diploid with 32 chromosomes (Shirakawa et al. 2011). The genome size, however, is still discussed. Shirakawa et al. (2011) report 706 Mbp, Jensen et al. (2015) report 2956 +/- 210 Mbp and Veleba et al. (2017) report 2.85 Gbp. Thomas Hackl estimated the genome size between 2.5 Gbp and 2.8 Gbp and reported that the large genome size is mostly caused by the presence of repetitive elements (Hackl 2016).

Some genes found in the transcriptome of *D. muscipula* show similarities to genes of the grapevine and tomato (Jensen et al. 2015).

A second transcriptomic study compared the gene expression of different tissues and gained several insights into the molecular mechanism of carnivory. The gene expression profiles of activated traps and glands are very similar to the profiles of roots, which is probably caused by the nutrient uptake functionality of both tissues. Many of the genes involved in the process of detecting, catching and digesting the prey are originated from defence mechanisms. For example, genes involved in Chitin sensing are used to detect herbivores in other plants and potential prey in *D. muscipula* (Bemm et al. 2016).

2.3. The waterwheel Plant

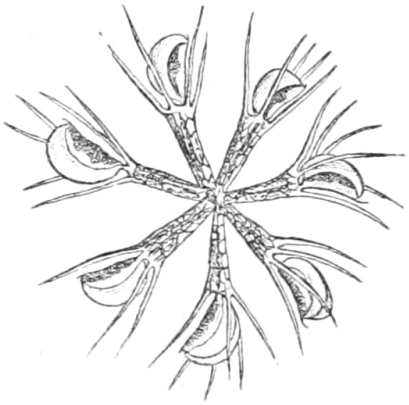


Figure 2.2.: **Drawing of the Waterwheel plant** (*A. vesiculosa*) by Ferdinand Julius Cohn. Public Domain

The waterwheel plant, *Aldrovanda vesiculosa*, is the closest living relative to *D. muscipula*. Darwin described it as a "miniature aquatic Dionaea" (Darwin 1875). On the first glance, this comes very close, as the waterwheel plant is small, has traps similar to the snap traps of the Venus Flytrap and lives in ponds. However, the actual functionality of the trap is quite different (Poppinga et al. 2013, Westermeier et al. 2018).

The name is derived from its morphological appearance, which consists of multiple "trap-wheels". These wheels contain several petioles with traps at the outer ends and are distributed evenly along a central stem (see fig. 2.2 and fig. 2.5d). An adult *A. vesiculosa* does not have any root (Adamec 2000). A second unusual morphology caused by the aquatic lifestyle is the lack of stomata. There are only a few very small stomata found on the abaxial surface of sepal leaves (Zaman et al. 2011).

In contrast to *D. muscipula*, the waterwheel is distributed all over the world. There are populations in Northern Europe, Africa, Asia and Australia (IUCN 2012). Despite the large distribution, the worldwide population of *A. vesiculosa* is rather small and decreasing. Therefore, the IUCN classified the waterwheel as "endangered" (IUCN 2012). While there have been several species of *Aldrovanda*, all but one are extinct, making *Aldrovanda* a monotypic genus (Degreef 1997).

The genome of *A. vesiculosa* is much smaller than the genome of *D. muscipula*. It is diploid and comprises 469 Mbp (Veleba et al. 2017) distributed over 48 chromosomes (Shirakawa et al. 2011). It has been suggested, that in contrast to the Venus Flytrap, *A. vesiculosa* does not show much variation in its genome (Hoshi et al. 2006).

2.4. The Sundew

Figure 2.3.: **Example of *D. spatulata*.** The spoon shaped leaves are clearly visible. Picture by flickr user Boaz-Ng, licensed under CC-BY-NC.

Unlike *Dionaea* and *Aldrovanda*, *Drosera* is not a monotypic genus. It contains more than 100 species with a high morphological diversity. The smallest, called pygmy *Drosera*, are 15 - 20 mm wide (Brittnacher 2018), while one of the largest examples, *Drosera magnifica*, is over 1 m tall (Gonella et al. 2015). The shape of the traps ranges from long and thin (*D. capensis*) to wide flat leaves (*Drosera erythrorhiza*). The traps are mostly covered with tentacles with a drop of sticky fluid on top (see fig. 2.3). Similar to *D. muscipula* electrical and jasmonate signalling is used to detect the presence of trapped prey (Krausko et al. 2017). Many *Drosera* species show some form of movement to aid capturing and digesting prey (Darwin 1875). The pimpernel sundew (*Drosera glanduligera*) even has special tentacles on the outer part that catapult the prey into the sticky inner part of the trap (Poppinga et al. 2012). Similar to other carnivorous plants, most *Drosera* species do not prey on their pollinators. This is achieved by a clear separation of trap and flower (El-Sayed et al. 2016).

The *Drosera* species studied here is *D. spatulata*. It is one of the smaller species and consists of a flat about 5 cm wide rosette of trap leaves. During flowering time, a scape grows in the middle with small flowers on the top. The name is derived from the latin word spathulatus which means "spoon-shaped" and reflects the shape of the traps (see fig. 2.3).

When an insect lands on the trap, it is fixed by the sticky fluids. Like many other sundews, the leaves of *D. spatulata* move towards the trapped prey. An enzyme cocktail is secreted which digests the insect and the nutrient-rich fluids are absorbed by the plant (Darwin 1875).

The genome of *D. spatulata* is diploid, 293 Mbp large (Veleba et al. 2017) and consists of 40 chromosomes (Shirakawa et al. 2011). This is similar to the already published *Drosera* genome from *D. capensis*.

2.5. Pollination in Tobacco

Figure 2.4.: **Botanical drawing of a Tobacco plant** (*N. tabacum*) by Franz Eugen Köhler (1897). Public Domain

Nicotiana tabacum is one of the most well known Solanaceae. It is an annual herb growing 1-2 m tall with large leaves. All green parts of the plant are covered in sticky gland hairs to protect the plant against herbivores. The flowers are long tubes of slightly red and white colour (Schmeil 1911). *N. tabacum* is mostly grown in cultivation and used to produce consumable tobacco. The genome of *N. tabacum* is allotetraploid, about 4.5 Gbp large and contains about 70 % repeats (Sierro et al. 2014, Zimmerman and Goldberg 1977). Because of its large flowers (see fig. 2.7), tobacco is well suited to study the fertilisation process.

When a flower gets pollinated, the pollen grows a pollen tube. This tube grows through the style and turns to the ovules as soon as it reaches them. Then the pollen tube bursts and the ovule is fertilised (see fig. 2.8). During this process, a lot of cell to cell signalling is necessary, which is mainly achieved by Cysteine-Rich-Proteins (CRP) (Marshall et al. 2011). In the first step, when the pollen tube penetrates the stigma, a member of the defensin-like (DEFL) subfamily is used as a signal to detect the plants own pollen and prevent self-fertilisation (Takayama et al. 2001,J. B. Nasrallah and M. E. Nasrallah 2014). The following growth of the pollen tube through the style is also guided by CRPs (Qu et al. 2015). One important group involved here are the LURE peptides, which belong to the DEFL subfamily. LUREs attract pollen tubes of their own species and can therefore guide the growth direction (Okuda et al. 2009). When reaching the ovule, the pollen tube bursts and releases its two sperm cells. This process is also regulated by members of the DEFL subfamily (Amien et al. 2010, Bircheneder and Dresselhaus 2016). After the fertilisation, CRPs are still upregulated indicating their involvement in the further development of the embryo (Huang et al. 2015, Bircheneder and Dresselhaus 2016).

It has been suggested, that the involvement of CRPs in fertilisation evolved from defence mechanisms (Bircheneder and Dresselhaus 2016). Originally, cell to cell communication via CRPs was used to defend plants against pathogens, i.e. fungi and bacteria (Marshall et al. 2011). When pollen evolved, it was important to protect these and the early seedlings from pathogens, which lead to an increased expression of CRPs in these tissues (Bircheneder and Dresselhaus 2016). Still, these are the tissues were the most CRPs are expressed (Huang et al. 2015). Additionally, the penetration of the pollen tube and fungal hyphae are morphologically similar and therefore requires similar communication mechanisms to detect harmful fungi and the wrong pollen (Bircheneder and Dresselhaus 2016).

While CRPs are quite diverse and share only little sequence similarity, the positions of their 6 to 8 cysteines are highly conserved within subfamilies and can be used to classify them. Members of the DEFL subfamily, for example, consist of 40 to 70 amino acids

and has 8 cysteines in a specific pattern (Silverstein et al. 2007).

In this study, we want to analyse the role of these CRPs, especially the DEFLs, in the fertilisation process in *N. tabacum*. To achieve this, we run a transcriptomic analysis of ovule and pollen tube tissues during different stages of the fertilisation process: Pollen tube only, ovule only, ovule of a pollinated and ovule of a fertilised flower. This approach allows us to find differences in the expression of CRPs in the different stages. Since the published genome of *N. tabacum* is based on the TN90 strain, and the plant material available to us is from the SR1 strain, we cannot use the reference based method. However, we can use the published genome as the reference for the genome-guided Trinity approach explained in section 1.3.

2.6. Plant Defence Mechanisms

During the evolution, plants developed many strategies to cope with different threats, mainly herbivores and pathogens (Purves 2006). The first line of defence is mechanical. Many plants build strong bark or cuticles to strengthen the outside of the plant and make it more difficult for attackers to penetrate the cells. Also, spikes and thorns are used by different plants to prevent being eaten by herbivores (Purves 2006). The second line of defence is chemical. Plants can produce substances that are toxic to their attackers or cause a repellent taste. An example of such a toxic substance is nicotine produced by tobacco (Steppuhn et al. 2004). An example of the repellent taste strategy can be found in peppers. The fruits of some pepper plants contain a substance called capsaicin, which is the cause of the spicy taste. It has been shown to protect the plant against fungi (Tewksbury et al. 2008) and many herbivores, especially mammals, as they are repelled by the spicy taste of the peppers (Tewksbury and Nabhan 2001).

Sometimes the aforementioned defence mechanisms take on additional duties within the plant. For example, birds can not taste the spicy capsaicin in the fruits what leads to them being eaten only by birds which then distribute the seeds over a large area (Tewksbury and Nabhan 2001). Another example of a neofunctionalisation of defence mechanisms is carnivory (Pavlovič and Mithöfer 2019). Here the plant goes one step further and actually killing the attacker and taking up the nutrients. In many stages in the process of capturing and digesting an animal, many mechanisms originally related to defence have been found. For example, sensing the attacker and releasing chemical substances that kill and digest the animal (Bemm et al. 2016, Hedrich and Neher 2018). A third example of new duties for defence mechanisms is the pollination process of plants. During this process, many genes of the DEFL subfamily are involved (Bircheneder and Dresselhaus 2016). Defensin genes are older than the divergence of animals and plants and evolved to protect early seedlings from pathogens (Dias and Franco 2015). Since the penetration of pollen and fungi is morphologically similar, these group of proteins was repurposed during evolution to aid the pollination process (Bircheneder and Dresselhaus 2016).

a) *Nepenthes*

b) *U. aurea*

c) *D. spatulata*

d) *A. vesiculosa*

e) *D. muscipula*

Figure 2.5.: **Examples of different carnivorous plants.** The pictures a) and c) are created by flickr user incidencematrix and licensed under CC-BY. Picture b) is created by Wikimedia user Michal Rubeš and licensed under CC-BY. Picture d) is created by cpitalia wiki user Sonia-80-pi and licensed under CC-BY-NC-ND. Picture e) is created by flickr user Scott Sherrill-Mix and licensed under CC-BY-NC.

a) B52

b) Crested Petioles

c) Holland Red

d) Microdent

Figure 2.6.: **Examples of *D. muscipula* cultivars.** All Pictures by cpitalia wiki user Sonia-80-pi and licensed under CC-BY-NC-ND.

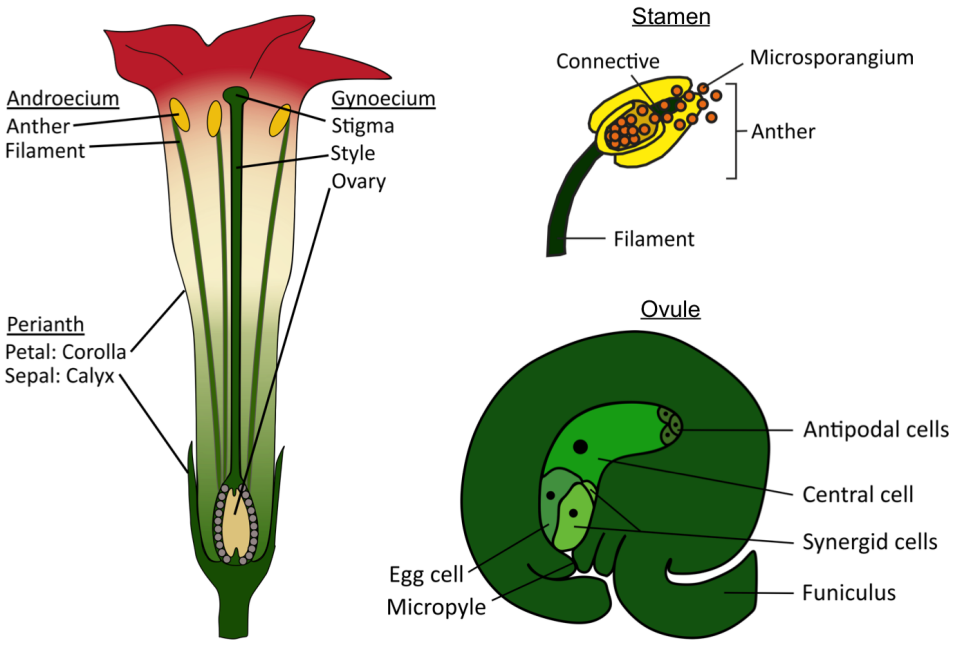


Figure 2.7.: **Flower morphology of *N. tabacum*.** Drawing by Katharina von Meyer, used with permission

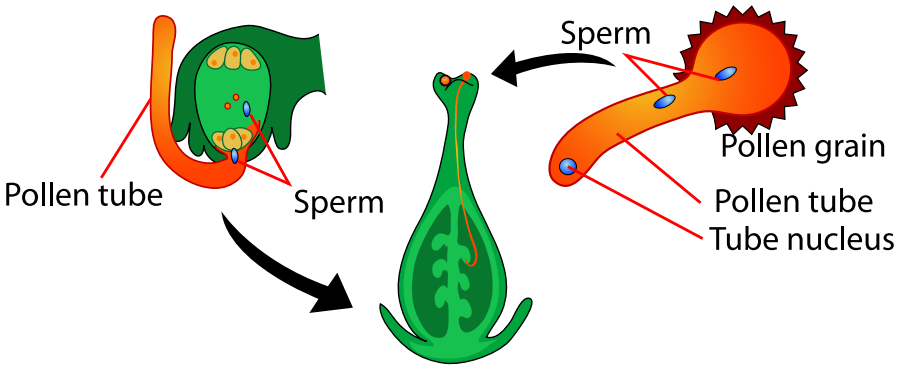


Figure 2.8.: **Fertilization Process of Angiosperms.** This drawing shows the growth of the pollen tube and how it reaches the ovule. Modified from a Drawing by Mariana Ruiz published in Wikimedia as public domain.

3. Objectives of this Study

This study is divided into three main parts, genomics of carnivorous Droseraceae, transcriptomics of pollen tube guidance in tobacco and the development of a software pipeline for repeat analysis. The first project aims at investigating the evolution of carnivory in plants based on the genome sequences of *D. muscipula*, *A. vesiculosa* and *D. spatulata*. The goal of this study is to have three annotated genome assemblies, which can then be used by my collaborators for further comparative studies. Since the *D. muscipula* genome contains a vast amount of repeats, that can not be analysed with available software, I developed the reper pipeline to address the challenges that arise in this particular scenario. The goal here was also to generalise the software and make it publicly available.

The tobacco project aims at describing the involvement of CRPs, and especially DEFLs, in the pollination process on a transcriptomic level. The first goal here is establishing the sequencing needs for our purpose. The next goal is to assemble and annotate the transcriptome, analyse the gene expression and find differentially expressed genes. The data generated here is then made available to my collaborators via TBro (Ankenbrand et al. 2016). In the third step, we want to identify CRPs in the transcriptome and focus our expression analyses on them.

In total, this study demonstrates two scenarios of neofunctionalisation of plant defence mechanisms and analyses their underlying behaviour on a genomic and transcriptomic scale.

Part II.

Genomics of Carnivory in Plants

4. Materials and Methods

4.1. Software and Data

In this project, I used 732 giga base pairs of sequencing data. For the *A. vesiculosa* genome, I used three Illumina and 23 PacBio libraries. These Illumina libraries were sequenced in June and October 2014 by LGC Genomics, the PacBio libraries in multiple runs between December 2016 and June 2017 by GATC. I also used one RNASeq library consisting of a whole adult non-flowering plant. This library was sequenced in July 2016 by LGC Genomics. For the *D. muscipula* analyses, I used four Illumina libraries, which were sequenced in August 2014 by LGC Genomics and provided by Thomas Hackl. Additionally, I used the RNASeq reads from the transcriptome project by Bemm et al. (2016). The *D. spatulata* libraries used were provided by Gergő Palfalvi and consist of three Illumina and 24 RNASeq libraries. An overview of all libraries is given in the supplemental table VI.1. I used various software for the analyses. A list is given in table 4.1.

4.2. Genome Assemblies

The *D. spatulata* assembly was provided by Gergő Pálfalvi (National Institute for Basic Biology, Okazaki, Japan). He used the same method for the assembly as I used for the assembly of *A. vesiculosa* (see section 5.1). The *D. muscipula* assembly used in this study was made by Thomas Hackl (Department of Civil and Environmental Engineering, Massachusetts Institute of Technology). He used the allpath assembler with digitally normalised Illumina data, as well as Illumina reads and artificial Illumina-like reads derived from PacBio data for scaffolding. For further refinement, he used Redundans (Pryszcz and Gabaldón 2016) and PBJelly (English et al. 2012). Details about the assembly workflow can be found in his PhD Thesis (Hackl 2016). The completeness of all three assemblies was evaluated using BUSCO version 3.0.1.

4.3. Annotation

The annotation of the three genomes was achieved using an iterative approach based on a MAKER tutorial (`http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial_for_WGS_Assembly_and_Annotation_Winter_School_2018`). This approach allows to use a variety of evidence data: I used all plant proteins in the swissprot database (downloaded on 6. Dec 2017), the proteins of Amaranth (*Amaranthus hypochondriacus*) and Quinoa (*Chenopodium quinoa*) and an augustus model of *A. thaliana*. Ad-

ditionally, I used transcriptome assemblies and individually trained snap HMMs for each plant.

The transcriptome assembly was built for each species using the cufflinks pipeline with the RNAseq reads (libraries listed in supplemental table VI.1). First, the reads were aligned to the genome assembly using star. Next, cufflinks and cuffmerge were used to analyse the mappings and to create the transcriptome assembly.

The first iteration of maker was run using the assembled transcriptome as gff file, the *A. thaliana* augustus model, as well as the plant proteins mentioned above as evidence. The resulting annotation was then used to train the snap HMMs. These HMMs were then used as evidence for the second maker iteration. Then, a second snap training was conducted using the genes from this maker run. Finally, a third maker run was started using the second snap results. The scripts used are available on GitHub (`https://github.com/Okazaki-Wuerzburg/analysis-scripts`, `doi:10.5281/zenodo.3228126`)

4.4. Functional Annotation

The functional annotation of the genes resulting from Maker was mostly based on Interpro. To assign Interpro IDs, I used Interproscan Version 5.25-64.0 with the `-goterms` and `-cpu 30` parameters and the and the maker protein fasta files as input. The resulting tsv file was then parsed to extract GO-terms for the proteins.

GO-enrichments were based on these results. To calculate them, I used ontologizer with the following parameter settings. `-n` to get the annotation output file, `-d 0.05` to set the p-value cutoff for the output plot and `-mtc "Benjamini-Hochberg"` to select the Benjamini-Hochberg multiple testing correction algorithm. The resulting GO-terms and p-values were then used as input for REVIGO to group the GO-terms and get additional graphical representations of the enrichments.

4.5. Repeat Annotation

The annotation of repetitive elements was done using different methods. First, Matthias Freund used the reper pipeline (Terhoeven et al. 2018), which I developed as part of this project, to find and quantify repeats. Reper uses kmer distributions to extract highly represented sequences, assembles and classifies them. The last step is a quantification of the detected repeats based on read mappings. For a detailed description of the software see part IV. The repeats for *A. vesiculosa* were annotated based on the 180 bp overlap library (L180), for *D. muscipula* based on the first of the LGC Illumina libraries (dm-il-01) and for *D. spatulata* based on the 130314_L4-16 library. More details about this can be found in the Master Thesis of Matthias Freund (Freund 2019). I used the resulting repeat libraries to mask the genomes using RepeatMasker.

For comparison, a second RepeatMasker based annotation was used. Here I used a workflow based on RepeatModeler to create a repeat library. The process is based on a protocol by MAKER (`http://weatherby.genetics.utah.edu/MAKER/wiki/index.`

`php/Repeat_Library_Construction-Basic`). First, an initial library is built with RepeatModeler which will be refined in the next steps. The first refining step aims at classifying sequences which were classified as "unknown" before. Here a blast search against the RepeatMasker library is used. The second step is removing gene fragments. To achieve this, a repeat free library of swissprot plant proteins was build and all repeat sequences matching this in a blast search were excluded from the repeat library.

4.6. Detection of Centromeres

The workflow used to detect the centromeres is based on the method used by Melters et al. (2013). First, the Illumina input library is downsampled to get one paired-end library with 100 thousand read pairs (100K-lib) and one library with one million read pairs (1M-lib). Then PRICE is used to assemble the 100K-lib using one of the 1M-lib files as seed. PRICE runs for 25 cycles. Since the *D. muscipula* cycles needed much longer run times and the times are increasing with each cycle, only 18 cycles were run. As the third step, trf was used to find tandem repeats in the assemblies. The resulting candidate centromere sequences were combined and quantified. This was done by using blast to align both 1M-lib files to the tandem repeat sequences and counting the matches for each. This allows estimating the genomic fraction of each centromere candidate. The candidate with the highest count was selected as the centromere sequence.

4.7. Heterozygosity

In order to detect sites of heterozygosity, I first mapped Illumina reads to the genome assemblies. I combined all of the *D. spatulata* libraries which resulted in a total coverage of 28 x. For *D. muscipula*, I used the LGC-01 library and for *A. vesiculosa*, I used the L180-overlap library. For comparison, I also included the *A. thaliana* library ERR1913322 and the *Beta vulgaris* library SRR3929720. All libraries were downsampled to match 28 x coverage. I then used samtools mpileup with the `-g` parameter to create a bcf file from which the base composition in the mapped reads can be extracted for each position in the genome. Next, I searched this data for positions that show an approximate 50/50 base pair disagreement. I achieved this by calculating the ratio of the two mostly seen bases. If this ratio is between 0.8 and 1.2, I consider this position heterozygous. Positions with less than 10 reads mapped, were excluded.

Table 4.1.: **Software and Databases used for assembly and annotation of the three carnivorous plant genomes**

Name	Version	Citation
augustus	3.2.3	Stanke and Waack 2003, Keller et al. 2011
bowtie2	2.3.1	Langmead and Salzberg 2012
BUSCO	3.0.1	Simão et al. 2015, Waterhouse et al. 2018
canu	1.5	Koren et al. 2017
cufflinks	2.2.1	Trapnell et al. 2010
FASTQC	0.11.5	Andrews 2016
Interpro DB	64.0	Finn et al. 2017
Interproscan	5.25-64.0	Jones et al. 2014
maker	2.31.9	Cantarel et al. 2007
ontologizer	2.1	Bauer 2016
pilon	1.22	Walker et al. 2014
PRICE	0.6	Ruby et al. 2013
RepeatMasker	4.0.7	Smit et al. 2013
RepeatModeler	1.0.11	Smit and Hubley 2008
REVIGO	May 2018	Supek et al. 2011
samtools	1.4	H. Li et al. 2009, Danecek et al. 2011
snap	2013-02-16	Korf 2004
SOAP	r240	Luo et al. 2012
SPAdes	3.8.2	Nurk et al. 2013, Vasilinetc et al. 2015
star	2.5	Dobin et al. 2013
swissprot	Dec 2016	The UniProt Consortium 2017
trf	409	Benson 1999

5. Results

5.1. The *A. vesiculosa* Genome Assembly

At the start of the project, we sequenced *A. vesiculosa* with the Illumina technology, resulting in three paired-end libraries with different insert sizes. A 180 bp overlap library, a 550 bp and a 3 Kbp jumping library. Initially, Felix Bemm worked on a draft assembly and experienced some issues with the 3 Kbp jumping library. I evaluated the quality of the three libraries and found that the 3 Kbp library did not meet our standards. The FASTQC report showed high adapter contamination (see fig. 5.1), which accounted for 65 % of the raw data. I mapped the remaining 35 % on a draft assembly provided by Felix Bemm. Based on these mappings, I could estimate the insert size to be 500 bp, rather than 3 Kbp (see fig. 5.1c). Therefore, the 3 Kbp library was discarded for further analyses.

I used different assembly tools, SPAdes and SOAP to assemble the two remaining libraries. The SPAdes assembly could not be completed as the memory requirements exceeded our resources. The quality of the SOAP assembly was comparable to the *D. muscipula* assembly (see table 5.6). However, since the genome is much smaller and contains fewer repeats, we expected to get a more continuous assembly. In order to achieve this, we decided to go for an additional sequencing of *A. vesiculosa* using PacBio technology.

For assembling the PacBio data, we chose the canu assembler followed by a polishing step using pilon with Illumina data. In the first step, canu was used to assemble the raw PacBio reads (version 1.5, default parameter settings, except for `genomeSize=500m`). The resulting assemblies showed N50 values of 705 Kbp for *D. spatulata* and 314 Kbp for *A. vesiculosa*. The reported BUSCO completeness was 82.3 % and 79.5 % respectively (see also table 5.6). Next, the Illumina reads were mapped to the assembly with bowtie2 (version 2.3.1; default parameter settings, 32 threads). The resulting mapping was then analysed with pilon (version 1.22; default parameter; 20 threads) to refine the genome assembly. This resulted in slightly larger N50 values and better BUSCO completeness scores (see table 5.6). An overview of the final genome assemblies is given in table 5.1.

Table 5.1.: **Overview of the final genome assemblies.** The table shows the overview metrics of the final genome assemblies of *A. vesiculosa*, *D. muscipula* and *D. spatulata*. The first column is the published experimentally estimated genome size, the second column the size of the assembly, column three to five show information about the contig lengths and the last column the percentage of complete hits in the BUSCO assessment

	genome size	assembly size	contigs	longest	N50	BUSCO
A. vesiculosa	469 Mbp	420 Mbp	2408	3.4 Mbp	314 Kbp	C:86.9%
D. muscipula	2.85 Gbp	1.5 Gbp	104847	1 Mbp	35 Kbp	C:83.6%
D. spatulata	293 Mbp	238 Mbp	1061	3.4 Mbp	705 Kbp	C:86.0%

5.2. Annotations

The annotation pipeline resulted in a comparable number of genes for the three carnivores (see table 5.2). *A. vesiculosa* contains the highest number of genes (25 K). The gene counts of *D. muscipula* and *D. spatulata* are slightly lower (21 K and 18 K). The BUSCO completeness of all three gene sets is between 76 and 84 %. *A. vesiculosa* shows an increased duplication rate (11.2 %). While the number of exons per gene is similar (see table 5.7), *D. muscipula* has significantly longer introns compared to *A. vesiculosa* and *D. spatulata* (both $p - value < 2e - 16$).

Based on the Interproscan, I assigned IPR IDs to 80 % of the *D. muscipula*, 86 % of the *A. vesiculosa* and 87 % of the *D. spatulata* genes. Respectively, 56 %, 65 % and 66 % of the genes could be assigned at least one GO-term.

Table 5.2.: **Overview of the protein annotation.** This table shows the number of proteins annotated in the three genome assemblies, the length of the longest protein and the number of proteins with at least one hit in the InterPro scan. The last column shows the percentage of complete hits (C), including singe (S) and duplicated (D) hits from the BUSCO assessment.

	num proteins	longest	with Interpro hits	BUSCO
Aldrovanda	25123	4918	24450	C:84.3%[S:73.1%,D:11.2%]
Dionaea	21135	5081	19873	C:76.2%[S:72.5%,D:3.7%]
Drosera	18111	5761	17645	C:83.6%[S:80.1%,D:3.5%]

5.3. The Centromeres

The identification and quantification of the centromeric regions (see section 4.6) resulted in candidate sequences for all three species (see listing VI.1). The metrics monomer length, monomer GC content and genomic fraction fit well into the boundaries established by Melters et al. (2013). This is also shown in figure 5.2. The most noticeable difference between the three species is the genomic fraction. About 380 Mbp (13.5 %) of the *D. muscipula* genome consists of centromeric regions. That is more than the complete genome size of *D. spatulata*. An additional difference is the higher GC content in *A. vesiculosa*.

Table 5.3.: **Overview of the centromere candidates.** This table shows the monomer length, the GC content and the fraction of the genome (in % and Mbp) of the centromere candidate hits.

	Monomer length [bp]	**GC content [%]**	**gen. fraction**	**gen. part [Mbp]**
Aldrovanda	154	59.74	0.03	1.6
Dionaea	880	30.68	13.5	385
Drosera	617	28.36	2.3	5.5

5.4. Repetitive Elements

The RepeatMasker analyses showed that about 40 % of the *D. muscipula* genome assembly consists of repeats. The *A. vesiculosa* also contains up to 50 % repeats. *D. spatulata* contains fewer repeats. Here the analysis resulted in about 20 %. An overview of the results is given in table 5.4.

Table 5.4.: **Fraction of repeats masked in the three genome assemblies.** Comparison of the repeats masked by RepeatMasker using a custom repeat library constructed with reper and with RepeatModeler.

	with reper lib	**with RepeatModeler lib**
A. vesiculosa	48.44	44.27
D. muscipula	46.59	37.82
D. spatulata	13.74	32.74

5.5. Heterozygosity

The Search for heterozygous sites in the genomes yielded the following results. The number of hits in *A. vesiculosa*, *D. spatulata* and *A. thaliana* is comparable. *D. muscipula* and *B. vulgaris*, however, show a much higher number of heterozygosity sites (see table 5.5).

Table 5.5.: **Results of the Heterozygosity analysis.** This table shows the number of hits (an equally divided variation of two bases) in context of the genome size. The last two columns show the relative number of heterozygosity sites in the genomes.

	hits	**assembly size**	**rate**	**hits per Mbp**
A. thaliana	6049	120 Mbp	5.1e-5	50.5
A. vesiculosa	26029	420 Mbp	6.2e-5	61.9
B. vulgaris	934144	567 Mbp	1.6e-3	1648.8
D. muscipula	1656475	1455 Mbp	1.1e-3	1138.7
D. spatulata	9173	238 Mbp	3.9e-5	38.6

5.6. GO-enrichments of carnivore specific-genes

Franziska Saul, a Master student, I co-supervised, ran orthology predictions on the annotated genes, as well as genes of *A. thaliana* and *B. vulgaris*. In this dataset she identified a set of 162 orthogroups shared by the three carnivores and a set of 136 orthogroups shared by *A. vesiculosa* and *D. muscipula*, the snap-trap group (Saul 2019).

I calculated GO-enrichments for the genes in the carnivorous and snap-trap orthogroups of each plant and visualised the results with REVIGO (see fig. 5.3 and section 1.3). Among others, I found enriched GO terms associated with response to endogenous stimulus, transport, fatty acid metabolism, binding and sodium transmembrane transport in the carnivore specific orthogroups. In the snap-trap specific orthogroups, GO terms associated with DNA damage response, hydrolase activity and ion and protein binding were enriched. In *D. muscipula*, I also found enriched "transposase activity". Additionally, nine of the 24 GO terms described as carnivory related by Wheeler and Carstens (2018) are present in the three carnivorous plants. four of them were also enriched in the comparisons mentioned above: serine-type carboxypeptidase activity (GO:0004185), polygalacturonase activity (GO:0004650), superoxide dismutase activity (GO:0004784) and phosphatase activity (GO:0016791).

5.7. Improving the *D. muscipula* Transcriptome

In 2016, the *D. muscipula* transcriptome was published (Bemm et al. 2016). This transcriptome was assembled *de-novo* and is based solely on RNAseq data. Therefore it contains multiple assembly artefacts and a high number of isoforms (300 K). I developed a method to improve the transcriptome by using the genome sequencing data and the genome assembly to flag low-quality isoforms.

I used a combination of three quality metrics to evaluate each isoform. The first metric is a mapping to the genome assembly. Here I used the map2assembly script, which is included in the maker package. Using this tool, I could map the isoforms to the genome assembly and extract the information whether the isoform could be mapped or not from the resulting gff file. The second metric is the coverage in the genomic reads. For this, I used all genomic Illumina libraries, mapped them to the isoforms and calculated the median coverage for each isoform and sequencing library. An isoform is considered as supported when the median coverage of this isoform equals at least the expected library coverage minus 10 %. The third metric is based on expression counts. Here I calculated the median expression of the replicates in each sample in all RNASeq experiments. An isoform is supported if at least one sample has a median expression of 15. Each isoform that passes at least two of these three metrics is considered as trusted.

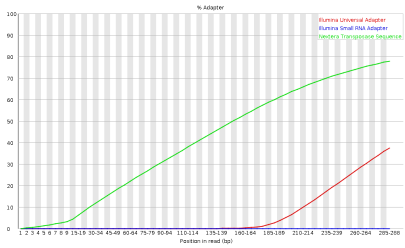
Using this method, I was able to flag about a third of the isoforms from the original transcriptome assembly as low quality. These can now be discarded from follow up analyses. In order to make this data accessible, I set up an internal instance of TBro (Ankenbrand et al. 2016) and uploaded the results there.

Table 5.6.: **Assembly statistics of the three carnivorous plant genomes using different assembly strategies.** This table shows different quality metrics of the assemblies that were created during the project. The table shows the number of contigs, the assembly size, the N50 length and the BUSCO quality summary. The BUSCO data consists of the percentages of complete hits (C), including single (S) and duplicated (D) hits, as well as the percentages of fragmented hits (F) and missing genes (M).

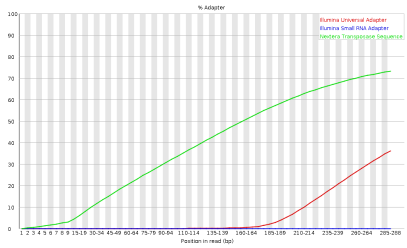
Assembly	**contigs**	**size**	**N50**	**BUSCO**
D. spatulata canu	1065	238 Mbp	705 Kbp	C:82.3%[S:79.7%,D:2.6%],F:3.5%,M:14.2%
D. spatulata pilon	1061	238 Mbp	705 Kbp	C:86.0%[S:82.9%,D:3.1%],F:2.8%,M:11.2%
A. vesiculosa soap	658725	516 Mbp	32 Kbp	C:97.9%[S:79.3%,D:18.6%],F:3.7%,M:17.1%
A. vesiculosa canu	2408	420 Mbp	313 Kbp	C:79.5%[S:72.2%,D:7.3%],F:3.8%,M:16.7%
A. vesiculosa pilon	2408	420 Mbp	314 Kbp	C:86.5%[S:76.6%,D:10.3%],F:1.8%,M:11.3%
D. muscipula	104847	1455 Mbp	35 Kbp	C:83.6%[S:80.5%,D:3.1%],F:3.8%,M:12.6%

Table 5.7.: **Overview of CDS annotation.** This table gives an overview of the coding sequences annotated in the three genome assemblies. The total number, the length of the longest CDS, the mean number of exons and the mean intron lengths are shown.

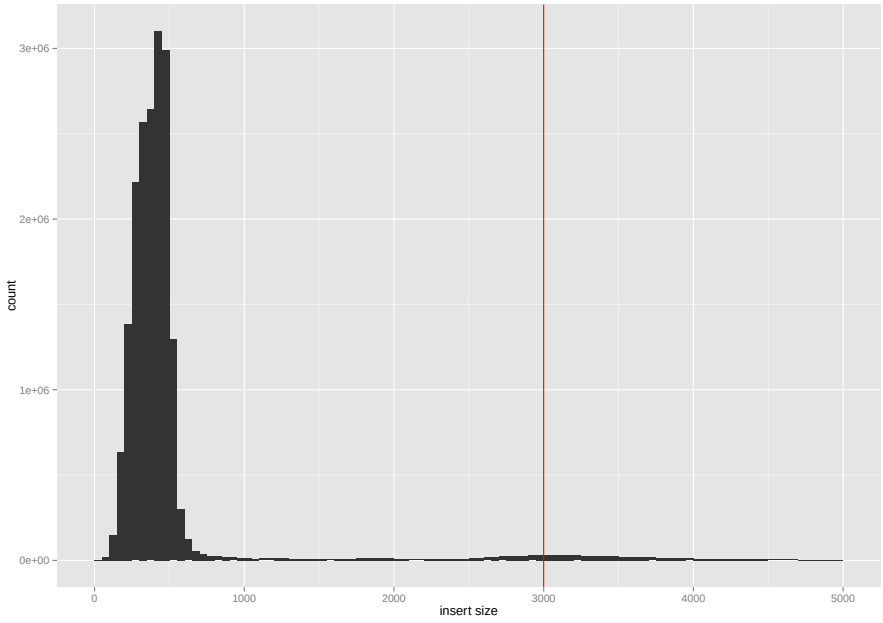
	num cds	**longest**	**mean num exons**	**mean intron length**
Aldrovanda	25123	14757	5.8	400.0
Dionaea	21135	15246	5.0	690.0
Drosera	18111	17286	6.0	400.6



a) Adapter content of reads



b) Adapter content of mates



c) Insert size estimation

Figure 5.1.: **Quality assessment of the *A. vesiculosa* 3 Kbp jumping library.** Figures a) and b) show the adapter contamination estimated using FASTQC. Figure c) shows the estimated insert sizes based on mapping the reads to a draft assembly. The red line indicates the target size of 3 Kbp.

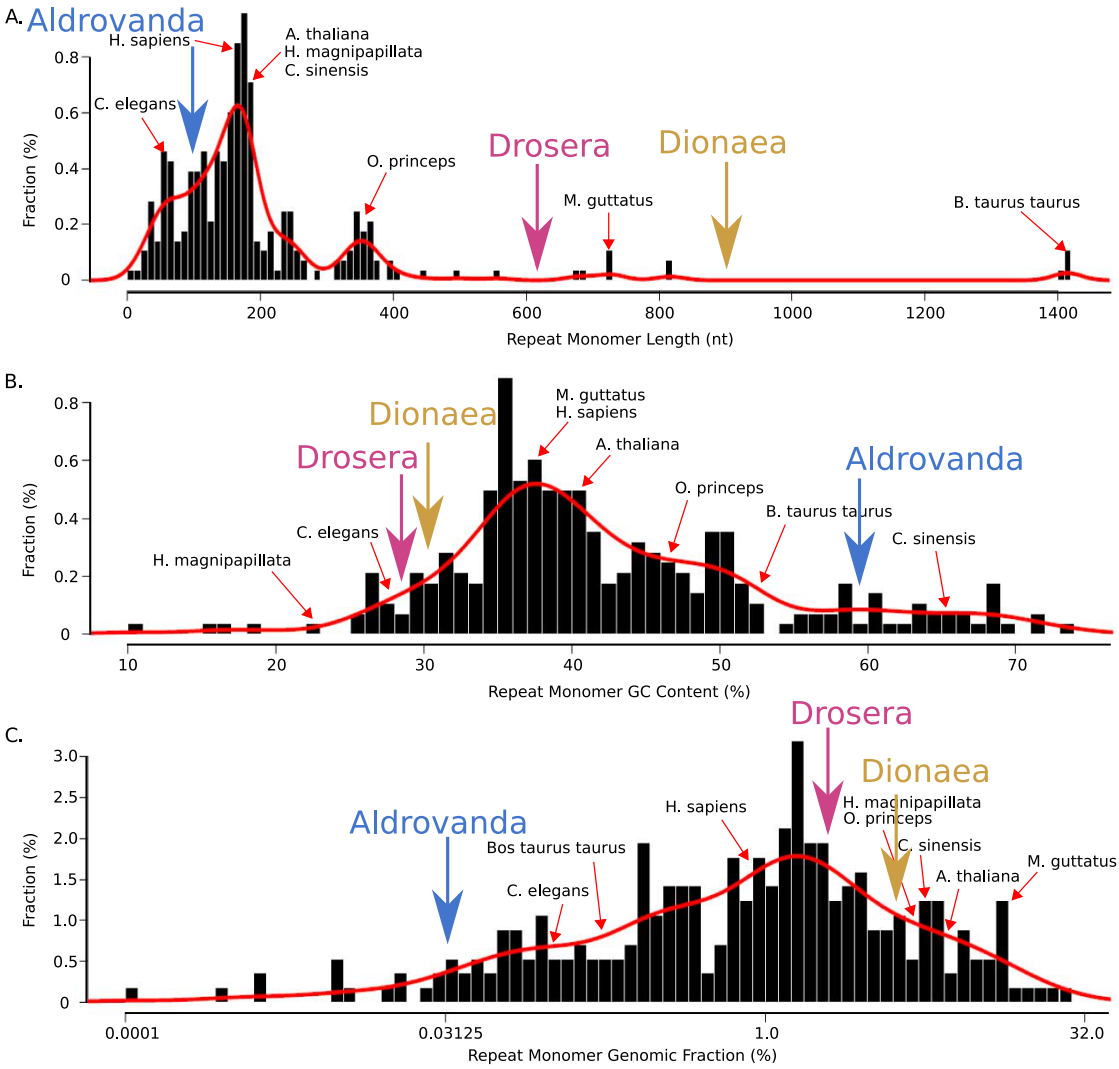
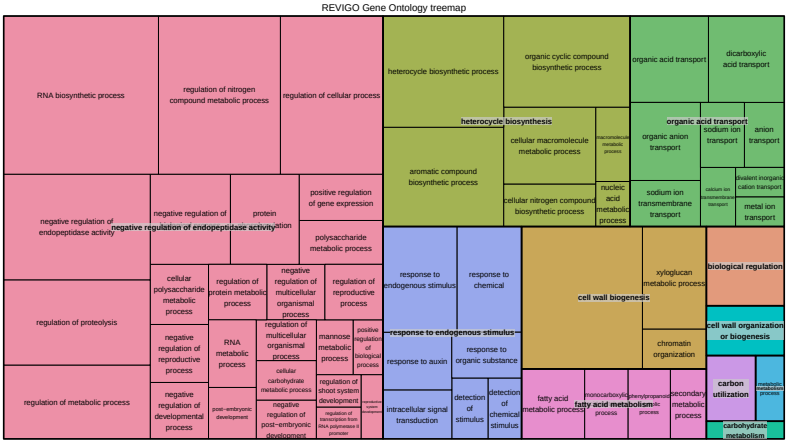
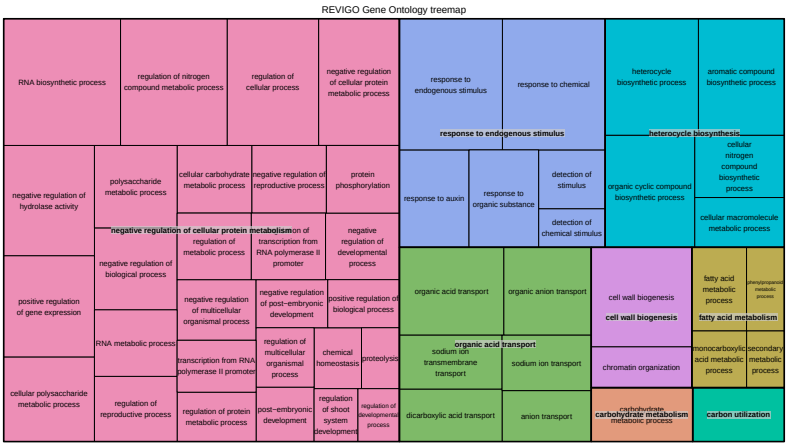


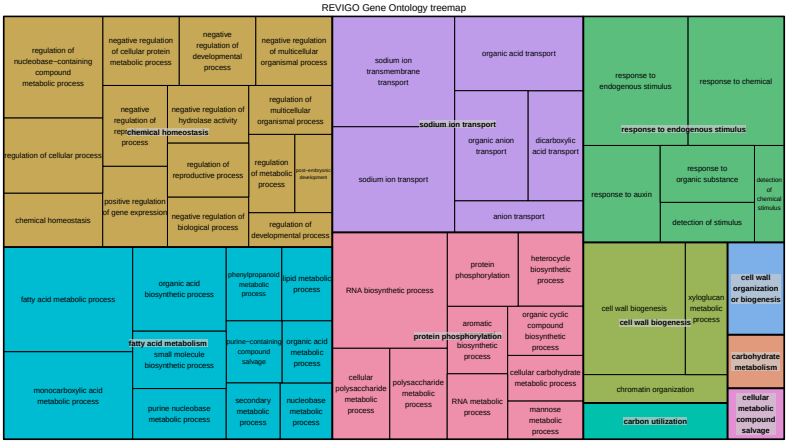
Figure 5.2.: **Comparison of centromere sequences.** Approximate positions of *A. vesiculosa*, *D. muscipula* and *D. spatulata* monomer length, GC content and genomic fraction of the centromeres in comparison to other organisms. This figure is a modified version of Figure 5 from Melters et al. (2013), provided under the CC-BY license



a) *A. vesiculosa*



b) *D. muscipula*



c) *D. spatulata*

Figure 5.3.: **Enriched Biological Process GO terms in the carnivorous orthogroups.** The three figures show treemaps of the GO terms enriched in the orthogroups specific for the three carnivourous plants. The figures have been created with REVIGO based on the GO enrichment results from ontologizer.

6. Discussion

6.1. An Overview of the Genomes

The assembly quality of *A. vesiculosa* and *D. spatulata* are good in terms of size, N50 and BUSCO completeness (see table 5.1). The *D. muscipula* assembly has some flaws regarding assembly size and N50 contig length. However, the BUSCO completeness of 83.6 % suggests that most of the protein-coding part of the genome was assembled and that the missing part consists mostly of repetitive elements. This is also supported by the protein annotation. The number of proteins, their number of Interpro hits, as well as the protein BUSCO completeness is similar for all three plants (see table 5.2) except that *A. vesiculosa* shows a higher duplication rate. This could be a result of a recent genome triplication (Gergő Palfalvi 2018, personal communication).

6.2. Repetitive Elements

Repetitive elements comprise a large proportion of many plant genomes (Feschotte et al. 2002). As the three carnivores have a large difference in genome size, the composition of repeats within these genomes is an interesting point. I used different methods to detect, classify and quantify repeats in our three genomes (see section 4.5), one of which was newly developed for this project (see part IV, Terhoeven et al. 2018). An overview of the results is given in table 5.4.

When interpreting the RepeatMasker results, it is important to know that compared to the RepeatModeler library, the reper library will miss some sequences, that look like repeats, e.g. have LTR regions, but are represented only one or two times in the genome. However, it will include sequences, that are present multiple times, but not incorporated into the assembly (mostly because of difficulties assembling repeats from short reads). Depending on the nature of the repeats in a plant, one of those categories may be larger. As shown in table 5.4, the fraction of bases masked in the assemblies of *A. vesiculosa* and *D. spatulata* is lower when using the reper library. That means, that most of the repeats are incorporated into the assembly and their similarity is too low for the detection by reper. In contrast to that, the reper library leads to a higher number of repeats in *D. muscipula*. That indicates a high similarity of these repeats and that some of them are not assembled correctly so, the RepeatModeler library misses them.

Comparing the reper based Repeatmasker results to the reper included quantification by Matthias Freund, another interesting feature emerges. In his analysis, Matthias Freund reports a repeat fraction estimate of the three genomes, 53 % for *D. muscipula*, 25 % for *A. vesiculosa* and 7 % for *D. spatulata* (Freund 2019). These numbers are based

on the genome size and not, like the RepeatMasker analysis, on the assembly size. This difference has the most impact on explaining the *D. muscipula* results. As the assembly size is just a bit more than half the genome size, the reper results give an estimate of the unassembled parts of the genome. In contrast to *D. muscipula*, the genome assemblies of *A. vesiculosa* and *D. spatulata* are almost complete. Therefore, a different reasoning applies here. As seen in the comparison between the reper and the RepeatModeler libraries above, the repeats in *A. vesiculosa* and *D. spatulata* are mostly complete and have low frequencies in the genome assembly. In this scenario, the reper quantification method underestimates the total number of repeats, as the alignment parameters are quite strict. RepeatMasker, however, uses less strict alignment parameters and therefore finds additional sequences in the assembly, that are not similar enough for the reper-only approach.

As the previous analyses suggest, the repeats of *D. muscipula* seem to be different from the other two plants. One possible explanation for this would be a recent expansion. Matthias Freund, a Master Student I co-supervised, developed a method to test this. He uses the Jukes-Cantor distance to quantify the similarity between LTRs. His results show a high density of evolutionary events with very low Jukes-Cantor distances in *D. muscipula*. The other two plants, especially *D. spatulata*, show much lower numbers in the low Jukes-Cantor distance regions and relatively high numbers for higher distances. That means, the transposons of *A. vesiculosa* and *D. spatulata* show much more variations than the transposons in *D. muscipula*, which indicates a recent expansion of repeats in the *D. muscipula* genome.

6.3. GO enrichment

The GO enrichment of the carnivore and snap-trap specific orthogroups showed various hints of carnivory and defence. It has been shown, that the Venus Flytrap uses chemical sensing, mostly chitin, to detect whether a captured object is actually a suitable prey (Darwin 1875, Hedrich and Neher 2018). The enriched GO terms "response to organic substance" and "detection of chemical stimulus" may relate to this. The "hydrolase activity", "fatty acid metabolism" and various transport related terms might be related to digestion and nutrient uptake. Of the 24 GO terms linked to carnivory by Wheeler and Carstens (2018), three were enriched in the carnivore specific groups and four in the snap-trap specific groups. Five more could be found in the dataset but were not enriched. Additionally, the term "transposase activity" was enriched in *D. muscipula*, which also hints to a high transposon activity in the genome of the Venus Flytrap.

6.4. Heterozygosity

It has been reported that *A. vesiculosa* has almost no variance within its population (Hoshi et al. 2006). In contrast to that, Thomas Hackl suggested that *D. muscipula* has a very high heterozygosity (Hackl 2016). With the now assembled genomes, I was able to test these hypotheses. Indeed, I found a very high number of heterozygous sites in the *D. muscipula* genome (see table 5.5). However, the rate of these sites in *A. vesiculosa* is not unusually low. In fact, it is comparable to *A. thaliana* and even higher than *D. spatulata*.

6.5. Differences in Genome Size

Compared to *A. vesiculosa* and *D. spatulata*, the genome of *D. muscipula* is huge. A common explanation for this are whole genome duplications. However, Gergő Palfalvi found no evidence for this in *D. muscipula* (Palfalvi 2018, personal communication). Taking a look at the genome composition, however, should provide insights here. Beside the differences in genome size, figure 6.1 shows this composition. As stated in section 5.2, *D. muscipula* has longer introns. However, this difference is negligible for the genome size differences. The differences in centromere size (see section 5.3) have a much bigger impact here. However, the most important difference is the amount of repetitive elements. An extensive analysis of these can be found in section 6.2. The *D. muscipula* LTR regions alone comprise more Megabases than the total genomes of *A. vesiculosa* and *D. spatulata* together. One possible explanation for this expansion could be stress induced effects caused by the artificial reproduction used in *D. muscipula* breeding. However, this theory could be discarded by genome size measurements of wild and cultured plants (Traud Winkelmann, Leibnitz Universität Hannover 2016, personal communication). She reported genome sizes between 3.16 Gbp and 3.25 Gbp for wild type and cultivated plants. These various genome sizes reported for *D. muscipula* are also indicated in fig. 6.1.

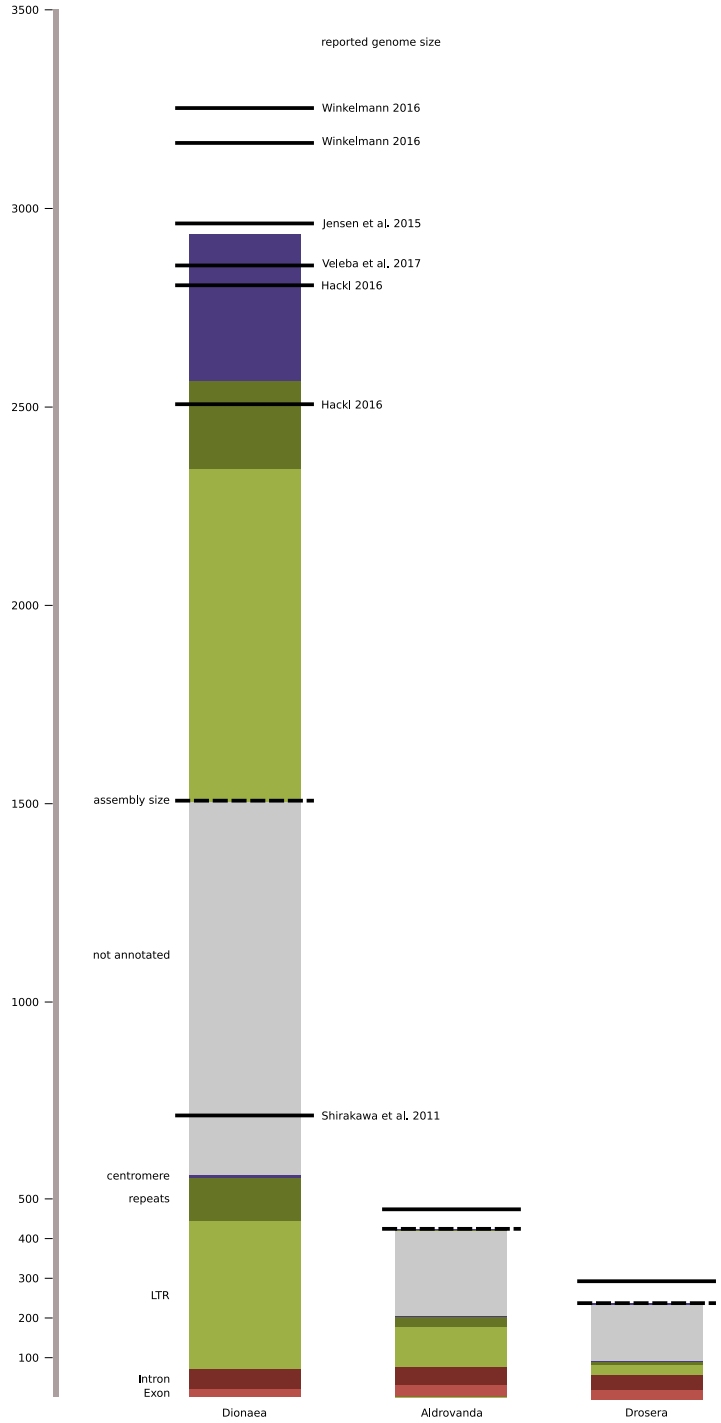


Figure 6.1.: **Genome composition of *A. vesiculosa*, *D. muscipula* and *D. spatulata*.** The height of the bars indicate the number of Mbp allocated for each group of features in the genome, Introns, Exons, LTRs, other repeats, centromere and not annotated regions. The dashed lines show the assembly size and the solid lines the genome sizes. For *D. muscipula*, multiple reported genome sizes are shown.

Part III.

Transcriptomics of Pollen Tube Guidance

7. Methods

7.1. estimating sequencing needs

To estimate the sequencing depth needed for this project, I ran RNASeq analyses with different *N. tabacum* datasets while reducing read length and coverage. I used 14 RNASeq libraries (see supplemental table VI.2) from the *N. tabacum* genome project (Sierro et al. 2014). This dataset consists of two experiments. An experiment comparing root with leaf tissue and an experiment comparing flower tissues of different age (young, mature, senescent). For the root/leaf experiment, I created nine subset libraries with coverage reduced to 75 % and 50 % and the read length reduced to 75 bp and 50 bp. The flower experiment was sampled into three subsets. The complete dataset, 75 bp, and 75 bp-50 % coverage.

The analysis of the root/leaf experiment was done using the reference based approach in the tuxedo suite toolset (Trapnell et al. 2010). Here tophat (Kim et al. 2013), cufflinks (Trapnell et al. 2010) and cummeRbund (Goff and Trapnell 2017) were used to find differentially expressed genes in each of the nine subsets. The workflow followed the one specified in Trapnell et al. 2012. The results were then compared between the nine different subsets.

The analysis of the flower-age experiment was done using a different approach. Here I mapped the reads to the genome assembly using tophat and calculated the differential expression using DESeq2 (Love et al. 2014).

7.2. RNASeq analysis

All scripts and commands used to analyse the RNASeq data were uploaded to GitHub and archived in Zenodo. The scripts can be found at `https://github.com/nterhoeven/tobacco-PT-guidance-scripts` (`doi:10.5281/zenodo.3228130`)

For this study, we used four different samples of *N. tabacum* SR1, namely pollen-tube (PT), ovule only (OV), ovule pollinated (OVP) and ovule fertilized (OVF). The samples were prepared by Katharina von Meyer (Lehrstuhl für Pflanzenphysiologie und Biophysik, Universität Würzburg) as follows. For the PT samples, the pollen was harvested and after growth of a pollen tube, the sample was frozen. For the OV samples, ovules were extracted from the anthers and then frozen. For the OVP and OVF samples, flowers were pollinated by hand and the growth of the pollen tube was monitored using fluorescence microscopy. After the pollen tube has grown halfway to the ovules (OVP sample) or the full way down and into the ovule (OVF sample), the ovules were extracted and frozen. After RNA extraction, quality control and DNA digestion, the samples were

sequenced on an Illumina HiSeq 3000 sequencer. For details about the wet-lab process see the dissertation of Katharina von Meyer.

We sequenced 12 libraries using paired-end Illumina technology. As some of these libraries did not meet our quality standards (see section 8.2), we sequenced additional four libraries. The libraries used for further analyses were then renamed to follow a consistent naming scheme. An overview of all sequencing libraries is given in table 7.1.

The sequencing libraries were quality and adapter trimmed using trimmomatic 0.33 (Bolger et al. 2014) (`trimmomatic-first-set.sh` and `trimmomatic-second-set.sh`). The quality was evaluated with fastqc (Andrews 2016) (`fastqc-first-set.sh` and `fastqc-second-set.sh`).

The transcriptome was assembled using Trinity Version v2.2.0 (Grabherr et al. 2011) in the genome-guided mode. We preferred this method in contrast to a typical reference-based assembly, because there was no genome sequence for the SR1 Strain available. Using Trinity, we could use the TN90 Strain genome assembly (Sierro et al. 2014) as a reference. The first step, mapping the reads to the genome was done using tophat v2.0.11 (Kim et al. 2013) (`run_mapping.sh`). The commands used for the Trinity step can be found in `run_trinity.sh`.

The assembled transcripts were annotated using the following approach: First, Trapid (Van Bel et al. 2013) was used to predict peptides. This was done using the Trapid web interface (`http://bioinformatics.psb.ugent.be/webtools/trapid/`) and plaza 2.5 (Van Bel et al. 2012) as reference database. Then peptides with less than 10 amino acids were removed and the remaining peptide sequences were subject to the functional annotation. We used the mercator (Lohse et al. 2014) web interface (`http://www.plabipd.de/portal/mercator-sequence-annotation`) to assign mapman (Usadel et al. 2009) categories to the peptides as well as InterPro scan Version 5.25-64.0 (Jones et al. 2014; Finn et al. 2017) for the annotation of various other features such as GO-terms (Ashburner et al. 2000; The Gene Ontology Consortium 2017) and Pfam domains (Finn et al. 2016) (`chunk-input-data.sh`, `run_Interproscan.sh`, `start-jobs.sh`).

The differential expression analysis was done using a typical DESeq approach. The transcripts were indexed and quantified using salmon version 0.7.2 (Patro et al. 2017) (`run_salmon-index.sh` and `run_salmon-quant.sh`) and these results were then analyzed with DESeq2 (Love et al. 2014) to calculate the differential expressions. The `de_seq.R` script used here was written based on the DESeq2 tutorial on bioconductor (`https://bioconductor.org/help/workflows/rnaseqGene/`).

For easy access and usage for following analyses, the annotated transcriptome and the expression data was uploaded to a TBro (Ankenbrand et al. 2016) instance.

Table 7.1.: **RNASeq Libraries used in the tobacco project.** This table gives an overview of the RNASeq libraries. The name (original and new), number of reads, total number of bases, longest and shortest read lengths and the N50 are shown.

name	original name	reads	bases	max	min	N50
OV-A	NtOV2_S40_L004_R1	63927395	8935904733	151	35	150
OV-A	NtOV2_S40_L004_R2	63927395	8940331171	151	35	150
OV-B	NtOV3_S38_L004_R1	46063023	6490724835	151	35	150
OV-B	NtOV3_S38_L004_R2	46063023	6493084338	151	35	150
OV-C	NtPT6_S42_L004_R1	57201131	8043024946	151	35	150
OV-C	NtPT6_S42_L004_R2	57201131	8046499666	151	35	150
OV-D	Wuerz1_1_S33_L002_R1	24176124	3308233134	151	35	149
OV-D	Wuerz1_1_S33_L002_R2	24176124	3310675643	151	35	149
OV-E	Wuerz1_2_S59_L004_R1	19052499	2612007701	151	35	149
OV-E	Wuerz1_2_S59_L004_R2	19052499	2614430570	151	35	150
OVF-A	NtOVF1_S43_L005_R1	48740369	6687225177	151	35	150
OVF-A	NtOVF1_S43_L005_R2	48740369	6689706589	151	35	150
OVF-B	NtOVF2_S41_L004_R1	43933140	5626912309	151	35	150
OVF-B	NtOVF2_S41_L004_R2	43933140	5629811490	151	35	150
OVF-C	NtOVF5_S45_L005_R1	47396865	6396543619	151	35	150
OVF-C	NtOVF5_S45_L005_R2	47396865	6399877682	151	35	150
OVP-A	NtOVP3_S44_L005_R1	50223124	7054946656	151	35	150
OVP-A	NtOVP3_S44_L005_R2	50223124	7056530914	151	35	150
OVP-B	NtOVP4_S39_L004_R1	52397399	7342789659	151	35	150
OVP-B	NtOVP4_S39_L004_R2	52397399	7345381104	151	35	150
OVP-C	NtOVP5_S47_L005_R1	99950641	13144577121	151	35	150
OVP-C	NtOVP5_S47_L005_R2	99950641	13147396334	151	35	150
PT-A	NtPT3_S46_L005_R1	68477981	9619021623	151	35	150
PT-A	NtPT3_S46_L005_R2	68477981	9622164344	151	35	150
PT-B	NtPT5_S37_L004_R1	82898552	11624742188	151	35	150
PT-B	NtPT5_S37_L004_R2	82898552	11630526750	151	35	150
PT-C	Wuerz2_1_S46_L003_R1	35208766	4850225064	151	35	149
PT-C	Wuerz2_1_S46_L003_R2	35208766	4853819225	151	35	149
PT-D	Wuerz2_2_S72_L005_R1	27605532	3783136382	151	35	149
PT-D	Wuerz2_2_S72_L005_R2	27605532	3786654184	151	35	149
-	NtOV1_S48_L005_R1	39956493	5261609131	151	35	150
-	NtOV1_S48_L005_R2	39956493	5264771484	151	35	150

8. Results and Discussion

8.1. Estimating sequencing needs

The root/leaf experiment resulted in many differentially expressed genes. Reducing the read length has almost no impact on the number of genes found. The sequencing depth, however, does show differences here (see fig. 8.1a). Comparing the lists of differentially expressed genes shows a large overlap and a variation of about 10-20 % (see fig. 8.1b).

The influence of read length and sequencing depth in the flower experiment is shown in a PCA plot (see fig. 8.2a). Compared to the complete library, the plot shows no difference for the 75 bp read length library and a slight difference when additionally reducing the sequencing depth to 50 %. It also shows that there might have been a mix up of a mature and a senescent sample, but that has no impact on this study because I am only interested in the differences between library sizes. This difference is a loss of about 1200 and a gain of 229 significant genes (see fig. 8.2b). The fraction of lost genes is higher (65 %) when looking at the 100 differentially expressed genes with the lowest expression values (see fig. 8.2c).

Both analyses show that the read length has almost no impact on the results. Therefore reducing the sequencing costs by choosing shorter reads should come at no quality loss here. The sequencing depth, however, has an impact on the results. This is especially true for genes with low expression values. As an overall conclusion, 40-50 Mio reads with a length of 75 bp should be a good amount for this project.

8.2. First Sequencing

While running the analysis pipeline, two issues were discovered. The first issue occurred while mapping the RNASeq reads to the genome assembly. All libraries had mapping rates of over 90 %. However, one library (OV-1) had a mapping rate of 0.4 %. Further investigation suggested that this library was contaminated with bacterial sequences.

The second issue was revealed in the PCA plot constructed during the DESeq analysis. As seen in figure 8.3a, one of the PT samples clusters clearly with the OV samples. This is also confirmed in the distance heatmap (fig. 8.3b). A single clear error like this indicates a mixed up labelling of the samples.

Since we had some more samples left, we decided to sequence four additional libraries: Two Ovule Only (Wuerz1_1 and Wuerz1_2) and two pollen tube (Wuerz2_1 and Wuerz2_2).

8.3. Second Sequencing

After the additional sequencing runs were complete, I ran the Trinity, salmon, DESeq analysis on the complete dataset. From this point on, I used the new library names mentioned in table 7.1. The *N. tabacum* transcriptome assembly resulted in 522736 isoforms comprising 451879 unigenes. In total, 299972 proteins were annotated, 261804 of which have at least one Interpro hit.

The DESeq2 analysis showed a huge difference between the pollen tube (PT) and the Ovule (OV, OVP, OVF) samples (see 8.4a). Since this difference superposes the differences between the three Ovule conditions, the analysis was rerun excluding the PT samples. Figure 8.4b shows a Principle-Component-Analysis of these samples. The plot shows a clear separation on the first axis between the OV only and the two OV conditions that came in contact with pollen tube. On the second axis, these two are clearly separated.

I uploaded the transcriptome assembly, the annotation, and the DESeq results to our internal TBro instance. That way, my collaborators were able to access the results, search for certain genes of interest, especially CRPs, and check their expression levels in the different conditions, as well as whether they are significantly differentially expressed.

8.4. Downstream Analyses

The dataset created in this study can now be used to gain various insights into the fertilisation process. One example is the role of Cysteine-Rich-Proteins (CRPs).

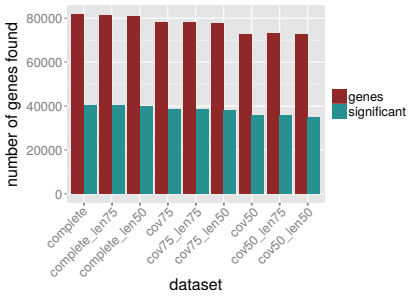
Using HMM profiles, Dirk Becker identified 953 CRPs in the protein set of *N. tabacum*. Based on the classification by Silverstein et al. (2007) and Huang et al. (2015), I assigned these into 201 classes, which can be further grouped into 17 categories. We did the same analysis for *A. thaliana*, the Chinese Cabbage (*Brassica rapa*, The Brassica rapa Genome Sequencing Project Consortium et al. 2011), maize (*Zea mays*, Schnable et al. 2009), rice (*Oryza sativa japonica*, Kawahara et al. 2013), tomato (*Solanum lycopersicum*, Wang et al. 2005) and the published *N. tabacum* genome. The results are shown in table 8.1.

I analysed the expression profiles of the CRPs found in our transcriptome assembly. As seen in fig. 8.5, the most significantly up- or down-regulated genes can be found in the comparison between the pollen tube and one of the three ovule samples. This is true for all CRPs and in compliance with the previously shown DESeq analysis.

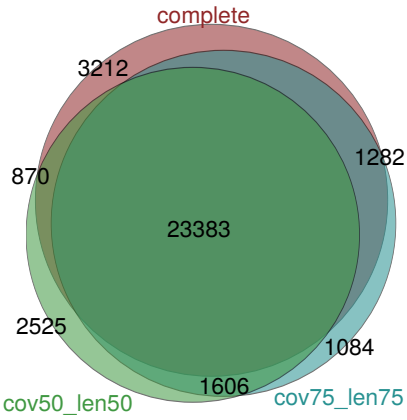
As it is previously known that DEFLs play an important role in pollination (Bircheneder and Dresselhaus 2016, Amien et al. 2010), we took a closer look at the expression of this group of CRPs. In a global overview (see fig. 8.5), the proportion of DEFL proteins that are up-regulated increases with progress in fertilisation. Based on these results, Katharina von Meyer selected six proteins for further validation by qPCR. She found significant gene expressions in Stigma, Style, Ovules and Placenta tissues. These results further validate the importance of these DEFLs in the fertilisation process.

Table 8.1.: **Number of CRPs found in the different species.** Tobacco SR1 is the data from our experiment, Tobacco TN90 is based on the data from the published genome. The other protein sets are based on the data in ensmbl plants (Monaco et al. 2014).

CRP Group	**SR1**	**TN90**	**Arabidopsis**	**Chin. Cabbage**	**Maize**	**Rice**	**Tomato**
Antimicrobial peptide MBP-1	3	0	1	0	4	14	1
Bowman Birk inhibitor	0	0	0	0	34	11	0
Defensin/DEFL	116	68	286	72	58	18	49
GASA/GAST/Snakin	44	38	21	16	14	11	20
Glutenin/gliadin/prolamin	0	0	0	0	3	30	0
Hevein	37	26	22	30	26	27	36
Kazal type inhibitor	10	2	2	3	4	1	2
Kunitz type inhibitor	28	28	8	33	1	1	16
LTP/2S Albumin/ECA 1	243	274	281	211	146	144	130
Maternally-expressed gene (MEG)/Ae1	12	3	15	3	35	8	2
No Name	284	224	256	233	190	142	120
Pollen Ole e I	56	76	41	63	37	43	24
Protease inhibitor II	26	14	1	1	2	6	12
RALF	44	15	42	36	27	33	10
Root cap/LEA	11	11	2	5	12	6	6
Stig1	6	11	6	7	3	1	11
Thionin	33	27	73	22	20	18	22

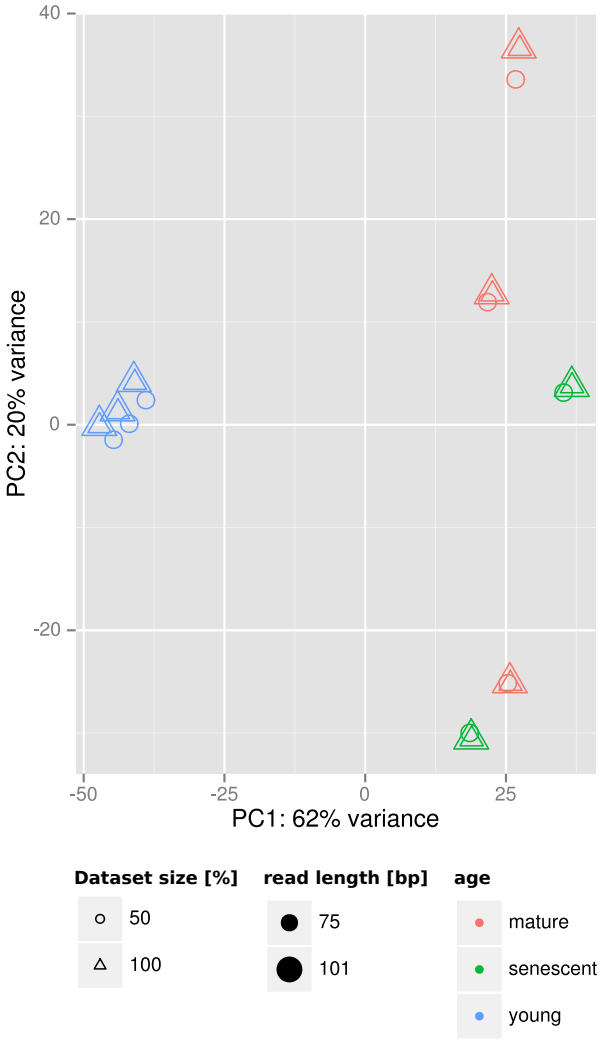


a) Number of differentially expressed genes

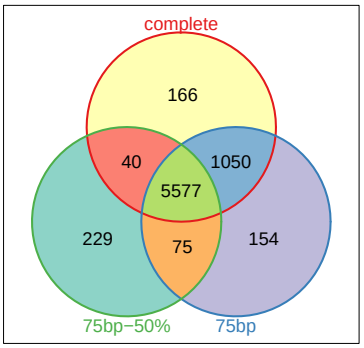


b) Comparison of genes found

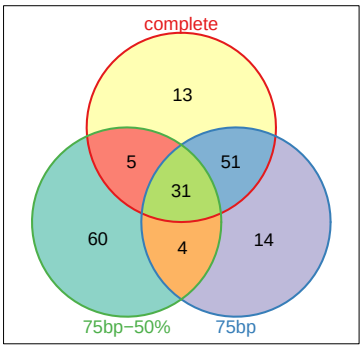
Figure 8.1.: **Overview of the root/leaf experiment.** The two figures show a comparison of the results of the root/leaf experiment with different datasets. The complete dataset and datasets where the coverage (cov) and the read length (len) were reduced to 75 or 50 %. Figure a) shows the total number of genes found along with the number of significant genes. The overlap of the significant genes found with three of these datasets is shown in the venn diagram (b))



a) PCA plot of the expression values

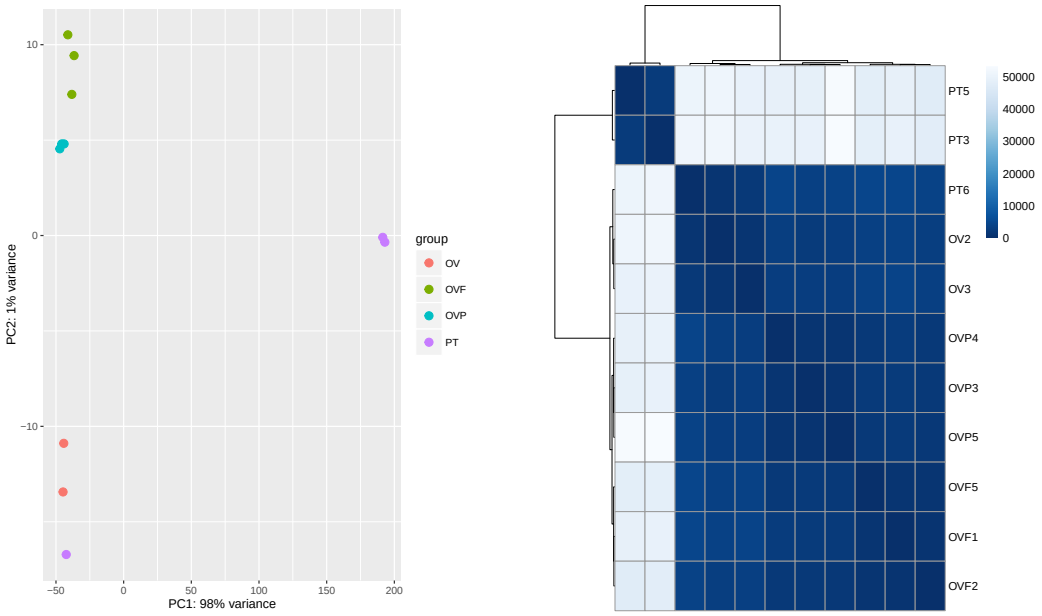


b) Overlap of the differentially expressed genes



c) Overlap of the 100 differentially expressed genes with the lowest counts

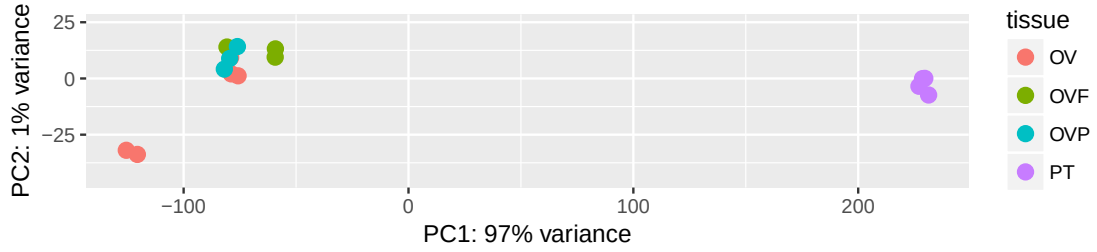
Figure 8.2.: **Overview of the flower experiment.** Figure a) is a PCA plot of the gene expression in the flower experiment across different dataset sizes. Figure b) and c) show the overlap of differentially expressed genes compared between the complete and two reduced datasets (75 bp with complete coverage and 75 bp with 50 % coverage).



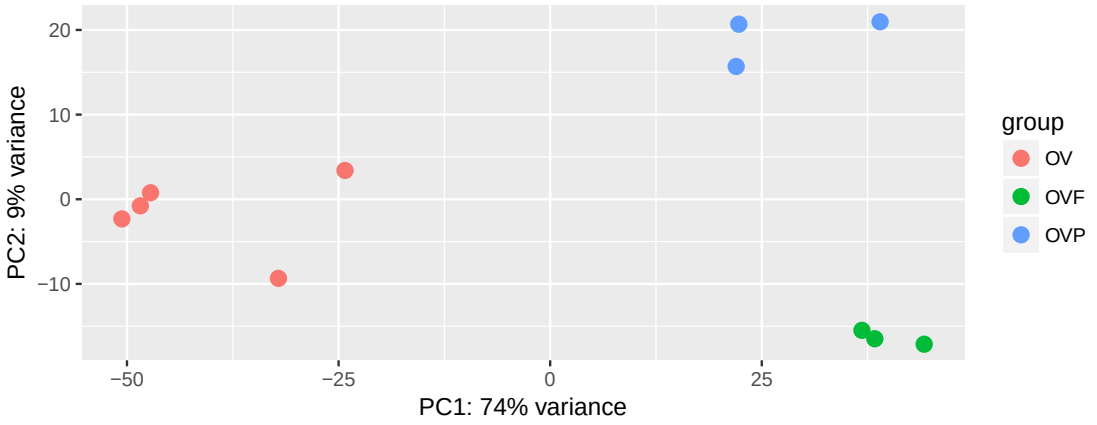
a) PCA plot of the expression values

b) Heatmap of the expression values

Figure 8.3.: **PCA plot and Heatmap of the first RNASeq analysis.** The two plots show that the PT-6 samples clusters with the OV samples. This indicates a mix-up labelling of the samples.



a) all samples



b) Pollen Tube excluded

Figure 8.4.: **PCA plots of the Expressions in the different Tobacco conditions.** These PCA plots show the differences between the individual samples of RNASeq experiment. Figure a) contains all samples. Because of the large difference between PT and the others, the differences within the OV samples cannot be seen. Figure b) shows this difference. This figure was the result of a new analysis where the PT samples were excluded.

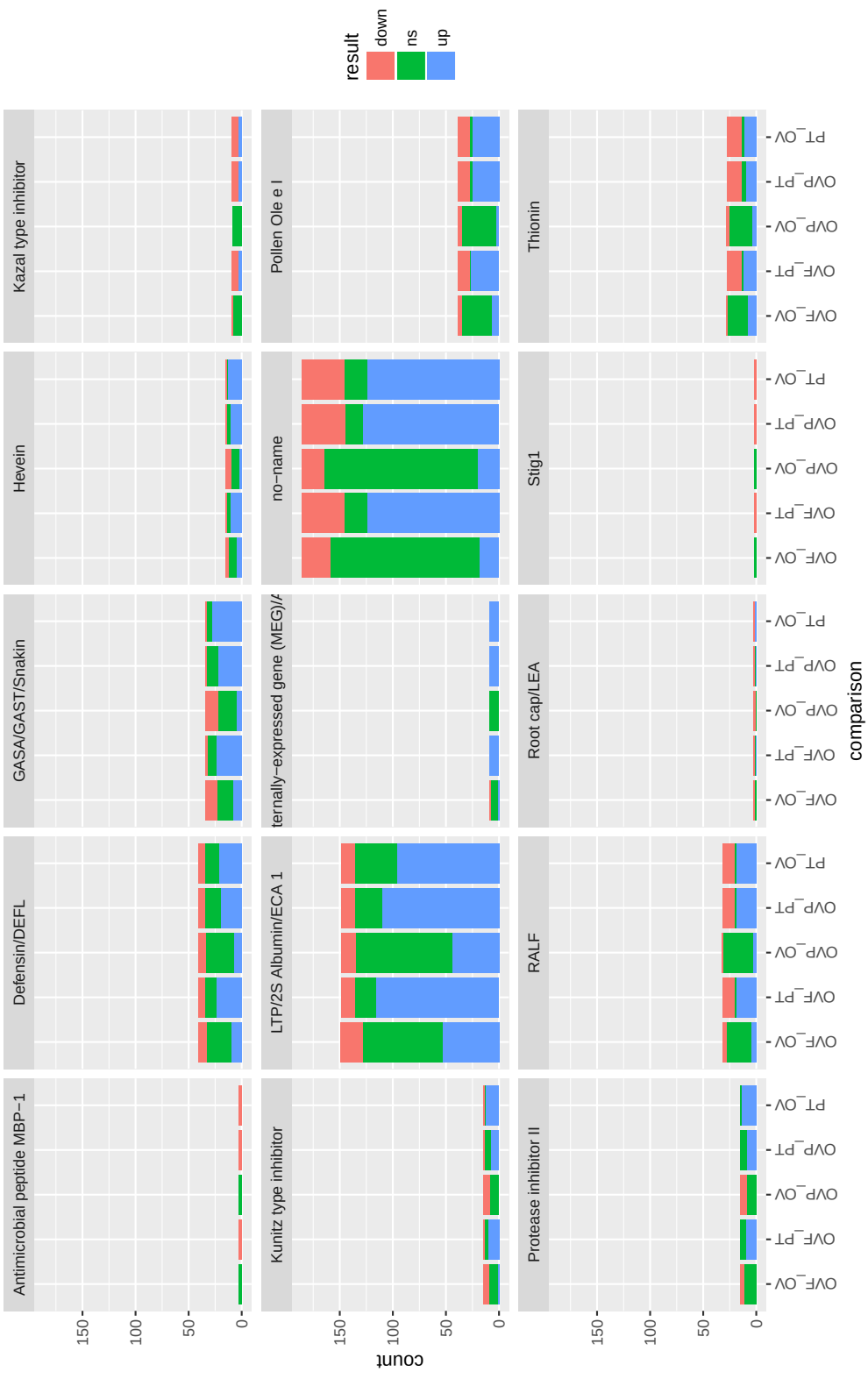


Figure 8.5.: **Expression Profiles of CRPs in the different Tobacco samples.** The height of the bars show the number of genes found for each CRP group. The colour shows how many of those are up regulated, down regulated or not significant. Genes that did ot show at least one significant difference, were excluded from this plot.

Part IV.

Reper: Genome-wide identification, classification and quantification of repetitive elements without an assembled genome

9. Implementing a kmer based repeat analysis workflow

The workflow of reper is illustrated in fig. 9.1. The input data reper needs is a paired-end Illumina library. Since the goal is the extraction of overrepresented sequences, a low coverage (i.e. 6 fold) is sufficient.

The first step is a kmer analysis. Here, jellyfish (Marçais and Kingsford 2011) is used to count 31-mers in the input data. Then the input reads are filtered based on these kmer counts. A read pair is kept if both reads fulfil the following requirement: At least 50 % of the kmers have counts that are at least five times the genomic coverage.

The remaining reads are then assembled using Trinity (Grabherr et al. 2011). In contrast to a genome assembler, Trinity can report multiple variants of a sequence (i.e. multiple isoforms of a unigene). This resembles the behaviour of different variants of a certain repeat. However, this is not perfect. To get a better grouping of the assembled sequences, I used cd-hit (W. Li and Godzik 2006; Fu et al. 2012) to create clusters. Two sequences are combined in a cluster if the alignment length is at least 90 % the length of the shorter sequence and the identity is at least 80 %. In each cluster, the longest sequence is chosen as representative and called exemplar.

These exemplars are used to classify the repeats. This is done using a blast (Camacho et al. 2009) searching against a reference database. The defaults here are REdat (Nussbaumer et al. 2012) and RefSeq (O'Leary et al. 2016). The blast results are then analysed as follows: First, all alignments are sorted into three confidence groups based on their e-value: high (e-value $<= 1e-3$), medium (e-value $<= 1$), and low (e-value > 1). These groups are then analysed from high to low until a decision is reached. If there are one or more hits in the high confidence group, the class is assigned based on the relative majority within these. If there are no hits in the high confidence group, the medium group is analysed. In this group, the class is also decided by the relative majority. However, if there are less than three hits voting for a class, the repeat is classified as "unknown". If a repeat has no hits in the high and medium group, the same procedure is applied to the low confidence group. However, the threshold for the majority is five hits in this group. If this is not reached or a repeat has no hits at all, it is classified as "unknown".

After the classification, the repeats are quantified. First, the original input reads are mapped on all assembled repeat sequences using bowtie2 (Langmead and Salzberg 2012). The results of these mappings are analysed with the script `build_repeat_landscape.pl`. This utilises samtools (H. Li et al. 2009) to extract the read counts for each sequence which is then used to calculate the number of base pairs the sequence contributes to the whole genome size. The final output tables (see examples 9.1, 9.2 and 9.3) which contain information on three levels of detail (class, cluster and sequence). The broadest

overview is given in the table on class level (table 9.1). For each class, it shows the accumulated read count, the number of sequences, their accumulated number of base pairs and the total amount of base pairs this class contributes to the complete genome. A more detailed view is given in the cluster level table (table 9.2). Here the cluster ID, the accumulated read count, the size of the cluster in base pairs and number of sequences, as well as its part of the genome in Mbp and the assigned class is given for each cluster. The sequence level table (table 9.3) shows information for each of the assembled sequences. It contains the sequence ID, the read count, the length of the sequence, its part of the genome as well as the cluster and class this sequence was assigned to.

The reper package also includes several helper scripts. For example, a script to prepare the giri repbase (Bao et al. 2015) for the use as a reference database in the classification step. It also includes a script to create a visual representation of the results using R (R Core Team 2015) and ggplot2 (Wickham 2016).

Table 9.1.: **reper example output on class level.** On the class level, the reper output table contains the read mapping count, the total number of base pairs and sequences as well as the genomic part (in Mbp) for each class found.

class	**count**	**num Bp**	**num Seqs**	**genomic part [Mbp]**
Retroelement	4090	5358	8	0.07
LINE	2259	1744	3	0.04
DNA	423424	189228	236	7.06

Table 9.2.: **reper example output on cluster level.** On the cluster level, the reper output table contains the read mapping count, the total number of base pairs and sequences as well as the genomic part (in Mbp) and the class for each cluster in the dataset.

cluster ID	**count**	**size [bp]**	**size [numSeqs]**	**genomic part [Mbp]**	**class**
1989	308	339	1	0.01	Unknown
138	2775	4835	2	0.05	DNA/En-Spm
2493	1724	494	2	0.03	LTR/Gypsy
3041	671	239	1	0.01	LTR/Gypsy
396	5000	4266	10	0.08	LTR

Table 9.3.: **reper example output on sequence level.** On the sequence level, the reper output table contains the read mapping count, the length of the sequence, the genomic part (in Kbp) as well as the cluster and the class for each sequence in the dataset.

sequence ID	count	length [bp]	gen. part [Kbp]	cluster ID	class
TRINITY_DN1739_c0_g5_i1	308	339	5.13	1989	Unknown
TRINITY_DN1739_c0_g3_i1	1679	2625	27.98	138	DNA/En-Spm
TRINITY_DN1739_c0_g3_i2	1096	2210	18.27	138	DNA/En-Spm
TRINITY_DN1478_c2_g2_i1	922	215	15.37	2493	LTR/Gypsy
TRINITY_DN1478_c2_g1_i3	802	279	13.37	2493	LTR/Gypsy

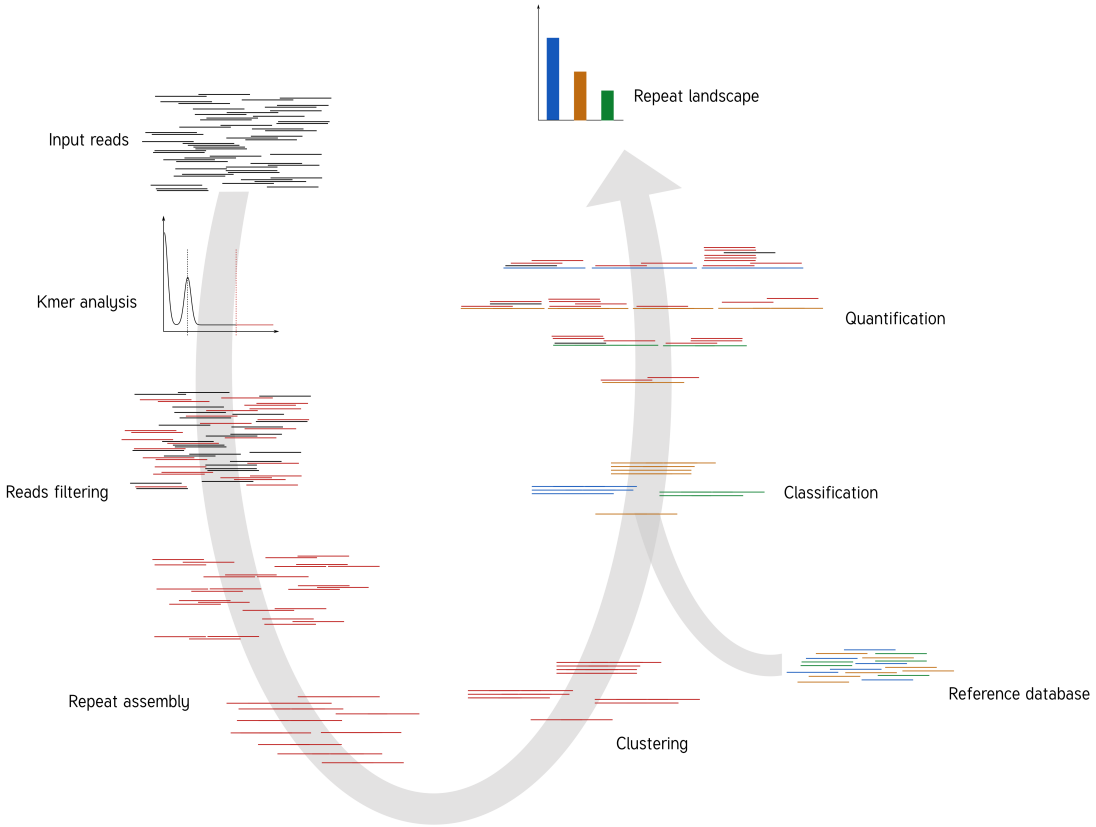


Figure 9.1.: **The reper workflow.** In the first step, a kmer analysis is used to label and filter the input reads. The extracted reads are then assembled and the resulting sequences are clustered. With the help of a reference database, these clusters are then classified. The reads are then mapped to the repeat sequences to quantify the repeats. As the last step, all information s gathered to produce the repeat landscape of the genome.

Part V.

Conclusion

10. Biological Insights and Lessons learned

Insights gained from my analyses

The carnivorous plant genomes project resulted in annotated genome assemblies of three carnivorous Droseraceae. That increases the total number of sequenced carnivorous plants to seven extending the variety of trap styles to include snap traps. The snap traps of *D. muscipula* and *A. vesiculosa* work differently (Westermeier et al. 2018). With the genome data, this could now be confirmed based on tissue-specific orthologous genes by my collaborators (personal communication). The analysis of repeat content in the *D. muscipula* genome confirmed Thomas Hackl's hypothesis that repeats are the main cause for the increased genome size. Other hypotheses, like unusually low heterozygosity in *A. vesiculosa* (Hoshi et al. 2006), could not be confirmed by my analyses. In combination with transcriptomic data, the genomes can now be used to study more aspects of these plants. For example, Anda Iosip is currently investigating the molecular process of the trap closure of *D. muscipula*.

Using these genomes and the gene predictions, a few other interesting studies could be done. *U. gibba* does not have any root and is missing root genes (Ibarra-Laclette et al. 2013). As adult *A. vesiculosa* plants also do not show roots (Adamec 2000), it would be interesting to know, whether the same root genes are lost here. The genome analysis of *D. capensis* did result in several new proteases linked to carnivory (Butts et al. 2016). It would be interesting to see if the same proteases occur in the genomes of the three Droseraceae sequenced in this study. With a similar focus, it would be possible to evaluate the presence of the 24 GO terms associated with carnivory (Wheeler and Carstens 2018). Some of them were found in the GO enrichment analyses of the carnivore and snap-trap specific orthogroups. However, this analysis does not provide a detailed insight into the whole GO term landscape of the three genomes. Further analyses might shed more light on this matter. Based on the genome annotations, my collaborators also analysed the expansion and contraction of gene families compared to several other plants. In the expanded gene families they found genes associated with peptidases and hydrolases as well as Jasmonic acid, a plant hormone associated with carnivory (Krausko et al. 2017). The contracted gene families contained genes associated with roots, which is particularly interesting since many carnivorous plants have lowly developed roots (Adamec 2000, Darwin 1875). My collaborators also showed that many of the carnivory associated genes expressed in *D. muscipula* traps are originally recruited from roots (manuscript in preparation).

The transcriptomic analysis of the fertilisation process in tobacco also confirmed prior hypotheses. We characterised the CRPs present in the tobacco transcriptome during fertilisation. We also found a rise in upregulated CRPs with proceeding fertilisation.

Additionally, the HMM analyses resulted in a large number of CRPs that could not be assigned to the classes based on Silverstein et al. (2007) and Huang et al. (2015). However, it has been suggested that many CRPs involved in fertilisation are not described yet (Bircheneder and Dresselhaus 2016). The not classified CRPs found here are potential candidates for novel classes. To gain further insights into their role, these candidates should be experimentally confirmed. Our results are in compliance with previous studies suggesting an important role of the defence originated CRPs in pollen tube guidance (J. B. Nasrallah and M. E. Nasrallah 2014, Qu et al. 2015, Bircheneder and Dresselhaus 2016).

In general, this study demonstrates two concepts of neofunctionalisation of defence mechanisms. Carnivory in plants evolved by reusing mechanisms to detect and defend against herbivores into trapping, killing and digesting prey (Hedrich and Neher 2018). While the work in this study is not focused on functional analyses, the annotated genome assemblies of *A. vesiculosa*, *D. spatulata* and *D. muscipula* are used as a base for further analyses by my collaborators investigating the evolution of carnivory. The second new duty for defence related mechanisms is shown in the *N. tabacum* transcriptome project. CRPs play an important role in cell-to-cell communication, with the DEFL subclass especially involved in defence against pathogens (Marshall et al. 2011). From their duty to protect seeds and early seedlings from pathogens, DEFLs have evolved to control and regulate most aspects of the fertilisation process (Bircheneder and Dresselhaus 2016). In our study, we could confirm their involvement in the pollen tube guidance on a transcriptomic level.

Lessons learned for future projects

This study also included some lessons learned on a technical basis. For one, working with plant genomes is challenging. The large size difference, mostly caused by whole genome duplications, multiploidy and repeats (Leushkin et al. 2013) can lead to high computational needs and difficulties in handling the data.

To address the challenge of a highly repetitive *D. muscipula* genome, I developed the reper software. With reper, we gained a method to analyse repeats in *de-novo* genome sequencing projects of non-model organisms. This allowed us to analyse the whole repeat content of the sequenced plants and gain insights about the parts that could not be assembled. To open reper to the public, we published a paper in the Journal of Open Source Software (JOSS) in February 2018 (Terhoeven et al. 2018). The complete software can be used under the terms of the MIT License and is available on GitHub (`https://github.com/nterhoeven/reper`).

Other challenges that arise in working with plant genomes could be addressed by newer methods. For example, it would have been useful to have nanopore sequencing data for the three carnivorous plant genomes. Unfortunately, this was not applicable at the time when the sequencing was done. At least, the improvement of the PacBio technology allowed us to use this as the main data source for the *A. vesiculosa* and *D. spatulata* assemblies. These long read technologies can have a great impact on the quality of plant genomes. A good example of this is the difference between the Illumina

only SOAP assembly and the PacBio based canu assembly of *A. vesiculosa*. The even longer nanopore reads could probably improve this further. The long reads can span repeat regions and whole genes, which is needed to assemble the individual parts in the correct order (Nakano et al. 2017). For future plant genome sequencing projects, I would recommend using a long read technology, like PacBio or nanopore, for the base assembly and an additional Illumina library for the refinement. This approach was successful for our *A. vesiculosa* and *D. spatulata* assemblies and has also been used in various other genome projects (e.g. Lan et al. 2017, Belser et al. 2018, Laverty et al. 2019). It would also be possible to start with a low coverage Illumina sequencing and use reper to estimate the repeat content before deciding on the optimal sequencing strategy.

Having an annotated genome is quite useful for transcriptomic experiments. The transcriptome only analyses by Bemm et al. (2016) had to cope with about 300 K isoforms. Using the genome data I was able to assign additional quality metrics to each isoform and unigene, resulting in about 100 K low-quality isoforms that can be excluded from further analyses. Additionally, the annotated genome assembly now allows the use of reference based transcriptome analyses. My collaborator Anda Iosip is currently working with this method to gain insights into the snapping mechanism of *D. muscipula*.

In contrast to that, I was not able to use the published genome assembly for a reference based analysis of the tobacco transcriptome since it is from a different strain. Therefore I had to build a new transcriptome assembly. With its 500 K isoforms, it shares the same issues as the original *D. muscipula* transcriptome. If we want to do more transcriptomic studies on the tobacco SR1 strain, it would be helpful to invest in a genome assembly first. This will make the transcriptome analyses easier and more comparable to each other. However, this will not be a fast and easy project, since *N. tabacum* is allotetraploid, 4.5 Gbp large and contains about 70 % repeats (Sierro et al. 2014).

A second possibility for improving the quality of transcriptome analyses could be the use of long read technologies, such as PacBio or nanopore (An et al. 2018). While this is still rare, it will probably become state of the art in the next few years. As in the genome sequencing, the long reads can span complete transcripts, which reduces the need for assemblies and brings new possibilities to analyse splice variants. However, due to the high error rates, a correction step using Illumina reads is still recommended (An et al. 2018).

The reason why I needed a new transcriptome assembly in the tobacco project, should also be considered for future studies in other plants. As *D. muscipula* shows a wide variety of phenotypes, there may be genomic differences similar to the different strains of tobacco. Therefore, different genome assemblies of the various strains would be helpful. This would also open other possibilities regarding population genomics and conservational biology. Both, *D. muscipula* and *A. vesiculosa* have small populations, but very different geographical distributions. While *A. vesiculosa* has low genetic variations (Hoshi et al. 2006), *D. muscipula* seems to be different, as seen from the heterozygosity analysis. More genomes from the global populations should allow more insights here, proof or disprove this hypothesis and help develop strategies for species conservation.

A base for future studies

In summary, I used and developed various methods for genomic and transcriptomic analyses of examples for neofunctionalisation of plant defence mechanisms. The resulting annotated genome assemblies of the three carnivorous Droseraceae are used as a foundation of various comparative analyses by my collaborators. Also, the reper pipeline was used by my collaborators in the carnivore genome project. The transcriptome analyses of Tobacco confirms the involvement of DEFLs in the fertilisation process and can also be used as a base for various future studies. While the work presented in the Thesis resulted in some interesting insights into the studied processes, it will also serve as a foundation for various future studies to gain further insights into the Kingdom of Plants and the repurposing of its defence mechanisms.

Part VI.

Appendix

Supplemental Materials

1. Evolution of Carnivory in Plants

1.1. Sequencing Libraries

Table VI.1.: **Sequencing libraries used in the carnivorous plants project.** A list of all sequencing libraries used in this part of the thesis: *A. vesiculosa* Illumina, PacBio and RNASeq libraries, *D. muscipula* genome and transcriptome libraries and *D. spatulata* genome and transcriptome libraries.

name	**reads**	**bases**	**max**	**min**	**N50**
A. vesiculosa genome Illumina					
Av_gen_J3Kb_P1.fastq	21599298	6479789400	300	300	300
Av_gen_J3Kb_P2.fastq	21599298	6479789400	300	300	300
Av_gen_L180_P1.fastq	201781384	20178138400	100	100	100
Av_gen_L180_P2.fastq	201781384	20178138400	100	100	100
Av_gen_L550_P1.fastq	139042165	13904216500	100	100	100
Av_gen_L550_P2.fastq	139042165	13904216500	100	100	100
TOTAL	724845694	81124288600	300	100	100
A. vesiculosa genome PacBio					
m170112_155407	97332	756530012	43059	35	12486
m170218_071121	101147	700925260	46299	35	11859
m170218_113014	60763	415655726	42358	35	12013
m170222_195627	89035	636794014	41248	35	11862
m170223_001644	100852	763580067	43302	35	11950
m170223_043729	99278	747260987	46542	35	11861
m170223_085732	94119	675858043	42040	35	11956
m170223_131917	93806	641524829	43564	35	11621
m170223_174024	93590	680485604	45335	36	12020
m170223_220343	82628	553125171	48506	36	11432
m170224_044228	73738	528169752	48260	37	12178
m170224_090309	73680	523587014	41237	35	12251
m170224_132447	66012	445715959	41222	35	11971
m170224_174606	63684	436638503	44534	35	11984
m170224_220738	60866	408727925	44252	36	11939
m170225_022801	65843	445678038	42458	35	11953
m170225_065152	68581	453431848	43713	36	11836
m170227_194829	47680	374459361	40894	35	11665
m170228_000950	74192	598390357	42339	35	11812

m170609_235905	159890	1009200149	41051	35	9664
m170614_190650	144960	987242164	46635	35	9903
m170614_232604	145639	979481606	43203	35	9680
m170615_034516	150980	1016948102	39366	35	9795
TOTAL	2014489	13575437873	48506	35	10771
		A. vesiculosa transcriptome			
Aldrovanda_R1.fastq	15397092	4619127600	300	300	300
Aldrovanda_R2.fastq	15397092	4619127600	300	300	300
TOTAL	30794184	9238255200	300	300	300
		D. muscipula genome			
dm-il-01_1.fq	342983444	34210566518	100	20	100
dm-il-01_2.fq	342983444	34208198231	100	20	100
dm-il-02_1.fq	346461493	34564374885	100	20	100
dm-il-02_2.fq	346461493	34562088172	100	20	100
dm-il-03_1.fq	344489284	34374751189	100	20	100
dm-il-03_2.fq	344489284	34372624536	100	20	100
dm-il-04_1.fq	441747363	44052248516	100	20	100
dm-il-04_2.fq	441747363	44048861064	100	20	100
TOTAL	2951363168	294393713111	100	20	100
		D. muscipula transcriptome			
DM_exp001_Fl_L1_P1.fastq	82344072	8316751272	101	101	101
DM_exp001_Fl_L1_P2.fastq	82344072	8316751272	101	101	101
DM_exp001_Fl_L2_P1.fastq	94576872	9552264072	101	101	101
DM_exp001_Fl_L2_P2.fastq	94576872	9552264072	101	101	101
DM_exp001_Fl_L3_P1.fastq	82183197	8300502897	101	101	101
DM_exp001_Fl_L3_P2.fastq	82183197	8300502897	101	101	101
DM_exp001_Gl_L1_P1.fastq	61104656	6171570256	101	101	101
DM_exp001_Gl_L1_P2.fastq	61104656	6171570256	101	101	101
DM_exp001_Gl_L2_P1.fastq	46414468	4687861268	101	101	101
DM_exp001_Gl_L2_P2.fastq	46414468	4687861268	101	101	101
DM_exp001_Gl_L3_P1.fastq	67264172	6793681372	101	101	101
DM_exp001_Gl_L3_P2.fastq	67264172	6793681372	101	101	101
DM_exp001_Pe_L1_P1.fastq	48249836	4873233436	101	101	101
DM_exp001_Pe_L1_P2.fastq	48249836	4873233436	101	101	101
DM_exp001_Pe_L2_P1.fastq	43846727	4428519427	101	101	101
DM_exp001_Pe_L2_P2.fastq	43846727	4428519427	101	101	101
DM_exp001_Pe_L3_P1.fastq	109174125	11026586625	101	101	101
DM_exp001_Pe_L3_P2.fastq	109174125	11026586625	101	101	101
DM_exp001_Ri_L1_P1.fastq	59582735	6017856235	101	101	101
DM_exp001_Ri_L1_P2.fastq	59582735	6017856235	101	101	101
DM_exp001_Ri_L2_P1.fastq	79822192	8062041392	101	101	101
DM_exp001_Ri_L2_P2.fastq	79822192	8062041392	101	101	101
DM_exp001_Ri_L3_P1.fastq	94468755	9541344255	101	101	101
DM_exp001_Ri_L3_P2.fastq	94468755	9541344255	101	101	101
DM_exp001_Ro_L1_P1.fastq	60765302	6137295502	101	101	101
DM_exp001_Ro_L1_P2.fastq	60765302	6137295502	101	101	101
DM_exp001_Ro_L2_P1.fastq	42496016	4292097616	101	101	101
DM_exp001_Ro_L2_P2.fastq	42496016	4292097616	101	101	101

DM_exp001_Ro_L3_P1.fastq	71275544	7198829944	101	101	101
DM_exp001_Ro_L3_P2.fastq	71275544	7198829944	101	101	101
DM_exp001_TrCor_L1_P1.fastq	68174440	6885618440	101	101	101
DM_exp001_TrCor_L1_P2.fastq	68174440	6885618440	101	101	101
DM_exp001_TrCor_L2_P1.fastq	76033562	7679389762	101	101	101
DM_exp001_TrCor_L2_P2.fastq	76033562	7679389762	101	101	101
DM_exp001_TrCor_L3_P1.fastq	51132019	5164333919	101	101	101
DM_exp001_TrCor_L3_P2.fastq	51132019	5164333919	101	101	101
DM_exp001_Tr_L1_P1.fastq	65538410	6619379410	101	101	101
DM_exp001_Tr_L1_P2.fastq	65538410	6619379410	101	101	101
DM_exp001_Tr_L2_P1.fastq	85082118	8593293918	101	101	101
DM_exp001_Tr_L2_P2.fastq	85082118	8593293918	101	101	101
DM_exp001_Tr_L3_P1.fastq	48264648	4874729448	101	101	101
DM_exp001_Tr_L3_P2.fastq	48264648	4874729448	101	101	101
TOTAL	2875587732	290434360932	101	101	101
	D. spatulata genome				
130308_L5-74_1.fq	6925745	692574500	100	100	100
130308_L5-74_2.fq	6925745	692574500	100	100	100
130314_L4-17_1.fq	7315882	1060802890	145	145	145
130314_L4-17_2.fq	7315882	1060802890	145	145	145
130314_L4-16_1.fq	11291266	1637233570	145	145	145
130314_L4-16_2.fq	11291266	1637233570	145	145	145
TOTAL	51065786	6781221920	145	100	145
	D. spatulata transcriptome				
idx1_R1_001.fastq	11906087	1500166962	126	126	126
idx1_R2_001.fastq	11906087	1500166962	126	126	126
idx2_R1_001.fastq	11749389	1480423014	126	126	126
idx2_R2_001.fastq	11749389	1480423014	126	126	126
idx3_R1_001.fastq	10803715	1361268090	126	126	126
idx3_R2_001.fastq	10803715	1361268090	126	126	126
idx4_R1_001.fastq	11306844	1424662344	126	126	126
idx4_R2_001.fastq	11306844	1424662344	126	126	126
idx5_R1_001.fastq	11423195	1439322570	126	126	126
idx5_R2_001.fastq	11423195	1439322570	126	126	126
idx6_R1_001.fastq	10164103	1280676978	126	126	126
idx6_R2_001.fastq	10164103	1280676978	126	126	126
idx7_R1_001.fastq	11347700	1429810200	126	126	126
idx7_R2_001.fastq	11347700	1429810200	126	126	126
idx8_R1_001.fastq	14813362	1866483612	126	126	126
idx8_R2_001.fastq	14813362	1866483612	126	126	126
idx9_R1_001.fastq	12994333	1637285958	126	126	126
idx9_R2_001.fastq	12994333	1637285958	126	126	126
idx10_R1_001.fastq	17696593	2229770718	126	126	126
idx10_R2_001.fastq	17696593	2229770718	126	126	126
idx11_R1_001.fastq	16741071	2109374946	126	126	126
idx11_R2_001.fastq	16741071	2109374946	126	126	126
idx12_R1_001.fastq	16424280	2069459280	126	126	126
idx12_R2_001.fastq	16424280	2069459280	126	126	126

idx13_R1_001.fastq	561	70686	126	126	126
idx13_R2_001.fastq	561	70686	126	126	126
idx14_R1_001.fastq	162	20412	126	126	126
idx14_R2_001.fastq	162	20412	126	126	126
idx15_R1_001.fastq	34	4284	126	126	126
idx15_R2_001.fastq	34	4284	126	126	126
idx16_R1_001.fastq	18	2268	126	126	126
idx16_R2_001.fastq	18	2268	126	126	126
idx18_R1_001.fastq	176	22176	126	126	126
idx18_R2_001.fastq	176	22176	126	126	126
idx19_R1_001.fastq	22	2772	126	126	126
idx19_R2_001.fastq	22	2772	126	126	126
idx20_R1_001.fastq	3	378	126	126	126
idx20_R2_001.fastq	3	378	126	126	126
idx21_R1_001.fastq	81	10206	126	126	126
idx21_R2_001.fastq	81	10206	126	126	126
idx22_R1_001.fastq	11	1386	126	126	126
idx22_R2_001.fastq	11	1386	126	126	126
idx23_R1_001.fastq	6	756	126	126	126
idx23_R2_001.fastq	6	756	126	126	126
idx25_R1_001.fastq	1556	196056	126	126	126
idx25_R2_001.fastq	1556	196056	126	126	126
idx27_R1_001.fastq	80	10080	126	126	126
idx27_R2_001.fastq	80	10080	126	126	126
TOTAL	314746764	39658092264	126	126	126

1.2. Centromere Candidates

Code VI.1: "Centromere sequences of the three carnivorous plant genomes"

```
>AV_centromere_candidate
GTACTGCGGACTCATCGAGTCGAGTCGCCTGTTCCTCGCGGGCGGCGAATCCATTGGTCT
GGACCAATGTTCGTCCCCCTCGTTGGCTCGGTTGACGGATCAAAGTCGGTTTCCCCATCA
GAGAGGCGTCGGTACTCGGAAACGGGGATCATGC
>DM_centromere_candidate
TTTCCATCCATTTCCATGATTATACTCTAGTTTTACATTGATTTCACATGGAAACGATGG
AATTAGCTTGTTTTCATCCATTTTCATTGTTTTCCATCCATTTCCATGATTATACTCTAG
TTTTACATTGATTTCACATGGAAACGATGGAATTAGCTTGTTTTCATCCATTTTCATTGT
TTTCCATCCATTTCCATGATTATACTCAAGTTTTACATTGATTTCACATGGAAACGATGG
AATTAGCTTGTTTTCATCCATTTTCATTGTTTTCCATCCATTTCCATGATAATACTCAAG
TCTACATTGATTTCACATGGAAATTACGTGTTAACATGTGAATCAATGTAAAACTAGAGT
ATAATCATGGAAATTGATGGTTTTCCATCCATTTCCATGATTATACTCTAGTTTTACATT
GATTTCACATGGAAACACTTAATTTCCATGTGAAATCAATGTGAGACTTGAGTATTATCA
TGGAAATTGATGGAAAACATGAAAATGGATGAAAACAAGCTAATTCCATCGTTTCCATGT
GAAATCAATGTAAAACTTGAGTATTATCATGGAAATGGATGGAAAACAATGAAAATGGAT
GAAAACAAGCTAATTCCATCGTTTCCATGTGAAATCAATGTAAAACTAGAGTATAATCAT
GGAAATGGATGGAAAACAATGAAAATGGATGAAAACAAGCTAATTCCATCGTTTCCATGT
GAAATCAATGTAAAACTAGAGTATAATCATGGAAATGGATGGAAAACAATGAAAATGGAT
GAAAATAAGCTAATTCCTTTGTTTCCTTGTGAAATCAATGTAAAACATTGATTTCACATG
GAAACAAAGGAATTAGCTTGTTTTCATCCATTTTCATTGT
>DS_centromere_candidate
TAGTTTCTTGCGTTTCCAAGTTAATTTGAGGTTGTTTTATAGTGTTTTTTAGGTTTTTGA
TAGATTTGAGTTGTTTTATGGTGCTTTTAAGGCTCTTCATGTGTTTTTGAGTGTTTTAAA
ATCATTTCAATGCTCTTTTGCTACATATCATGCGTTTTTAAGTGATTTGAAGACATTTAA
GTTGTTTTGTGGTGTTTTTAGAATCATCAAGTGTTTTTGAGTATTTTAAGCTTATTTCAA
TGCTCTTTTCATACTTTATTGCGTTTCCAAGTTATTTAAGATTGTTTTTGAATGTTTTGA
GCCCTTTTCATGAGTTTTTGATTGTTTTAAGGTCTATTCAAGGCTGTTTTCATAGTTTCT
TGCGGTTTGAATAGATTTGAGGTTGTTTTATAATGTTTTACGACTTTTCATGTGTTTTTA
AGTGTTTTAATGTCATTTCAAAGCTCATTTTCATAGTTTTTGTTGTTTTTAGCCCTTTTA
TGAGTTTTTGAGTGTTTTAAGGTGTATCAAAGCTCTTTTCATAGTTTCTTGCGTTTTAAA
GTAATTTGAGATTGTTTTATAATGTTTTAAGCCTTTTCAGGTTTAAGTGTTTTAAGGTCA
TTTCAAAGTCATTTTCA
```

1.3. GO enrichment Plots

a) *A. vesiculosa* Molecular Function

b) *A. vesiculosa* Cellular Compound

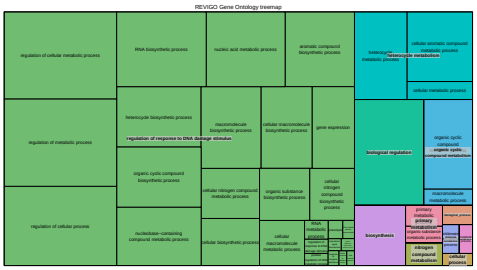
c) *D. muscipula* Molecular Function

d) *D. muscipula* Cellular Compound

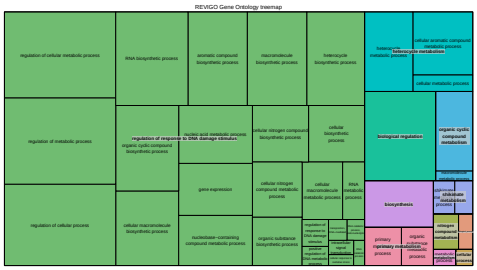
e) *D. spatulata* Molecular Function

f) *D. spatulata* Cellular Compound

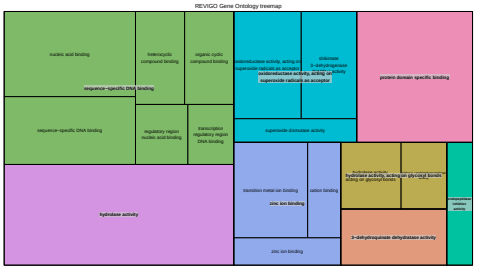
Figure VI.1.: **Treemaps of GO enrichments of the carnivorous orthogroups.**



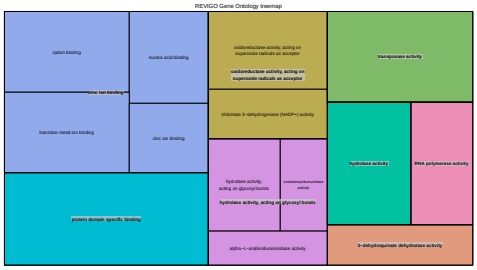
a) *A. vesiculosa* Biological Process



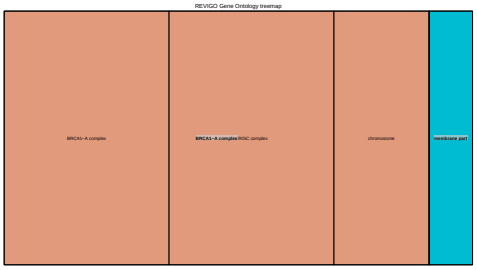
b) *D. muscipula* Biological Process



c) *A. vesiculosa* Molecular Function



d) *D. muscipula* Molecular Function



e) *A. vesiculosa* Cellular Compound

f) *D. muscipula* Cellular Compound

Figure VI.2.: **Treemaps of GO enrichments of the snap-trap orthogroups.**

2. Transcriptomics of Pollen Tube Guidance

2.1. Sequencing Libraries

Table VI.2.: **RNASeq libraries used for estimating the sequencing needs.** The libraries originate from the Tobacco genome sequencing project (Sierro et al. 2014).

Run ID	**Sample Name**	**Tissue**	**Size [Mbp]**
SRR955761	Ntab-TN90-L1	Leaf	10512
SRR955762	Ntab-TN90-L2	Leaf	15888
SRR955763	Ntab-TN90-L3	Leaf	11542
SRR955764	Ntab-TN90-R1	Root	9508
SRR955765	Ntab-TN90-R2	Root	8841
SRR955766	Ntab-TN90-R3	Root	10052
SRR1199197	Ntab-TN90-YF1	Immature Flower	9430
SRR1199198	Ntab-TN90-YF2	Immature Flower	8935
SRR1199199	Ntab-TN90-YF3	Immature Flower	7859
SRR1199069	Ntab-TN90-MF1	Mature Flower	9713
SRR1199070	Ntab-TN90-MF2	Mature Flower	8330
SRR1199071	Ntab-TN90-MF3	Mature Flower	8506
SRR1199124	Ntab-TN90-SF1	Senescent Flower	5519
SRR1199125	Ntab-TN90-SF2	Senescent Flower	4432

List of Figures

1.1. Structures of different repeat types 7

2.1. Drawing of the Venus Flytrap 9
2.2. Drawing of *A. vesiculosa* 11
2.3. Example of *D. spatulata* 12
2.4. Drawing of *N. tabacum* 13
2.5. Examples of different carnivorous plants 15
2.6. Examples of *D. muscipula* cultivars 16
2.7. Flower morphology of *N. tabacum* 17
2.8. Fertilization Process of Angiosperms 17

5.1. Quality assessment of the *A. vesiculosa* 3 Kbp jumping library 30
5.2. Comparison of centromere sequences 31
5.3. Enriched Biological Process GO terms in the carnivorous orthogroups. 32

6.1. Genome composition of *A. vesiculosa*, *D. muscipula* and *D. spatulata* 36

8.1. Overview of the root/leaf experiment 44
8.2. Overview of the flower experiment 45
8.3. PCA plot and Heatmap of the first RNASeq analysis 46
8.4. PCA plots of the Expressions in the different Tobacco conditions 47
8.5. Expression Profiles of CRPs in the different Tobacco samples 48

9.1. The reper workflow 52

VI.1. Treemaps of GO enrichments of the carnivorous orthogroups 63
VI.2. Treemaps of GO enrichments of the snap-trap orthogroups 64

List of Tables

4.1. Software and Databases used for assembly and annotation of the three carnivorous plant genomes 23

5.1. Overview of the final genome assemblies 25
5.2. Overview of the protein annotation 25
5.3. Overview of the centromere candidates 26
5.4. Fraction of repeats masked in the three genome assemblies 26
5.5. Results of the Heterozygosity analysis 27
5.6. Assembly statistics of the three carnivorous plant genomes using different assembly strategies 29
5.7. Overview of CDS annotation 29

7.1. RNASeq Libraries used in the tobacco project 40

8.1. Number of CRPs found in the different species 43

9.1. reper example output on class leve 51
9.2. reper example output on cluster level 51
9.3. reper example output on sequence level 52

VI.1. Sequencing libraries used in the carnivorous plants project 59
VI.2. RNASeq libraries used for estimating the sequencing needs 65

List of Source Codes

VI.1. "Centromere sequences of the three carnivorous plant genomes" 62

References

Adamec, L. (Mar. 2000). "Rootless Aquatic Plant Aldrovanda Vesiculosa: Physiological Polarity, Mineral Nutrition, and Importance of Carnivory". In: *Biologia Plantarum* 43.1, pp. 113–119. ISSN: 1573-8264. DOI: 10.1023/A:1026567300241. URL: https://doi.org/10.1023/A:1026567300241.

Adlassnig, W. et al. (Feb. 2011). "Traps of carnivorous pitcher plants as a habitat: composition of the fluid, biodiversity and mutualistic activities". In: *Annals of Botany* 107.2, pp. 181–194. ISSN: 1095-8290, 0305-7364. DOI: 10.1093/aob/mcq238. URL: https://academic.oup.com/aob/article-lookup/doi/10.1093/aob/mcq238 (visited on 02/15/2019).

Albert, V. et al. (Sept. 11, 1992). "Carnivorous plants: phylogeny and structural evolution". In: *Science* 257.5076, pp. 1491–1495. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1523408. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1523408 (visited on 09/28/2018).

Alonso-Blanco, C. et al. (July 2016). "1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana". In: *Cell* 166.2, pp. 481–491. ISSN: 00928674. DOI: 10.1016/j.cell.2016.05.063. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867416306675 (visited on 11/20/2018).

Amborella Genome Project et al. (Dec. 20, 2013). "The Amborella Genome and the Evolution of Flowering Plants". In: *Science* 342.6165, pp. 1241089–1241089. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1241089. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1241089 (visited on 11/22/2018).

Amien, S. et al. (June 1, 2010). "Defensin-Like ZmES4 Mediates Pollen Tube Burst in Maize via Opening of the Potassium Channel KZM1". In: *PLoS Biology* 8.6. Ed. by F. Berger, e1000388. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1000388. URL: https://dx.plos.org/10.1371/journal.pbio.1000388 (visited on 11/12/2018).

An, D. et al. (Jan. 18, 2018). "Isoform Sequencing and State-of-Art Applications for Unravelling Complexity of Plant Transcriptomes". In: *Genes* 9.1, p. 43. ISSN: 2073-4425. DOI: 10.3390/genes9010043. URL: http://www.mdpi.com/2073-4425/9/1/43 (visited on 12/13/2018).

Andrews, S. (2016). *FastQC: A quality control tool for high throughput sequence data.* URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Ankenbrand, M. J. et al. (2016). "TBro: visualization and management of *de novo* transcriptomes". In: *Database* 2016, baw146. ISSN: 1758-0463. DOI: 10.1093/database/baw146. URL: https://academic.oup.com/database/article-lookup/doi/10.1093/database/baw146 (visited on 02/02/2018).

Ashburner, M. et al. (May 2000). "Gene Ontology: tool for the unification of biology". In: *Nature Genetics* 25.1, pp. 25–29. ISSN: 1061-4036. DOI: 10.1038/75556. URL: http://www.nature.com/doifinder/10.1038/75556 (visited on 02/02/2018).

Bao, W. et al. (Dec. 2015). "Repbase Update, a database of repetitive elements in eukaryotic genomes". In: *Mobile DNA* 6.1. ISSN: 1759-8753. DOI: 10.1186/s13100-015-0041-9. URL: http://www.mobilednajournal.com/content/6/1/11 (visited on 09/26/2018).

Bauer, S. (June 28, 2016). *Ontologizer*. Version 2.1. URL: http://ontologizer.de.

Belser, C. et al. (Nov. 2018). "Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps". In: *Nature Plants* 4.11, pp. 879–887. ISSN: 2055-0278. DOI: 10.1038/s41477-018-0289-4. URL: http://www.nature.com/articles/s41477-018-0289-4 (visited on 04/03/2019).

Bemm, F. et al. (June 2016). "Venus flytrap carnivorous lifestyle builds on herbivore defense strategies". In: *Genome Research* 26.6, pp. 812–825. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.202200.115. URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.202200.115 (visited on 01/31/2018).

Benson, G. (Jan. 1, 1999). "Tandem repeats finder: a program to analyze DNA sequences". In: *Nucleic Acids Research* 27.2, pp. 573–580. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/27.2.573. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/27.2.573 (visited on 12/21/2017).

Bircheneder, S. and T. Dresselhaus (Aug. 2016). "Why cellular communication during plant reproduction is particularly mediated by CRP signalling". In: *Journal of Experimental Botany* 67.16, pp. 4849–4861. ISSN: 0022-0957, 1460-2431. DOI: 10.1093/jxb/erw271. URL: https://academic.oup.com/jxb/article-lookup/doi/10.1093/jxb/erw271 (visited on 11/12/2018).

Böhm, J. et al. (Feb. 2016). "The Venus Flytrap Dionaea muscipula Counts Prey-Induced Action Potentials to Induce Sodium Uptake". In: *Current Biology* 26.3, pp. 286–295. ISSN: 09609822. DOI: 10.1016/j.cub.2015.11.057. URL: https://linkinghub.elsevier.com/retrieve/pii/S0960982215015018 (visited on 09/28/2018).

Bolger, A. M. et al. (Aug. 1, 2014). "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15, pp. 2114–2120. ISSN: 1460-2059, 1367-4803. DOI: 10.1093/bioinformatics/btu170. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu170 (visited on 02/02/2018).

Brittnacher, J. (2018). *Growing pygmy Drosera*. International Carnivorous Plant Society. URL: http://www.carnivorousplants.org/grow/guides/PygmyDrosera (visited on 11/29/2018).

Butts, C. T. et al. (Oct. 2016). "Novel proteases from the genome of the carnivorous plant *Drosera capensis* : Structural prediction and comparative analysis: Novel Proteases from Drosera capensis". In: *Proteins: Structure, Function, and Bioinformatics* 84.10, pp. 1517–1533. ISSN: 08873585. DOI: 10.1002/prot.25095. URL: http://doi.wiley.com/10.1002/prot.25095 (visited on 11/27/2018).

Camacho, C. et al. (2009). "BLAST+: architecture and applications". In: *BMC Bioinformatics* 10.1, p. 421. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-421. URL: http://www.biomedcentral.com/1471-2105/10/421 (visited on 12/15/2017).

Campbell, M. S. et al. (Feb. 1, 2014). "MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations". In: *PLANT PHYSIOLOGY* 164.2, pp. 513–524. ISSN: 0032-0889, 1532-2548. DOI: 10.1104/pp.113.230144. URL: http://www.plantphysiol.org/cgi/doi/10.1104/pp.113.230144 (visited on 09/26/2018).

Cantarel, B. L. et al. (Nov. 21, 2007). "MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes". In: *Genome Research* 18.1, pp. 188–196. ISSN: 1088-9051. DOI: 10.1101/gr.6743907. URL: http://www.genome.org/cgi/doi/10.1101/gr.6743907 (visited on 10/23/2018).

Clavijo, B. J. et al. (May 2017). "An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations". In: *Genome Research* 27.5, pp. 885–896. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.217117.116. URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.217117.116 (visited on 12/15/2017).

Danecek, P. et al. (Aug. 1, 2011). "The variant call format and VCFtools". In: *Bioinformatics* 27.15, pp. 2156–2158. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btr330. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr330 (visited on 10/23/2018).

Darwin, C. (1875). *Insectivorous plants*. 1st ed. London: John Murray.

Dayhoff, M. O. and R. Eck (1966). *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation. URL: https://books.google.de/books?id=imoKAQAAMAAJ.

Degreef, J. D. (1997). "Fossil Aldrovanda". In: *Carnivorous Plant Newsletter* 26, pp. 93–97. URL: http://bestcarnivorousplants.com/aldrovanda/papers_online/Fossil.htm.

Dias, R. d. O. and O. L. Franco (Oct. 2015). "Cysteine-stabilized ab defensins: From a common fold to antibacterial activity". In: *Peptides* 72, pp. 64–72. ISSN: 01969781. DOI: 10.1016/j.peptides.2015.04.017. URL: https://linkinghub.elsevier.com/retrieve/pii/S0196978115001369 (visited on 03/26/2019).

Dobin, A. et al. (Jan. 2013). "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1, pp. 15–21. ISSN: 1460-2059, 1367-4803. DOI: 10.1093/bioinformatics/bts635. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts635 (visited on 10/23/2018).

Ellinghaus, D. et al. (2008). "LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons". In: *BMC Bioinformatics* 9.1, p. 18. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-18. URL: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-18 (visited on 05/27/2019).

Ellis, J. (1770). *Directions for Bringing over Seeds and Plants, from the East Indies and Other Distant Countries, in a State of Vegetation: Together with a Catalogue of Such Foreign Plants as Are Worthy of Being Encouraged in Our American Colonies, for the Purposes of Medicine, Agriculture, and Commerce. To Which is Added, the Figure*

and Botanical Description of a New Sensitive Plant, Called Dionaea muscipula: or, Venus's Fly-trap. L. Davis.

English, A. C. et al. (Nov. 21, 2012). "Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology". In: *PLoS ONE* 7.11. Ed. by Z. Liu, e47768. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0047768. URL: https://dx.plos.org/10.1371/journal.pone.0047768 (visited on 10/23/2018).

Feschotte, C. et al. (May 1, 2002). "Plant Transposable Elements: Where Genetics meets Genomics". In: *Nature Reviews Genetics* 3.5, pp. 329–341. ISSN: 14710056. DOI: 10.1038/nrg793. URL: http://www.nature.com/doifinder/10.1038/nrg793 (visited on 12/15/2017).

Finn, R. D. et al. (Jan. 4, 2016). "The Pfam protein families database: towards a more sustainable future". In: *Nucleic Acids Research* 44 (D1), pp. D279–D285. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkv1344. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1344 (visited on 02/02/2018).

Finn, R. D. et al. (Jan. 4, 2017). "InterPro in 2017—beyond protein family and domain annotations". In: *Nucleic Acids Research* 45 (D1), pp. D190–D199. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkw1107. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1107 (visited on 02/02/2018).

Forterre, Y. et al. (Jan. 2005). "How the Venus flytrap snaps". In: *Nature* 433.7024, pp. 421–425. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature03185. URL: http://www.nature.com/articles/nature03185 (visited on 02/19/2018).

Fracasso, A. et al. (Dec. 2016). "Drought stress tolerance strategies revealed by RNA-Seq in two sorghum genotypes with contrasting WUE". In: *BMC Plant Biology* 16.1. ISSN: 1471-2229. DOI: 10.1186/s12870-016-0800-x. URL: http://bmcplantbiol.biomedcentral.com/articles/10.1186/s12870-016-0800-x (visited on 11/22/2018).

Freund, M. (Feb. 2019). "Comparative genomics of carnivorous Droseraceae". Master Thesis. Universität Würzburg.

Fu, L. et al. (Dec. 2012). "CD-HIT: accelerated for clustering the next-generation sequencing data". In: *Bioinformatics* 28.23, pp. 3150–3152. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/bts565. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts565 (visited on 12/15/2017).

Fukushima, K. et al. (Feb. 6, 2017). "Genome of the pitcher plant Cephalotus reveals genetic changes associated with carnivory". In: *Nature Ecology & Evolution* 1.3, p. 0059. ISSN: 2397-334X. DOI: 10.1038/s41559-016-0059. URL: http://www.nature.com/articles/s41559-016-0059 (visited on 11/27/2018).

Gaume, L. et al. (Mar. 2016). "Different pitcher shapes and trapping syndromes explain resource partitioning in *Nepenthes* species". In: *Ecology and Evolution* 6.5, pp. 1378–1392. ISSN: 20457758. DOI: 10.1002/ece3.1920. URL: http://doi.wiley.com/10.1002/ece3.1920 (visited on 11/28/2018).

Gibson, T. C. and D. M. Waller (Aug. 2009). "Evolving Darwin's 'most wonderful' plant: ecological steps to a snap-trap". In: *New Phytologist* 183.3, pp. 575–587. ISSN: 0028646X, 14698137. DOI: 10.1111/j.1469-8137.2009.02935.x. URL: http://doi.wiley.com/10.1111/j.1469-8137.2009.02935.x (visited on 11/28/2018).

Goff, L. and C. Trapnell (2017). *cummeRbund*. DOI: 10.18129/B9.bioc.cummeRbund.

Gonella, P. M. et al. (July 24, 2015). "Drosera magnifica (Droseraceae): the largest New World sundew, discovered on Facebook". In: *Phytotaxa* 220.3, p. 257. ISSN: 1179-3163, 1179-3155. DOI: 10.11646/phytotaxa.220.3.4. URL: https://biotaxa.org/Phytotaxa/article/view/phytotaxa.220.3.4 (visited on 11/29/2018).

Goubert, C. et al. (Apr. 2015). "De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes aegypti)". In: *Genome Biology and Evolution* 7.4, pp. 1192–1205. ISSN: 1759-6653. DOI: 10.1093/gbe/evv050. URL: https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evv050 (visited on 12/15/2017).

Grabherr, M. G. et al. (May 15, 2011). "Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data". In: *Nature Biotechnology* 29.7, pp. 644–652. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.1883. URL: http://www.nature.com/doifinder/10.1038/nbt.1883 (visited on 12/15/2017).

Hackl, T. (2016). "A draft genome for the Venus flytrap, Dionaea muscipula: Evaluation of assembly strategies for a complex Genome – Development of novel approaches and bioinformatics solutions". Dissertation. Würzburg: Universität Würzburg. URL: https://opus.bibliothek.uni-wuerzburg.de/frontdoor/index/index/docId/13314 (visited on 02/19/2018).

Hedrich, R. and E. Neher (Mar. 2018). "Venus Flytrap: How an Excitable, Carnivorous Plant Works". In: *Trends in Plant Science* 23.3, pp. 220–234. ISSN: 13601385. DOI: 10.1016/j.tplants.2017.12.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S1360138517302807 (visited on 11/30/2018).

Heubl, G. et al. (June 30, 2006). "Molecular Phylogeny and Character Evolution of Carnivorous Plant Families in Caryophyllales — Revisited". In: *Plant Biology* 8.6, pp. 821–830. DOI: 10.1055/s-2006-924460. URL: https://onlinelibrary.wiley.com/doi/abs/10.1055/s-2006-924460.

Hodick, D. and A. Sievers (Aug. 1989). "On the mechanism of trap closure of Venus flytrap (Dionaea muscipula Ellis)". In: *Planta* 179.1, pp. 32–42. ISSN: 0032-0935, 1432-2048. DOI: 10.1007/BF00395768. URL: http://link.springer.com/10.1007/BF00395768 (visited on 11/30/2018).

Hoshi, Y. et al. (2006). "Nucleotide sequence variation was unexpectedly low in an endangered species, Aldrovanda vesiculosa L. (Droseraceae)". In: *Chromosome Botany* 1.1, pp. 27–32. ISSN: 1881-5936, 1881-8285. DOI: 10.3199/iscb.1.27. URL: http://joi.jlc.jst.go.jp/JST.JSTAGE/iscb/1.27?from=CrossRef (visited on 09/28/2018).

Huang, Q. et al. (June 2015). "Active role of small peptides in *Arabidopsis* reproduction: Expression evidence: Active role of small peptides in Arabidopsis reproduction". In: *Journal of Integrative Plant Biology* 57.6, pp. 518–521. ISSN: 16729072. DOI: 10.1111/jipb.12356. URL: http://doi.wiley.com/10.1111/jipb.12356 (visited on 09/29/2018).

Ibarra-Laclette, E. et al. (June 2013). "Architecture and evolution of a minute plant genome". In: *Nature* 498.7452, pp. 94–98. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/

nature12132. URL: http://www.nature.com/articles/nature12132 (visited on 11/27/2018).

IUCN (June 30, 2000). *Dionaea muscipula: Schnell, D., Catling, P., Folkerts, G., Frost, C., Gardner, R. & et al.: The IUCN Red List of Threatened Species 2000: e.T39636A10253384*. type: dataset. International Union for Conservation of Nature. DOI: 10.2305/IUCN.UK.2000.RLTS.T39636A10253384.en. URL: http://www.iucnredlist.org/details/39636/0 (visited on 02/19/2018).

IUCN (Apr. 30, 2012). *Aldrovanda vesiculosa: Cross, A.: The IUCN Red List of Threatened Species 2012: e.T162346A901031*. type: dataset. International Union for Conservation of Nature. DOI: 10.2305/IUCN.UK.2012.RLTS.T162346A901031.en. URL: http://www.iucnredlist.org/details/162346/0 (visited on 09/28/2018).

Jensen, M. K. et al. (Apr. 17, 2015). "Transcriptome and Genome Size Analysis of the Venus Flytrap". In: *PLOS ONE* 10.4. Ed. by L. Zane, e0123887. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0123887. URL: http://dx.plos.org/10.1371/journal.pone.0123887 (visited on 02/19/2018).

Jones, P. et al. (May 1, 2014). "InterProScan 5: genome-scale protein function classification". In: *Bioinformatics* 30.9, pp. 1236–1240. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btu031. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu031 (visited on 02/02/2018).

Kawahara, Y. et al. (2013). "Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data". In: *Rice* 6.1, p. 4. ISSN: 1939-8433. DOI: 10.1186/1939-8433-6-4. URL: http://thericejournal.springeropen.com/articles/10.1186/1939-8433-6-4 (visited on 11/12/2018).

Keller, O. et al. (Mar. 15, 2011). "A novel hybrid gene prediction method employing protein multiple sequence alignments". In: *Bioinformatics* 27.6, pp. 757–763. ISSN: 1460-2059, 1367-4803. DOI: 10.1093/bioinformatics/btr010. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr010 (visited on 10/23/2018).

Kim, D. et al. (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". In: *Genome Biology* 14.4, R36. ISSN: 1465-6906. DOI: 10.1186/gb-2013-14-4-r36. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-4-r36 (visited on 02/02/2018).

Koren, S. et al. (May 2017). "Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation". In: *Genome Research* 27.5, pp. 722–736. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.215087.116. URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.215087.116 (visited on 10/23/2018).

Korf, I. (May 2004). "Gene finding in novel genomes". In: *BMC Bioinformatics* 5.1, p. 59. ISSN: 1471-2105. DOI: 10.1186/1471-2105-5-59. URL: https://doi.org/10.1186/1471-2105-5-59.

Krausko, M. et al. (Mar. 2017). "The role of electrical and jasmonate signalling in the recognition of captured prey in the carnivorous sundew plant *Drosera capensis*". In: *New Phytologist* 213.4, pp. 1818–1835. ISSN: 0028646X. DOI: 10.1111/nph.14352. URL: http://doi.wiley.com/10.1111/nph.14352 (visited on 11/28/2018).

Lan, T. et al. (May 30, 2017). "Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome". In: *Proceedings of the National Academy of Sciences* 114.22, E4435–E4441. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1702072114. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1702072114 (visited on 02/07/2019).

Langmead, B. and S. L. Salzberg (Mar. 4, 2012). "Fast gapped-read alignment with Bowtie 2". In: *Nature Methods* 9.4, pp. 357–359. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.1923. URL: http://www.nature.com/doifinder/10.1038/nmeth.1923 (visited on 12/15/2017).

Laverty, K. U. et al. (Jan. 2019). "A physical and genetic map of *Cannabis sativa* identifies extensive rearrangements at the THC/CBD acid synthase loci". In: *Genome Research* 29.1, pp. 146–156. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.242594.118. URL: http://genome.cshlp.org/lookup/doi/10.1101/gr.242594.118 (visited on 04/03/2019).

Leushkin, E. V. et al. (2013). "The miniature genome of a carnivorous plant Genlisea aurea contains a low number of genes and short non-coding sequences". In: *BMC Genomics* 14.1, p. 476. ISSN: 1471-2164. DOI: 10.1186/1471-2164-14-476. URL: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-14-476 (visited on 11/27/2018).

Li, H. et al. (Aug. 15, 2009). "The Sequence Alignment/Map format and SAMtools". In: *Bioinformatics* 25.16, pp. 2078–2079. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btp352. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352 (visited on 12/15/2017).

Li, W. and A. Godzik (July 1, 2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences". In: *Bioinformatics* 22.13, pp. 1658–1659. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btl158. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btl158 (visited on 12/15/2017).

Lohse, M. et al. (May 2014). "Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data: Mercator: sequence functional annotation server". In: *Plant, Cell & Environment* 37.5, pp. 1250–1258. ISSN: 01407791. DOI: 10.1111/pce.12231. URL: http://doi.wiley.com/10.1111/pce.12231 (visited on 09/26/2018).

Love, M. I. et al. (Dec. 2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8 (visited on 02/02/2018).

Luo, R. et al. (Dec. 2012). "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler". In: *GigaScience* 1.1. ISSN: 2047-217X. DOI: 10.1186/2047-217X-1-18. URL: https://academic.oup.com/gigascience/article-lookup/doi/10.1186/2047-217X-1-18 (visited on 10/23/2018).

Marçais, G. and C. Kingsford (Mar. 15, 2011). "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers". In: *Bioinformatics* 27.6, pp. 764–770. ISSN: 1460-2059, 1367-4803. DOI: 10.1093/bioinformatics/btr011. URL: https://

academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr011 (visited on 12/15/2017).

Marshall, E. et al. (Mar. 1, 2011). "Cysteine-Rich Peptides (CRPs) mediate diverse aspects of cell-cell communication in plant reproduction and development". In: *Journal of Experimental Botany* 62.5, pp. 1677–1686. ISSN: 0022-0957, 1460-2431. DOI: 10.1093/jxb/err002. URL: https://academic.oup.com/jxb/article-lookup/doi/10.1093/jxb/err002 (visited on 02/19/2019).

Melters, D. P. et al. (2013). "Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution". In: *Genome Biology* 14.1, R10. ISSN: 1465-6906. DOI: 10.1186/gb-2013-14-1-r10. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-1-r10 (visited on 09/26/2018).

Mendel, G. (1866). *Versuche über Pflanzen-Hybriden*. Brünn : Im Verlage des Vereines, DOI: 10.5962/bhl.title.61004. URL: http://www.biodiversitylibrary.org/bibliography/61004 (visited on 11/20/2018).

Monaco, M. K. et al. (Jan. 2014). "Gramene 2013: comparative plant genomics resources". In: *Nucleic Acids Research* 42 (D1), pp. D1193–D1199. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkt1110. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1110 (visited on 11/09/2018).

Montenegro, J. D. et al. (June 2017). "The pangenome of hexaploid bread wheat". In: *The Plant Journal* 90.5, pp. 1007–1013. ISSN: 09607412. DOI: 10.1111/tpj.13515. URL: http://doi.wiley.com/10.1111/tpj.13515 (visited on 12/15/2017).

Nakano, K. et al. (July 2017). "Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area". In: *Human Cell* 30.3, pp. 149–161. ISSN: 1749-0774. DOI: 10.1007/s13577-017-0168-8. URL: http://link.springer.com/10.1007/s13577-017-0168-8 (visited on 02/22/2019).

Nasrallah, J. B. and M. E. Nasrallah (Apr. 1, 2014). "*S* -locus receptor kinase signalling". In: *Biochemical Society Transactions* 42.2, pp. 313–319. ISSN: 0300-5127, 1470-8752. DOI: 10.1042/BST20130222. URL: http://biochemsoctrans.org/lookup/doi/10.1042/BST20130222 (visited on 03/04/2019).

NHGRI (June 7, 2016). *The Cost of Sequencing a Human Genome*. URL: https://www.genome.gov/sequencingcosts/ (visited on 10/22/0208).

Nowoshilow, S. et al. (Jan. 24, 2018). "The axolotl genome and the evolution of key tissue formation regulators". In: *Nature* 554.7690, pp. 50–55. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature25458. URL: http://www.nature.com/doifinder/10.1038/nature25458 (visited on 11/22/2018).

Nurk, S. et al. (2013). "Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads". In: *Research in Computational Molecular Biology*. Ed. by M. Deng et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 158–170. ISBN: 978-3-642-37195-0.

Nussbaumer, T. et al. (Nov. 29, 2012). "MIPS PlantsDB: a database framework for comparative plant genome research". In: *Nucleic Acids Research* 41 (D1), pp. D1144–D1151. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gks1153. URL: http://academic.oup.com/nar/article/41/D1/D1144/1063372/MIPS-PlantsDB-a-database-framework-for-comparative (visited on 12/15/2017).

Nystedt, B. et al. (May 2013). "The Norway spruce genome sequence and conifer genome evolution". In: *Nature* 497.7451, pp. 579–584. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature12211. URL: http://www.nature.com/articles/nature12211 (visited on 11/22/2018).

O'Leary, N. A. et al. (Jan. 4, 2016). "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". In: *Nucleic Acids Research* 44 (D1), pp. D733–D745. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkv1189. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1189 (visited on 12/15/2017).

Okuda, S. et al. (Mar. 2009). "Defensin-like polypeptide LUREs are pollen tube attractants secreted from synergid cells". In: *Nature* 458.7236, pp. 357–361. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature07882. URL: http://www.nature.com/articles/nature07882 (visited on 03/04/2019).

Patro, R. et al. (Apr. 2017). "Salmon provides fast and bias-aware quantification of transcript expression". In: *Nature Methods* 14.4, pp. 417–419. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.4197. URL: http://www.nature.com/articles/nmeth.4197 (visited on 02/02/2018).

Pavlovič, A. and A. Mithöfer (Apr. 24, 2019). "Jasmonate signalling in carnivorous plants: copycat of plant defence mechanisms". In: *Journal of Experimental Botany*. ISSN: 0022-0957, 1460-2431. DOI: 10.1093/jxb/erz188. URL: https://academic.oup.com/jxb/advance-article/doi/10.1093/jxb/erz188/5479295 (visited on 05/24/2019).

Payne, A. et al. (Jan. 1, 2018). "Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files." In: *bioRxiv*. DOI: 10.1101/312256. URL: http://biorxiv.org/content/early/2018/05/03/312256.abstract.

Poppinga, S. et al. (Sept. 26, 2012). "Catapulting Tentacles in a Sticky Carnivorous Plant". In: *PLoS ONE* 7.9. Ed. by P. V. A. Fine, e45735. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0045735. URL: https://dx.plos.org/10.1371/journal.pone.0045735 (visited on 02/15/2019).

Poppinga, S. et al. (July 2013). "Trap diversity and evolution in the family Droseraceae". In: *Plant Signaling & Behavior* 8.7, e24685. ISSN: 1559-2324. DOI: 10.4161/psb.24685. URL: http://www.tandfonline.com/doi/abs/10.4161/psb.24685 (visited on 11/28/2018).

Pryszcz, L. P. and T. Gabaldón (July 8, 2016). "Redundans: an assembly pipeline for highly heterozygous genomes". In: *Nucleic Acids Research* 44.12, e113–e113. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkw294. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw294 (visited on 10/23/2018).

Purves, W. K., ed. (2006). *Biologie*. 7. Aufl. OCLC: 836620362. München: Elsevier, Spektrum Akad. Verl. 1577 pp. ISBN: 978-3-8274-1630-8.

Qin, C. et al. (Apr. 8, 2014). "Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization". In: *Proceedings of the National Academy of Sciences* 111.14, pp. 5135–5140. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1400975111. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1400975111 (visited on 11/13/2018).

Qu, L.-J. et al. (Aug. 2015). "Peptide signalling during the pollen tube journey and double fertilization". In: *Journal of Experimental Botany* 66.17, pp. 5139–5150. ISSN: 0022-0957, 1460-2431. DOI: 10.1093/jxb/erv275. URL: https://academic.oup.com/jxb/article-lookup/doi/10.1093/jxb/erv275 (visited on 03/04/2019).

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: https://www.R-project.org/.

Ruby, J. G. et al. (May 2013). "PRICE: Software for the Targeted Assembly of Components of (Meta) Genomic Sequence Data". In: *G3: Genes|Genomes|Genetics* 3.5, pp. 865–880. ISSN: 2160-1836. DOI: 10.1534/g3.113.005967. URL: http://g3journal.org/lookup/doi/10.1534/g3.113.005967 (visited on 10/23/2018).

Sanger, F. et al. (Dec. 1, 1977). "DNA sequencing with chain-terminating inhibitors". In: *Proceedings of the National Academy of Sciences* 74.12, pp. 5463–5467. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.74.12.5463. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.74.12.5463 (visited on 11/20/2018).

Saul, F. (Jan. 2019). "Tracing the evolution of carnivorous Droseraceae - Comparative genomics of Dionaea muscipula, Aldrovanda vesiculosa and Drosera spatulata". Master Thesis. Universität Würzburg.

El-Sayed, A. M. et al. (Aug. 2016). "Pollinator-prey conflicts in carnivorous plants: When flower and trap properties mean life or death". In: *Scientific Reports* 6.1. ISSN: 2045-2322. DOI: 10.1038/srep21065. URL: http://www.nature.com/articles/srep21065 (visited on 11/28/2018).

Schmeil, O. (1911). *Lehrbuch der Botanik*. 28th ed. Verlag von Quelle und Meyer, Leipzig. 534 pp.

Schnable, P. S. et al. (Nov. 20, 2009). "The B73 Maize Genome: Complexity, Diversity, and Dynamics". In: *Science* 326.5956, pp. 1112–1115. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1178534. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1178534 (visited on 11/12/2018).

Shirakawa, J. et al. (2011). "Chromosome differentiation and genome organization in carnivorous plant family Droseraceae". In: *Chromosome Botany* 6.4, pp. 111–119. ISSN: 1881-5936, 1881-8285. DOI: 10.3199/iscb.6.111. URL: http://joi.jlc.jst.go.jp/JST.JSTAGE/iscb/6.111?from=CrossRef (visited on 02/19/2018).

Sierro, N. et al. (Dec. 2014). "The tobacco genome sequence and its comparison with those of tomato and potato". In: *Nature Communications* 5.1. ISSN: 2041-1723. DOI: 10.1038/ncomms4833. URL: http://www.nature.com/articles/ncomms4833 (visited on 10/04/2018).

Silverstein, K. A. et al. (July 2007). "Small cysteine-rich peptides resembling antimicrobial peptides have been under-predicted in plants: *Under-predicted cysteine-rich peptides in plants*". In: *The Plant Journal* 51.2, pp. 262–280. ISSN: 09607412. DOI: 10.1111/j.1365-313X.2007.03136.x. URL: http://doi.wiley.com/10.1111/j.1365-313X.2007.03136.x (visited on 09/29/2018).

Simão, F. A. et al. (Oct. 1, 2015). "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs". In: *Bioinformatics* 31.19, pp. 3210–3212. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btv351. URL:

https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv351 (visited on 10/23/2018).

Smit, A. and R. Hubley (2008). *RepeatModeler Open-1.0*. URL: http://www.repeatmasker.org.

Smit, A. et al. (2013). *RepeatMasker Open-4.0*. URL: http://www.repeatmasker.org.

Stanke, M. and S. Waack (Sept. 27, 2003). "Gene prediction with a hidden Markov model and a new intron submodel". In: *Bioinformatics* 19 (Suppl 2), pp. ii215–ii225. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btg1080. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btg1080 (visited on 10/23/2018).

Steppuhn, A. et al. (Aug. 17, 2004). "Nicotine's Defensive Function in Nature". In: *PLoS Biology* 2.8. Ed. by Michael Levine, e217. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0020217. URL: https://dx.plos.org/10.1371/journal.pbio.0020217 (visited on 04/04/2019).

Supek, F. et al. (July 18, 2011). "REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms". In: *PLoS ONE* 6.7. Ed. by C. Gibas, e21800. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0021800. URL: https://dx.doi.org/10.1371/journal.pone.0021800 (visited on 10/23/2018).

Takayama, S. et al. (Oct. 2001). "Direct ligand–receptor complex interaction controls Brassica self-incompatibility". In: *Nature* 413.6855, pp. 534–538. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/35097104. URL: http://www.nature.com/articles/35097104 (visited on 03/04/2019).

Terhoeven, N. (Aug. 28, 2014). "New approaches for repeat content and structure analysis in the complex genome of the Venus Flytrap Dionaea muscipula". Master Thesis. Universität Würzburg.

Terhoeven, N. et al. (Feb. 8, 2018). "reper: Genome-wide identification, classification and quantification of repetitive elements without an assembled genome". In: *The Journal of Open Source Software* 3.22, p. 527. ISSN: 2475-9066. DOI: 10.21105/joss.00527. URL: http://joss.theoj.org/papers/10.21105/joss.00527 (visited on 02/09/2018).

Tewksbury, J. J. et al. (Aug. 19, 2008). "Evolutionary ecology of pungency in wild chilies". In: *Proceedings of the National Academy of Sciences* 105.33, pp. 11808–11811. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0802691105. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.0802691105 (visited on 03/25/2019).

Tewksbury, J. J. and G. P. Nabhan (July 2001). "Directed deterrence by capsaicin in chillies". In: *Nature* 412.6845, pp. 403–404. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/35086653. URL: http://www.nature.com/articles/35086653 (visited on 03/25/2019).

The Brassica rapa Genome Sequencing Project Consortium et al. (Oct. 2011). "The genome of the mesopolyploid crop species Brassica rapa". In: *Nature Genetics* 43.10, pp. 1035–1039. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.919. URL: http://www.nature.com/articles/ng.919 (visited on 11/12/2018).

The Gene Ontology Consortium (Jan. 4, 2017). "Expansion of the Gene Ontology knowledgebase and resources". In: *Nucleic Acids Research* 45 (D1), pp. D331–D338. ISSN:

0305-1048, 1362-4962. DOI: 10.1093/nar/gkw1108. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1108 (visited on 02/02/2018).

The UniProt Consortium (Jan. 4, 2017). "UniProt: the universal protein knowledgebase". In: *Nucleic Acids Research* 45 (D1), pp. D158–D169. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkw1099. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1099 (visited on 10/23/2018).

Trapnell, C. et al. (May 2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". In: *Nature Biotechnology* 28.5, pp. 511–515. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.1621. URL: http://www.nature.com/articles/nbt.1621 (visited on 09/26/2018).

Trapnell, C. et al. (Mar. 1, 2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks". In: *Nature Protocols* 7.3, pp. 562–578. ISSN: 1754-2189, 1750-2799. DOI: 10.1038/nprot.2012.016. URL: http://www.nature.com/doifinder/10.1038/nprot.2012.016 (visited on 11/05/2018).

Usadel, B. et al. (Sept. 2009). "A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize". In: *Plant, Cell & Environment* 32.9, pp. 1211–1229. ISSN: 01407791, 13653040. DOI: 10.1111/j.1365-3040.2009.01978.x. URL: http://doi.wiley.com/10.1111/j.1365-3040.2009.01978.x (visited on 09/26/2018).

Van Bel, M. et al. (Feb. 1, 2012). "Dissecting Plant Genomes with the PLAZA Comparative Genomics Platform". In: *PLANT PHYSIOLOGY* 158.2, pp. 590–600. ISSN: 0032-0889, 1532-2548. DOI: 10.1104/pp.111.189514. URL: http://www.plantphysiol.org/cgi/doi/10.1104/pp.111.189514 (visited on 02/02/2018).

Van Bel, M. et al. (2013). "TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes". In: *Genome Biology* 14.12, R134. ISSN: 1465-6906. DOI: 10.1186/gb-2013-14-12-r134. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-12-r134 (visited on 02/02/2018).

Vasilinetc, I. et al. (Oct. 15, 2015). "Assembling short reads from jumping libraries with large insert sizes". In: *Bioinformatics* 31.20, pp. 3262–3268. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btv337. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv337 (visited on 10/23/2018).

Veleba, A. et al. (Feb. 2017). "Evolution of genome size and genomic GC content in carnivorous holokinetics (Droseraceae)". In: *Annals of Botany* 119.3, pp. 409–416. ISSN: 0305-7364, 1095-8290. DOI: 10.1093/aob/mcw229. URL: https://academic.oup.com/aob/article-lookup/doi/10.1093/aob/mcw229 (visited on 02/19/2018).

Walker, B. J. et al. (Nov. 19, 2014). "Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement". In: *PLoS ONE* 9.11. Ed. by J. Wang, e112963. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0112963. URL: https://dx.plos.org/10.1371/journal.pone.0112963 (visited on 10/23/2018).

Wang, Y. et al. (Dec. 2005). "Characteristics of the tomato nuclear genome as determined by sequencing undermethylated EcoRI digested fragments". In: *Theoretical and Ap-*

plied Genetics 112.1, pp. 72–84. ISSN: 0040-5752, 1432-2242. DOI: 10.1007/s00122-005-0107-z. URL: http://link.springer.com/10.1007/s00122-005-0107-z (visited on 11/12/2018).

Waterhouse, R. M. et al. (Mar. 1, 2018). "BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics". In: *Molecular Biology and Evolution* 35.3, pp. 543–548. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msx319. URL: https://academic.oup.com/mbe/article/35/3/543/4705839 (visited on 10/23/2018).

Westermeier, A. S. et al. (May 16, 2018). "How the carnivorous waterwheel plant (*Aldrovanda vesiculosa*) snaps". In: *Proceedings of the Royal Society B: Biological Sciences* 285.1878, p. 20180012. ISSN: 0962-8452, 1471-2954. DOI: 10.1098/rspb.2018.0012. URL: http://rspb.royalsocietypublishing.org/lookup/doi/10.1098/rspb.2018.0012 (visited on 11/28/2018).

Wheeler, G. L. and B. C. Carstens (Jan. 31, 2018). "Evaluating the adaptive evolutionary convergence of carnivorous plant taxa through functional genomics". In: *PeerJ* 6, e4322. ISSN: 2167-8359. DOI: 10.7717/peerj.4322. URL: https://peerj.com/articles/4322 (visited on 01/11/2019).

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: http://ggplot2.org.

Youngsteadt, E. et al. (Apr. 2018). "Venus Flytrap Rarely Traps Its Pollinators". In: *The American Naturalist* 191.4, pp. 539–546. ISSN: 0003-0147, 1537-5323. DOI: 10.1086/696124. URL: https://www.journals.uchicago.edu/doi/10.1086/696124 (visited on 02/13/2019).

Zaman, M. et al. (July 13, 2011). "Ecology, morphology and anatomy of *Aldrovanda vesiculosa* L. (Droseraceae) from Bangladesh". In: *Bangladesh Journal of Botany* 40.1. ISSN: 2079-9926, 0253-5416. DOI: 10.3329/bjb.v40i1.8002. URL: http://www.banglajol.info/index.php/BJB/article/view/8002 (visited on 11/28/2018).

Zimmerman, J. L. and R. B. Goldberg (1977). "DNA sequence organization in the genome of Nicotiana tabacum". In: *Chromosoma* 59.3, pp. 227–252. ISSN: 0009-5915, 1432-0886. DOI: 10.1007/BF00292780. URL: http://link.springer.com/10.1007/BF00292780 (visited on 10/04/2018).

Acknowledgement

First of all, I want to thank my primary supervisor Jörg Schultz. You were a great mentor and supervisor since I started my Bachelor course eight years ago. Thank you for creating a great work environment and your scientific input and guidance during all this time.

Further, I want to thank my supervisors from the Botany department, Rainer Hedrich and Dirk Becker. You provided excellent supervision and lots of valuable input for my work. I also want to thank you for giving me the opportunity to work on my projects.

I also want to thank Markus Ankenbrand, Alexander Keller and Frank Förster. You accompanied me through most of my academic career as colleagues, supervisors and friends. It was a great pleasure working with you and I will never forget our interesting discussions during several lunch and coffee breaks.

A big thank you also to all other members of the carnivore project, especially "my" Masters students Franziska Saul and Matthias Freund. It was a great time working with you. Also Gergő Pálfalvi, Thomas Hackl, Anda Iosip and Ines Kreuzer. Thank you very much for providing me with various information, data and general input for this work. You are great colleagues and it was a pleasure working with you in this large, collaborative project.

Regarding the tobacco project, I want to thank Katharina von Meyer. It was great working with you. Another thank you goes to all my collaborators and co-authors on various other projects, Sonja Hohlfeld, Simon Pfaff, Jan Freudenthal, Ludwig Leidinger and many more.

Further, I want to thank and all other members of the Botany department and the CCTB. You created a great work environment and many interesting discussions during seminars, WUBSyB, HackyHour and coffee breaks.

Last, but not least, I want to thank my family and friends. Your constant support and motivation was a huge help in completing this thesis. Special thanks to my wife Inge. I couldn't have done this without you.

Publication List

in preparation	Author list not final, yet Genomes of the Venus Flytrap and close relatives spot the roots of plant carnivory
in preparation	Meyer K von, **Terhoeven N**, Schultz J et al. Transcriptomic Analyses of Pollen Tube Guidance in Tobacco shed light on the role of DEFLs
in preparation	Freudenthal J, Pfaff S, **Terhoeven N**, et al. The Landscape of Chloroplast Assembly Tools
2018	**Terhoeven N**, Schultz J, Hackl T reper: Genome-wide identification, classification and quantification of repetitive elements without an assembled genome. Journal of Open Source Software, 3(22), 527
2018	Ankenbrand MJ, Pfaff S, **Terhoeven N**, et al. chloroExtractor: extraction and assembly of the chloroplast genome from whole genome shotgun data. Journal of Open Source Software, 3(21), 464A
2017	Ankenbrand MJ, **Terhoeven N**, Hohlfeld S, et al. biojs-io-biom, a BioJS component for handling data in Biological Observation Matrix (BIOM) format.[version 2; referees: 1 approved, 2 approved with reservations]. F1000Research 2017, 5:2348
2013	Schultz, J, **Terhoeven, N** The bilaterian roots of cordon-bleu., BMC Res. Notes 6, 393A

Curriculum Vitæ

CV omitted for online publication.

Affidavit / Eidesstattliche Erklärung

Affidavit

I hereby confirm that my thesis entitled "Genomics of carnivorous Droseraceae and Transcriptomics of Tobacco pollination as case studies for neofunctionalisation of plant defence mechanisms" is the result of my own work. I did not receive any help or support from commercial consultants. All sources and/or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor similar form.

Place, Date	Signature

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, die Dissertation "Genomik karnivorer Droseraceae und Transkriptomik der Befruchtung von Tabak als Fallstudien zur Umfunktionierung pflanzlicher Verteidigungsmechanismenëigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Ort, Datum	Unterschrift