Plant Carnivory and the Evolution of Novelty in *Sarracenia alata*

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in the Graduate School of the Ohio State University

By

Gregory Lawrence Wheeler, MS

Graduate Program in Evolution, Ecology, and Organismal Biology

The Ohio State University

2018

Dissertation Committee:

Dr. Bryan C. Carstens, Adviser

Dr. Marymegan Daly

Dr. Zakee Sabree

Dr. Andrea Wolfe

Abstract

Most broadly, this study aimed to develop a better understanding of how organisms evolve novel functions and traits, and examine how seemingly complex adaptive trait syndromes can convergently evolve. As an ideal example of this, the carnivorous plants were chosen. This polyphyletic grouping contains taxa derived from multiple independent evolutionary origins, in at least five plant orders, and has resulted in striking convergence of niche and morphology.

First, a database study was performed, with the goal of understanding the evolutionary trends that impact carnivorous plants as a whole. Using carnivorous and non-carnivorous plant genomes available from GenBank. An *a priori* list of Gene Ontology-coded functions implicated in plant carnivory by earlier studies was constructed via literature review. Experimental and control samples were tested for statistical overrepresentation of these functions. It was found that, while some functions were significant in some taxa, there was no overall shared signal of plant carnivory, with each taxon presumably having selected for a different subset of these functions.

Next, analyses were performed that targeted *Sarracenia alata* specifically. A reference genome for *S. alata* was assembled using PacBio, Illumina, and BioNano data and annotated using MAKER-P with additional preliminary database filtration. From these, it was found that *Sarracenia alata* possesses significant and substantial overrepresentation of genes with functions associated with plant carnivory, at odds with the hypothesis that the plant primarily relies on symbioses.

Finally, pitcher fluid was collected from *S. alata* in the field. RNA was extracted from the fluid, sequenced via Illumina, and assembled with Trinity. Sequences were sorted into host plant and microbiome based on BLAST match to the *S. alata* reference genome. It was found that, while *S. alata* contributes two-thirds of the transcripts, these encode no digestive enzymes and a very limited set of transport channel proteins; however, these functions were identified in microbe-originated transcripts. A large portion of *S. alata*'s transcripts were instead found to encode anti-microbial peptides (AMPs). These short proteins are known to play a role in modulating gut microflora in animals, and while they are documented in plants, their role had never been addressed in carnivorous plants.

From these findings, I have concluded that there are a large number of evolutionary paths that lead to highly similar adaptive strategies; specific relevant functions may be identified, but the subset of these used by a given lineage will vary greatly. In *Sarracenia alata* specifically, it appears that at one point in time there was strong selection favoring the retention of genes associated with prey digestion. However, at present these do not appear to be expressed, with microbial symbioses instead responsible for the bulk of the digestive process. Instead, the plant has likely evolved to specialize in a regulatory role, modulating the microbial composition of its fluid via the production of AMPs. This shows that not only do lineages evolve via different pathways, but that the same lineage may change its adaptive specialization at different points in its history.

Acknowledgments

I would like to begin with a general thanks not only to the members of my department, but also to the members of the community. Without the outpouring of support I've received following certain unexpected setbacks, I doubt I could have managed the completion of this program. The experience has certainly reaffirmed my decision to stay in Columbus, OH long-term, and I hope I can find some way to give back to others the kind of assistance I've received.

I give my thanks to Dr. Bryan Carstens, my advisor, for his mentorship, support, and tolerance. He trusted my judgement and ability in undertaking this complex project, and continued to support and accommodate me as I began the pursuit of a career that took me out of the lab. Thanks also to Dr. Andrea Wolfe, Dr. Marymegan Daly, and Dr. Zakee Sabree, for serving on my doctoral committee, providing guidance and constructive criticism, as well as providing review for this document. I also give my appreciation to Dr. Elizabeth Marschall and Dr. John Freudenstein, who served as department heads during my time in Evolution, Ecology, & Organismal Biology at The Ohio State University. They both did an excellent job making the department an enjoyable place to work and learn.

I greatly appreciate the help of Dr. John Horner and Dr. Richard Miller, for collecting the pitcher fluid samples that made Chapter 4 possible. Thanks also to Lois Ochs and Jeff Gold for sending live *Sarracenia* plants and to Kinjie Coe for sending tissue samples of the elusive *Darlingtonia*. And thanks to Abbie Zimmer, who in her time working with the lab did some great scientific illustrations, one of which is published within!

Next, I thank my colleagues in the department, in particular the members of the Carstens Lab: Ariadna Morales, Megan Smith, Coleen Thompson, Drew Duckett, Jordan Satler, and Tara Pelletier. With willingness to provide help, feedback, and the strictly-friendly occasional competition, I can truly say I will miss working alongside all of you. (An extra special thanks to Megan, who carried heavy containers of water up to the greenhouse for me for an entire semester.) Thanks also to my officemates Alejandro Otero Bravo and Ben Jahnes, for keeping things interesting and putting up with my general messiness and office pets; and also to all my other friends in the department, notably (but not limited to) Naava Honer, Drew Spacht, Cody Cardenas, and Jessie Lanterman.

I feel I must show my special appreciation to those at the Ohio Supercomputer Center. Without their willingness to support my research with the continued allocation of large amounts of computational resources, none of this would have been possible. I deeply appreciate the amount of leeway I was given, as my work was never interrupted by computational limits even as I went hundreds of thousands of hours over-budget, and filled over 140 TB of storage space. You guys are great! I'm sorry for taking advantage *a little* too much … and apologies to the students who come after me.

Now, on a more personal note, I would like to give my heartfelt love and appreciation to my parents, Laurie and Larry Wheeler, and my sister and brother-in-law, Emily and Timothy Clark. The continued unconditional support, even through occasional misgivings, has meant a lot to me, and has helped provide me with the confidence to push through the difficult times. To my parents, thank you for nurturing my love of science and technology which got me here today. To my sister, thank you for looking up to me; it has made me work harder to try and set the best example I possibly can.

Finally, I give my love, appreciation, and admiration to my partner Jordan Waltz. While my academic adventure was already well underway when we met, having you in my life has helped me keep up the drive to succeed in this and other difficult pursuits, and you're always quick to help take some of the weight off my shoulders when I struggle. You don't realize how often I stop to think about just how lucky I am, and just knowing that you're proud of me for what I've accomplished means the world to me. I only hope I can continue making you proud for a long time to come.

Vita

2011.....B.S. Biological Sciences,
Mississippi State University

2013.....M.S. Botany,
Mississippi State University

2018 – Present.....Computational Genomics Bioinformatics Scientist
Nationwide Children's Hospital, Columbus, OH

2014 – 2015; 2017 ..... Distinguished University Fellow
Department of Evolution, Ecology, and Organismal Biology
The Ohio State University

2015 – 2016; 2018 ..... Graduate Teaching Assistant
Department of Evolution, Ecology, and Organismal Biology
The Ohio State University

2015 – 2018.....Delegate to The Ohio State University Council of Graduate Students
Evolution, Ecology, & Organismal Biology

2015, 2017, & 2018 ..... Ohio Supercomputer Center Resource Awards
(84,900 Total RUs)

2016 .....Janice Carson Beatley Fund Award

Publications

9. Ranathunge, C., **Wheeler, G. L.**, Chimahusky, M., Perkins, A. D., Pramod, S., & Welch, M. E. (2018). Transcribed microsatellites often influence gene expression in natural sunflower populations. *bioRxiv*, 339903. Preprint.

8. Ranathunge, C., **Wheeler, G. L.**, Chimahusky, M. E., Kennedy, M. M., Morrison, J. I., Baldwin, B. S., ... & Welch, M. E. (2018). Transcriptome profiles of sunflower reveal the potential role of microsatellites in gene expression divergence. *Molecular ecology*, *27*(5), 1188-1199.

7. **Wheeler, G. L.**, & Carstens, B. C. (2018). Evaluating the adaptive evolutionary convergence of carnivorous plant taxa through functional genomics. *PeerJ*, *6*, e4322.

6. Wallace, L. E., **Wheeler, G. L.**, McGlaughlin, M. E., Bresowar, G., & Helenurm, K. (2017). Phylogeography and genetic structure of endemic Acmispon argophyllus and A. dendroideus (Fabaceae) across the California Channel Islands. *American journal of botany*, *104*(5), 743-756.

5. Gruenstaeudl, M., Reid, N. M., **Wheeler, G. L.**, & Carstens, B. C. (2016). Posterior predictive checks of coalescent models: P2C2M, an R package. *Molecular ecology resources*, *16*(1), 193-205.

4. **Wheeler, G. L.**, Dorman, H. E., Buchanan, A., Challagundla, L., & Wallace, L. E. (2014). A review of the prevalence, utility, and caveats of using chloroplast simple sequence repeats for studies of plant biology. *Applications in plant sciences*, *2*(12).

3. McGlaughlin, M. E., Wallace, L. E., **Wheeler, G. L.**, Bresowar, G., Riley, L., Britten, N. R., & Helenurm, K. (2014). Do the island biogeography predictions of MacArthur and Wilson hold when examining genetic diversity on the near mainland California Channel

Islands? Examples from endemic Acmispon (Fabaceae). *Botanical journal of the Linnean Society*, *174*(3), 289-304.

2. **Wheeler, G. L.** (2013). *Population-level genetic structure of Acmispon argophyllus on the Channel Islands of California*. Mississippi State University.

1. **Wheeler, G. L.**, McGlaughlin, M. E., & Wallace, L. E. (2012). Variable length chloroplast markers for population genetic studies in Acmispon (Fabaceae). *American journal of botany*, *99*(10).

Fields of Study

Major Field: Evolution, Ecology, and Organismal Biology

Minor Field: Statistics

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Overview

The source of new traits and adaptations has been a contentious issue in evolutionary biology, and was once one of the major stumbling blocks of the field. It was argued that, due to the allegedly irreducible complexity of structures, the probability of any new feature appearing was nearly zero. Improved understanding of genetics, developme

nt, homology, and ecology has revealed several avenues by which new features can appear, for enzymatic pathways (Jensen, 1976) as well as macroscopic structures such as eyes, limbs, and arthropod ornamentation (Shubin, Tabin & Carroll, 2009).

While it is true that complex new traits can appear as a result of the creation of new sequence via the build-up of mutations, this process is incredibly slow; due to a nearly infinite number of possible genetic states, it is also unlikely to repeatedly arrive at the same result. This is at odds with observations of the natural world, where convergent evolution, sometimes on short evolutionary timescales, is a frequent occurrence. Instead, it has become clear that through their evolutionary history, organisms have repeatedly adapted by repurposing existing "tools". Some genes performing one function in an ancestral organism, through small changes in protein structure, localization, or expression, are able to perform vastly different functions in their descendants (McLennan, 2008). Raw material on which selective processes can act is often provided through the duplication of these genes, with the resulting copies then being co-opted to serve a new functional role. For plants in particular, it is common for entire genomes to be duplicated; lineages become polyploid, with subsequent genomic reduction (Flagel & Wendel, 2009). As the lineage returns to the diploid state, genes that have evolved into novel adaptive

states are preferentially retained, while copies that are unnecessary are lost (Sémon & Wolfe, 2007). This leaves behind a genomic signal of overrepresentation of genes that have been selected for after a duplication event.

Another path a lineage can take in adapting to a new niche is that of beneficial symbioses with other organisms (Law & Dieckmann, 1998). While an ancestral organism may not possess the genetic material to easily adapt, it may circumvent this by forming relationships with other species that can provide these assets. In plants, the most common examples of microbial symbioses involve root associations with nitrogen-fixing bacteria as well as mycorrhizal fungi with superior ability to access water and nutrients (Reynolds et al., 2003). Mutualistic interactions with macroorganisms are also ubiquitous in the angiosperms, with a high percentage of species relying on insects for transfer of genetic material in pollen or vertebrates to disperse propagules (Wheelwright & Orians, 1982).

The Carnivorous Plants

In this study, I use carnivorous plants as an ideal example of how evolutionary convergence can occur. While once grouped together due to their most striking features – adaptations to trap and digest insects – carnivorous plants are an extreme case of polyphyly (Givnish, 2015). Despite similar adaptive strategies, in many cases even replicating specific shapes and structures, carnivorous plants have arisen independently in five angiosperm orders, with possible cases of carnivory outside the “higher plants” (Hess, Frahm & Theisen, 2005). This has occurred in response to similar ecological conditions, defined by an abundance of water,

sunlight, and invertebrates, but a lack of nitrogen, phosphorus, and mineral nutrients (Ellison & Gotelli, 2001).

Previous studies have isolated and identified some of the proteins that give some carnivorous plant species their digestive abilities (Scherzer et al., 2013); however, at the genetic level, it is not yet known how different lineages were initially able to gain their carnivorous abilities. For lineages in which particular digestive enzymes have been identified, it is typically hypothesized that these novel functions have arisen from modifications to the ancestor's evolutionary "toolkit". A specific example is the enzyme chitinase, which some carnivorous plant taxa use to digest the chitin that composes the outer shell of insect prey (Renner & Specht, 2012). While superficially this may seem a strange enzyme for a plant to produce, in fact many plants produce chitinase enzymes – as part of their immune response against fungal pathogens, which are also composed of chitin (Bemm et al., 2016). Some carnivorous plants are also known to obligately rely on symbioses to obtain resources. One clear example of this is found in *Roridula*, a genus of sticky-trap plants with no documented enzymatic activity. Instead, they absorb the feces of a capsid bug, *Pameridea roridulae*, which in turn relies on the plants to ensnare its prey (Midgley & Stock, 1998).

The Pale Pitcher Plant, *Sarracenia alata*

Within the carnivorous plants, this study will focus on one species in particular: *Sarracenia alata* Alph. Wood. This species, commonly known as the North American pale pitcher plant, is found in four states along the U.S. Gulf Coast; it possesses a disjunct distribution, split into eastern and western groups by the Mississippi River (Carstens & Satler,

2013). These plants are restricted to wet, acidic, nutrient-poor soils where their carnivory allows them to outcompete typical plants, which rarely thrive in these conditions. Notably, they are

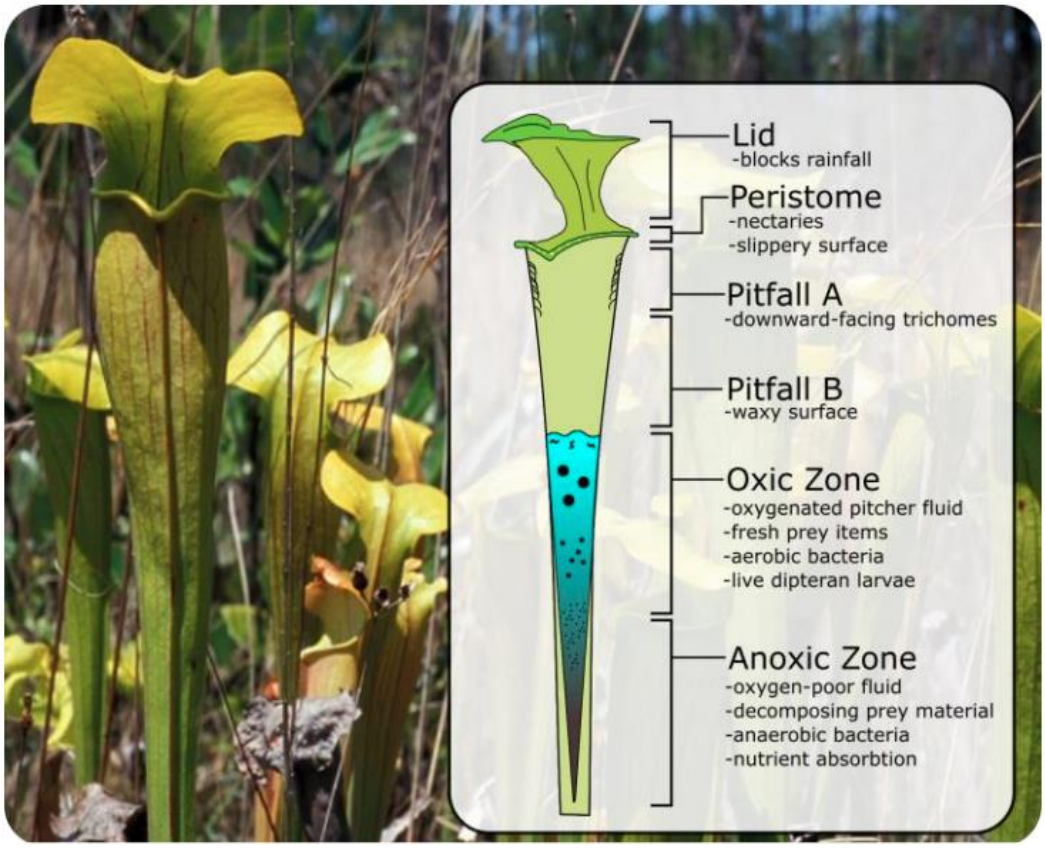

Figure 1: The structure and appearance of Sarracenia alata. *A typical specimen is depicted in its natural habitat (left), and represented diagrammatically in terms of morphology and functional zones (right). Photograph by Dr. Barry Rice, used with permission.*

common in coastal pine savannah, where in addition to the ideal soil conditions, they may also benefit from fire-control of competitors (Barker & Williamson, 1988). While *S. alata* is not recognized as a protected species, it, like most wetland species, is at risk of habitat destruction due to human activity and changing climate.

Visible as rosettes of tall, highly-modified leaves emerging from an underground rhizome, *S. alata* has sacrificed photosynthetic leaf area for prey-trapping ability. The pitchers possess several notable features and are divided into functional zones (Figure 1). Pitchers possess a lid, which blocks rainfall to regulate the level of liquid in the traps. Below this, the peristome (lip) of the pitcher attracts insect prey – primarily ants – with a sugary reward from extrafloral

nectaries. When wet, this surface has very low friction, causing insects to slip and fall into the trap. Downward-facing trichomes (stiff hairs) and a waxy surface prevent prey from climbing out of the pitfall; due to exhaustion, they eventually drown in the fluid. In the fluid's oxygenated upper layer, aerobic bacteria and invertebrates break down the insect corpses, which then sink into the unoxygenated portion of the pitcher fluid where they are finally reduced to bioavailable nutrients by anaerobic bacteria and absorbed by the plant.

Sarracenia alata was chosen as an ideal focal organism for this study for a number of reasons. First, despite research going back decades, little genetic work has been done in this species. The taxon has been included in phylogenetic analyses and population-level studies, but until recently only using a very limited set of neutral markers (Ellison et al., 2012; Stephens et al., 2015). In the genus as a whole, most studies have only considered the most widespread taxon, *Sarracenia purpurea* L., and functional research has been limited to older methods such as protein crystallization and radioisotope markers (Gallie & Chang, 1997; Butler, Gotelli & Ellison, 2008). This makes this system ripe for study by an approach based in next-gen sequencing, where any new data is likely to provide a wealth of new information. Second, while the mechanisms behind the function of more famous carnivorous plant taxa such as Venus's flytrap (*Dionaea muscipula* Sol. *ex.* J.Ellis) are now becoming well-understood, the elements of *S. alata*'s digestive process are likely more complex. Far from a simple secretion of digestive enzymes, *S. alata*'s pitchers contain an entire microenvironment, with invertebrates, microbes, and microfauna each filling a niche (whether it be beneficial, neutral, or parasitic to the host plant) (Koopman et al., 2010). This complex network of organisms and pathways increases the number of questions that can be addressed with the data collected, potentially allowing new studies for years to come with limited additional sampling. Finally, in the *Sarraceniaceae*, there

exists a long and ongoing debate about the exact origins of the plants' digestive abilities. Studies have strongly suggested that microbes do play a vital role in the digestive process (Luciano & Newell, 2017a), but some have gone as far as to conclude that the host plant is digestively inert, relying on these microbes entirely (Adams & Smith, 1977). This situation would represent a fascinating approach to plant carnivory, with comparatively little documentation.

In this study, I hope to address several questions, about the evolution of carnivorous plant taxa in general, and about *Sarracenia alata* in particular. First, have carnivorous plants convergently evolved by repeatedly using the same set of gene functions? If so, can these functions be predicted, and can a signal of their adaptive role be detected at the genomic level? Second, can genome sequencing and assembly reveal the adaptive history of *Sarracenia alata*? Does its genome show the signs of carnivory (if any) found in other taxa, or does it instead show signs of having evolved down a different path (e.g., relying on mutualism)? Finally, can the meta-transcriptome of *S. alata*, paired with an annotated reference genome, reveal in detail the associations and interactions between the plant and the microbial environment it hosts? Does this information, along with the genomic results, give a definitive answer to the questions of the origin of the carnivorous novelty in this species and its relatives? Using the latest bioinformatics methods, I hope to answer these questions and more in the following chapters.

Chapter 2: Evaluating the adaptive evolutionary convergence of carnivorous plant taxa through functional genomics

Note: This chapter was first published in PeerJ 6, e4322 (Wheeler & Carstens, 2018). Minor changes have been made to better fit this format.

Abstract

Carnivorous plants are striking examples of evolutionary convergence, displaying complex and often highly similar adaptations despite lack of shared ancestry. Using available carnivorous plant genomes along with non-carnivorous reference taxa, this study examines the convergence of functional overrepresentation of genes previously implicated in plant carnivory. Gene Ontology (GO) coding was used to quantitatively score functional representation in these taxa, in terms of proportion of carnivory-associated functions relative to all functional sequence. Statistical analysis revealed that, in carnivorous plants as a group, only two of the 24 functions tested showed a signal of substantial overrepresentation. However, when the four carnivorous taxa were analyzed individually, 11 functions were found to be significant in at least one taxon. Though carnivorous plants collectively may show overrepresentation in functions from the predicted set, the specific functions that are overrepresented vary substantially from taxon to taxon. While it is possible that some functions serve a similar practical purpose such that one taxon does not need to utilize both to achieve the same result, it appears that there are multiple approaches for the evolution of carnivorous function in plant genomes. Our approach could be

applied to tests of functional convergence in other systems provided on the availability of genomes and annotation data for a group.

Keywords

carnivorous plants, Gene Ontology, functional genomics, convergent evolution

Introduction

Convergent evolution provides some of the strongest support for the theory of evolution through natural selection. In the case of evolutionary convergence, organisms that may have very different evolutionary history (as measured phylogenetically), are driven by similar selective pressures to a highly similar phenotype (Losos, 2011). These selective pressures repeatedly create the same adaptive syndrome – a set of characteristics which come together to allow a specific lifestyle or perform a certain task (Reich et al., 2003). In many instances in the past, convergent evolutionary syndromes have confounded taxonomists, who (for example) mistakenly grouped New-World and Old-World vultures (Seibold & Helbig, 1995), all marine mammals (Foote et al., 2015), and many disparate lineages of microscopic organisms (Palenik & Haselkorn, 1992; Scamardella, 1999; Gupta, 2000), into clades which ultimately proved to be paraphyletic.

While phenotypic features of convergent taxa will appear superficially similar, they are not expected to share genomic similarity due to their evolutionary independence. A large number of possible sequence combinations can result in the same protein (Storz, 2016) and potentially large number of protein forms and combinations of multiple proteins that can produce the same effect (Bork, Sander & Valencia, 1993; Doolittle, 1994), so objectively defining an evolutionary

syndrome using genomic data is challenging. One possible solution is to define these syndromes as a set of discrete functions rather than as a set of nucleotide sequences. In this way, convergent syndromes are described in the same way they have evolved – adaptively by function – and can be evaluated as convergent or not based on sequence similarity. Gene Ontology (GO) coding (Ashburner et al., 2000a) provides an objective system by which to achieve this goal. By designating numerical codes for all possible biological activities and components, ranked hierarchically from general to specific, synonymy of function can easily be measured between even distantly related organisms. (Throughout this text, when a discrete GO term is being referenced, it will be presented in italics, whereas when functions are being referenced in the more general sense, it will be presented in plain text.) Using either experimentally determined gene/protein function or sequence similarity to previously identified functions, the activities of individual genes are paired with specific numeric codes. Gene Ontology analyses have been used in other studies to determine the functional components to a variety of traits, adaptations and physiologies of interest, including adaptation to high altitudes (Qiu et al., 2012), depth tolerance in deep-sea bacteria (Vezzi et al., 2005), and a number of human disorders (Ahn et al., 2003; Holmans et al., 2009); however, these have identified known genes of interest and then drawn conclusions of function *post hoc*. Rather than assigning the Gene Ontology codes first and subsequently determining the functions of particular interest as has been done previously, we can select functions of expected relevance *a priori* in order to allow for quantitative testing of their adaptive relevance by comparing functions in genomes in species that exhibit a convergent function. To the best of our knowledge, this is a novel approach.

Plant carnivory

One particularly notable convergent polyphyletic group is that of the carnivorous (alternatively, insectivorous) plants. Carnivorous plant taxa were originally classified as a single group due to their most striking and apparent feature, while disregarding features that would typically be used to define a botanical group (e.g., floral morphology; Primack, 1987). Subsequent work has demonstrated that a substantial number of phylogenetically distant plant lineages have evolved a carnivorous lifestyle (Givnish, 2015), presumably in response to similar selective pressures. As different lineages (or branches of the same lineage) have approached the process of insect trapping and digestion in different ways, this has in some cases made the defining of a plant as carnivorous or non-carnivorous difficult (Lloyd, 1934).

Givnish *et al.* (1984) defines a carnivorous plant as one that fulfills two requirements: it must gain some detectable fitness benefit from animal remains in contact with its surfaces, and it must possess adaptations that facilitate the attraction, capture, or digestion of these prey animals. By considering only functional attributes, this definition allows a wide range of variability in the evolutionary histories and routes of adaptation of plants that are considered carnivores. Currently, nearly 600 angiosperm species are recognized as carnivorous, representing as many as nine independent origins across five families (Givnish, 2015). In addition, investigations into possible carnivorous traits in non-vascular plants such as liverworts are ongoing (Hess, Frahm & Theisen, 2005), suggesting that evolutionary shifts in nutrient acquisition strategies are perhaps even more common that currently recognized. The multiple origins and evolutionary convergence demonstrated by radiations such as those in *Nepenthes* and *Sarracenia* indicate that plant carnivory is not phylogenetically constrained; rather, it is likely that these plants are limited by their specific nutrient economics (Bloom, Chapin, & Mooney, 1985), which allow them to

outcompete more typical nutrient acquisition strategies only in specific habitats (Ellison & Gotelli, 2001; Ellison et al., 2003).

Carnivorous plants occupy habitat where there is little competition for sunlight. Previous studies have shown that, by leaf mass, many carnivorous plants have poor photosynthetic yield (Ellison & Farnsworth, 2005; Ellison, 2006), a likely consequence of the adaptions of their leaves for the capture of insect prey. Additionally, some carnivorous plants invest photosynthetic carbon in the fluids or secretions utilized for prey capture. In *Drosera*, some 3-6% this carbon, which would otherwise be expended on reproduction or vegetative growth, is used to capture prey (Adamec, 2002). As a result of these compromises, carnivorous plants only outcompete other plants in habitat where the resources that they sacrifice as a consequence of the carnivorous lifestyle (carbon, water, sunlight) are plentiful, while the resources they specialize in obtaining (nitrogen, phosphorus) are scarce. These environments are likely to be wet and sunny, with acidic, nutrient-deficient soils (Givnish et al., 1984; Ellison & Gotelli, 2001).

Carnivory-associated functions

The most apparent trait of carnivorous plants is their ability to break down prey items using digestive enzymes. As digesting animal tissue is presumably not in the repertoire of ancestral angiosperms, a question of interest is how these enzymes have evolved. In many cases, genes for digestive enzymes are apparent modifications of genes utilized in resistance and correspond to pre-existing pathways related to herbivores and pathogens (Schulze et al., 2012; Fukushima, Fang, et al., 2017) or other processes present in most plants. Examples of such enzymes include chitinases, which were modified from anti-fungal and insect herbivore deterrence enzymes (Hatano & Hamada, 2008; Renner & Specht, 2012), proteases, likely derived

from those involved in bacterial resistance (Mithöfer, 2011), and lipases, which are involved in metabolizing stored energy (Seth et al., 2014). Furthermore, it appears that enzymes with similar functions have evolved convergently in taxa with independent carnivorous origins (Fukushima, Fang, et al., 2017), suggesting that it may not be difficult to evolve into the carnivorous niche. However, digestive enzymes may also be obtained through symbiotic interactions with micro- (Koopman et al., 2010; Caravieri et al., 2014) or macroorganisms (Midgley & Stock, 1998; Anderson & Midgley, 2003), suggesting that it may be possible to evolve into the carnivorous niche in part by appropriating the digestive enzymes of other species. While these plants fit Givnish et al.'s (1984) definition of carnivores, these digestion-associated genes would not be identifiable in the plant itself and thus would not contribute to functional overrepresentation in genomic analyses.

In addition to modifications or resistance genes or the appropriation of enzymes produced by symbionts, evidence suggests that genes used in nutrient transport are particularly important to the carnivorous lifestyle. Plant genomes possess as many as 10 times the number of peptide transport genes compared to other eukaryotes (Stacey et al., 2002), in addition to a wide variety of transport pathways for nitrate and ammonium (Williams & Miller, 2001). In carnivorous plants, the relative number of these pathways may be even higher. For example, in a transcriptomic analysis of *Utricularia gibba* L., a carnivorous bladderwort with a minute genome of only 80 megabases, 77 unique sequences corresponding to nitrogen transport were identified (Ibarra-Laclette et al., 2011). Modification and specialization has also occurred in transporters for other resources. For plants with traps involving rapid movement such as *Dionaea muscipula* Sol. ex J.Ellis, uptake of prey nutrients may be coupled to a trap's electrical potential (Scherzer et al., 2013). Modified pathways for osmolite uptake have been identified in *D. muscipula*, which

uses the HKT1-type ion channel to absorb sodium without disrupting the action potential of the trap (Böhm et al., 2016). Similar adaptations may benefit less active traps as well, as for example in *Sarracenia flava* L. amino acid uptake is dependent on a potassium ion gradient (Plummer & Kethley, 1964).

Genomics represent a new approach to investigate the evolution of novel organismal function. While the origin of novel biological functions and their role in adaptation to new habitats and ecological niches has been an important topic in evolutionary biology since the inception of the field (Darwin & Darwin, 1889), we now know that genes may be preferentially duplicated and modified, a common route to increased complexity and the possibility of new structures (Vandenbussche et al., 2003) and pathways (Monson, 2003). In more extreme cases, a whole-genome duplication event precedes an episode of major adaptive change (Soltis et al., 2009), leaving a lineage with thousands of redundant additional genes on which evolutionary processes can act. Gene copies with adaptive value are preferentially retained, while others are silenced and eventually lost (Adams & Wendel, 2005). If this general pattern is true of the genes involved in plant carnivory, such genes should be identifiable on the basis of function and would be expected to show a signal of overrepresentation in the genome.

Gene Ontology coding is an essential tool for resolving the issue of relating functionally similar (but non-homologous) genes – by design, genes that differ substantially in ancestry but provide the same function should be assigned the same Gene Ontology code(s). These descriptors are originally assigned based on experimental studies of specific genes in model organisms, which later allows non-experimental assignment using sequence homology; however, as automated annotation must be based on the content of a reference database, known biases in these databases must be considered. For example, studies addressing multiple genes often focus

on a specific gene class within a specific organism, resulting in an overemphasis of that class in that organism and its relatives; experiment-based annotations will be far more common for model organisms or those of economic interest; and, as more sequences are assigned function through extrapolation rather than experimentation, those assignments can be further propagated, progressively increasing the distance from the original experimental basis (Thomas et al., 2012; Altenhoff et al., 2012). In particular, due to these biases and methods of accurately matching samples to references, there is concern that functional divergence may be missed in cases where divergent sequences remain similar, or conversely, that erroneous function may be assigned when there is substantial divergence from the nearest-matching reference sequence. Despite this, it has been previously shown that known functionally-divergent paralogs also diverged (by 32% on average) in GO codes assigned by automation (Blanc & Wolfe, 2004) and that genes typically retain highly similar functions at amino acid identity levels as low as 40% (Sangar et al., 2007). Thus, there is reason to believe that identifying function from sequence data should be sufficiently accurate at our desired level of specificity.

Hypotheses

This study seeks to test for a functional genetic signal of evolutionary convergence at the level of the genome. Specifically, it seeks to test whether or not a convergently evolved functional syndrome (i.e. metabolic pathways of carnivory) will rely on the same functions across lineages (as seen in Yang et al., 2015). Three possibilities will be considered. First, organisms sharing this syndrome may not be genomically distinct from others. This is possible if the functional changes required for this syndrome are not substantial at the genome level (e.g. changes based on slight modification of regulatory elements or alternative splicing), or if neutral

variation among taxa is so substantial that the changes fall within the range of normal lineages. In this case, no signal should be detected differentiating experimental taxa from control samples (i.e., GO codes matching to expected carnivory-associated functions are not overrepresented). Second, a syndrome may require a specific set of functions at high representational levels in every lineage where it arises. This would be expected if the use of certain molecular machinery were unavoidable for a task, preventing evolution of the syndrome by any other pathways. In this case, it would be expected that a strong signal would be detected for functions across all experimental taxa (i.e., GO codes matching to expected carnivory-associated functions are uniformly overrepresented across carnivorous taxa). Lastly, a syndrome may indeed make use of some functions from a set list each time it arises, but not necessarily the same functions in each case. This would occur where there are several ways to address the same problem. A result where many of the predicted functions show strong signal, but with greatly different findings in each taxon, would support this model (i.e., GO codes matching to different carnivory-associated functions are overrepresented in each carnivorous taxon).

Materials & Methods

Identification of carnivory-associated functions

A literature review was conducted to develop a reference set of functions previously found to be associated with plant carnivory. A topic search was performed on Web of Science in December, 2016 with the following parameters: ("carnivorous plant" OR "insectivorous plant") AND ("gene" OR "genome" OR "transcriptome" OR "protein") AND ("digestion" OR "transport"), producing 21 results. Publications discussing specific genes (Owen et al., 1999; An, Fukusaki & Kobayashi, 2002; Scherzer et al., 2013; Böhm et al., 2016) or overviews of

putatively carnivory-associated genes (Ibarra-Laclette et al., 2011; Schulze et al., 2012; Rottloff et al., 2016) in sequenced carnivorous plant taxa were included (details in Table 10). From *Dionaea muscipula,* functions of proteins identified via proteomic analysis of the trap fluid (Schulze et al., 2012) were listed, with the addition of genes related to transport that had been specifically targeted by other studies (Owen et al., 1999; Böhm et al., 2016). Similarly, in *Nepenthes*, proteomic analysis of trap fluid released a list of functions likely to be associated with plant carnivory (Rottloff et al., 2016), with other studies assaying for a specific digestion-associated enzyme and detecting transport activity via traps' glandular symplasts (An, Fukusaki & Kobayashi, 2002 and Scherzer et al., 2013, respectively). In the bladderwort *Utricularia gibba* L., transcriptomic analysis was used to detect statistically increased expression of genes in traps and leaves putatively associated with carnivory (Ibarra-Laclette et al., 2011). Gene function terms, as given by these publications, were cross-referenced with the AmiGO2 Gene Ontology Database (Balsa-Canto et al., 2016) and matched to discrete GO codes that accurately represent their functions (Table 1). Of 54 total terms selected, 36 final GO codes were matched, with 5 terms synonymized and combined with matched terms and 13 terms having no appropriate match.

Gene Ontology Term	GO Code	Gene Ontology Term	GO Code
actin filament	GO:0005884	*heat shock protein activity*	GO:0042026; GO:0006986; GO:0034620
alpha-galactosidase activity	GO:0004557	*lipase activity*	GO:0016298
alternative oxidase activity	GO:0009916	*lipid transport*	GO:0006869
ammonium transmembrane transport	GO:0008519; GO:0072488	*methylammonium channel activity*	GO:0015264
aspartic-type endopeptidase activity	GO:0004190	*peroxidase activity*	GO:0004601
ATP:ADP antiporter activity	GO:0005471	*phosphatase activity*	GO:0016791
ATPase activity	GO:0016887	*phospholipase activity*	GO:0004620
beta-galactosidase activity	GO:0004565	*polygalacturonase activity*	GO:0004650
beta-glucanase activity	GO:0052736	*polygalacturonase inhibitor activity*	GO:0090353
chitinase activity	GO:0004568	*protein homodimerization activity*	GO:0042803
cinnamyl-alcohol dehydrogenase activity	GO:0045551	*ribonuclease activity*	GO:0004540
cyclic-nucleotide phosphodiesterase activity	GO:0004112	*serine-type carboxypeptidase activity*	GO:0004185
cysteine-type peptidase activity	GO:0008234	*sodium ion transmembrane transporter activity*	GO:0022816
endonuclease complex	GO:1905348	*superoxide dismutase activity*	GO:0004784
formate dehydrogenase complex	GO:0009326	*symplast*	GO:0055044
fructose-bisphosphate aldolase activity	GO:0004332	*thioglucosidase activity*	GO:0019137
glucosidase complex	GO:1902687	*water channel activity*	GO:0015250
glutathione transferase activity	GO:0004364	*xylanase activity*	GO:0097599

Table 1: Carnivory-associated functions identified via literature review. *Functions were matched to Gene Ontology terms and codes using the AmiGO2 database (Balsa-Canto et al., 2016). In cases where multiple GO codes are given, they are equivalent to or deprecated from the best-matching current term. See Table 10 for more information.*

Taxon sampling

GenBank's list of assessed plant genomes was surveyed for the inclusion of plants historically considered to be carnivorous. Four were available: *Cephalotus follicularis* (Fukushima, Fang, et al., 2017), *Drosera capensis* (Butts, Bierma & Martin, 2016), *Genlisea aurea* (Leushkin et al., 2013), and *Utricularia gibba* (Lan et al., 2017). The carnivorous taxa sampled represent three independent origins of plant carnivory (*Genlisea* and *Utricularia* likely sharing a single origin) in three plant orders (Caryophyllales, Oxalidales, and Lamiales). They also exemplify four different strategies for prey-capture. *Cephalotus* is a pitcher/pitfall trap, using a nectar lure, slippery rim, and downward-facing projections to guide prey into a digestive soup and prevent their escape; this strategy is also seen in *Nepenthes,* most *Sarracenia*, and some carnivorous bromeliads. *Drosera* is a sticky-trap plant, with glandular trichomes on its leaves that secrete both sticky compounds to prevent prey's escape and digestive enzymes to break them down; *Pinguicula* and *Byblis* also use this strategy. *Genlisea* is considered a lobster-pot trap, where prey species are guided to a small, funnel-like opening, through which exit is impossible; *Sarracenia psittacina* and, arguably, *Darlingtonia californica* employ this strategy. Lastly, *Utricularia gibba*, an aquatic carnivorous plant, uses a number of air-filled bladders to capture and digest prey. A trigger hair is stimulated as potential prey investigates the trap, releasing an air bubble contained within; the resulting vacuum pulls the prey inside, and the trap closes behind them. While no other carnivorous taxa possess this specific form, it does share some characteristics (a fast-moving trap activated by the stimulation of a trigger hair) with *Aldrovanda* and *Dionaea muscipula*. The trap characteristics, floral morphology, and overall growth form of the carnivorous taxa included in this study are depicted in Figure 2.

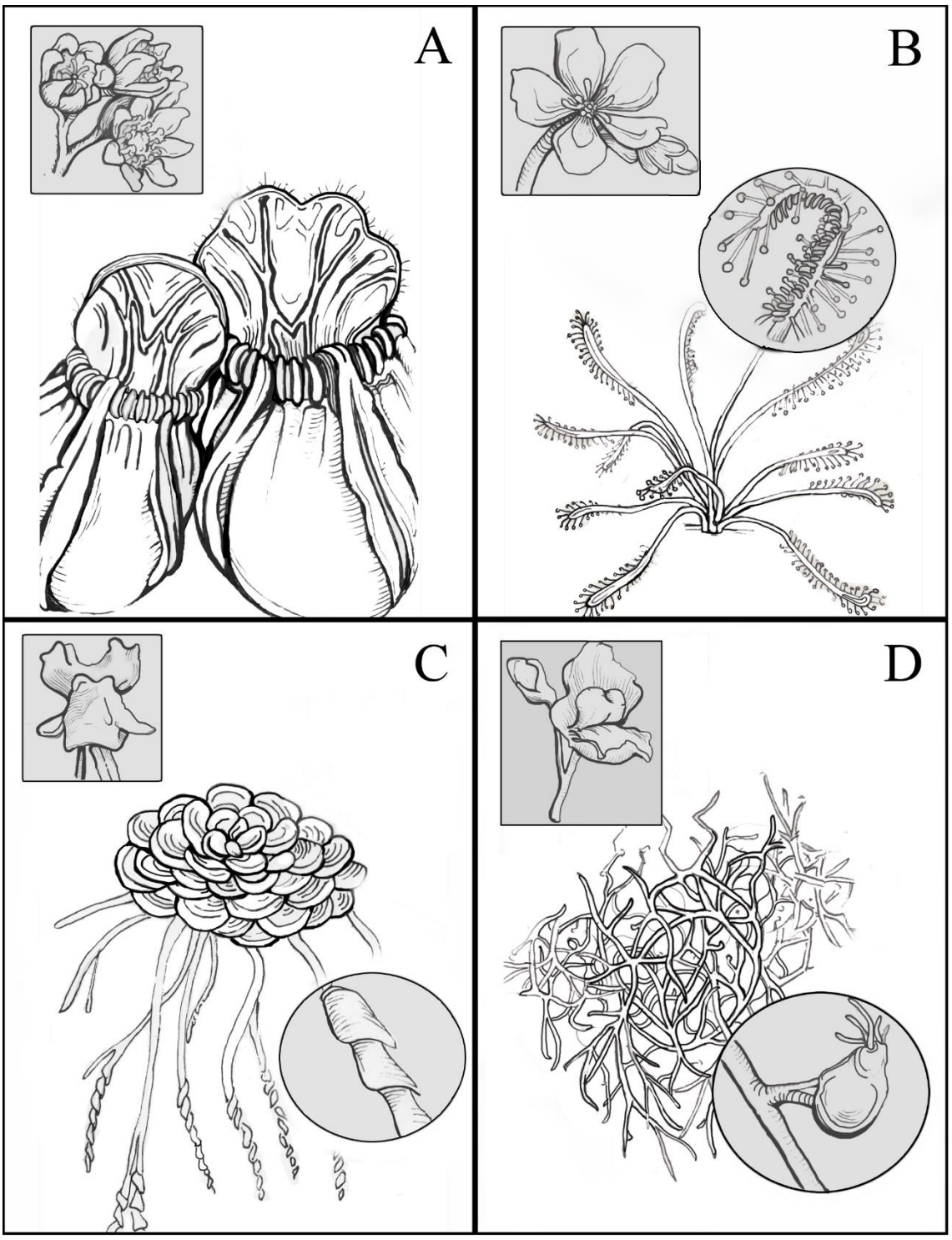

Figure 2: Illustrations of the carnivorous taxa included in this study. *Floral characteristics (square inset) and trap morphology (circle inset) are shown, as well as overall growth form. Taxa shown are (A) Cephalotus follicularis, (B) Drosera capensis,(C) Genlisea aurea, and (D) Utricularia gibba. Illustrations by Abbie Zimmer, 2017, included with permission.*

Non-carnivorous plants were also surveyed in order to establish a control range of "typical" flowering plants. All assessed plant genomes for which Gene Ontology-coded annotations are already available were included: *Arabidopsis thaliana* (Swarbreck et al., 2008), *Boea hygrometrica* (Xiao et al., 2015), *Glycine soja* (Kim et al., 2010; Qi et al., 2014), and *Oryza sativa* (Ohyanagi, 2006). Note that one of the carnivorous taxa, *Genlisea aurea*, also possessed GO annotations. Lastly, the genomes of the two non-carnivorous plants most closely related to carnivorous taxa were included: *Ocimum tenuiflorum* (Upadhyay et al., 2015), closest sequenced relative of *Byblis, Genlisea, Pinguicula,* and *Utricularia*; and *Actinidia chinensis* (Huang et al., 2013), closest sequenced relative of *Darlingtonia, Heliamphora, Roridula,* and *Sarracenia. Boea hygrometrica,* included for its available annotations, is also within the order of *Genlisea* and *Utricularia*. The reference-range taxa selected cover five orders (Brassicales, Ericales, Fabales, Lamiales, and Poales), including both Monocots and Eudicots; thus, these samples can be considered a reasonable representation of the diversity and variation of angiosperms as a whole (Figure 3). While pairwise sampling and analysis of related carnivorous and non-carnivorous taxa would be optimal to explicitly control for phylogenetic effects, this is unfortunately not possible at present due to the lack of sequenced genomes for many plant orders and the scarcity of annotated plant genomes in general. However, we expect our current reference sampling design, which includes both non-carnivorous representatives from several carnivore-containing orders and a wide phylogenetic range of taxa overall, to somewhat mitigate this potential source of bias.

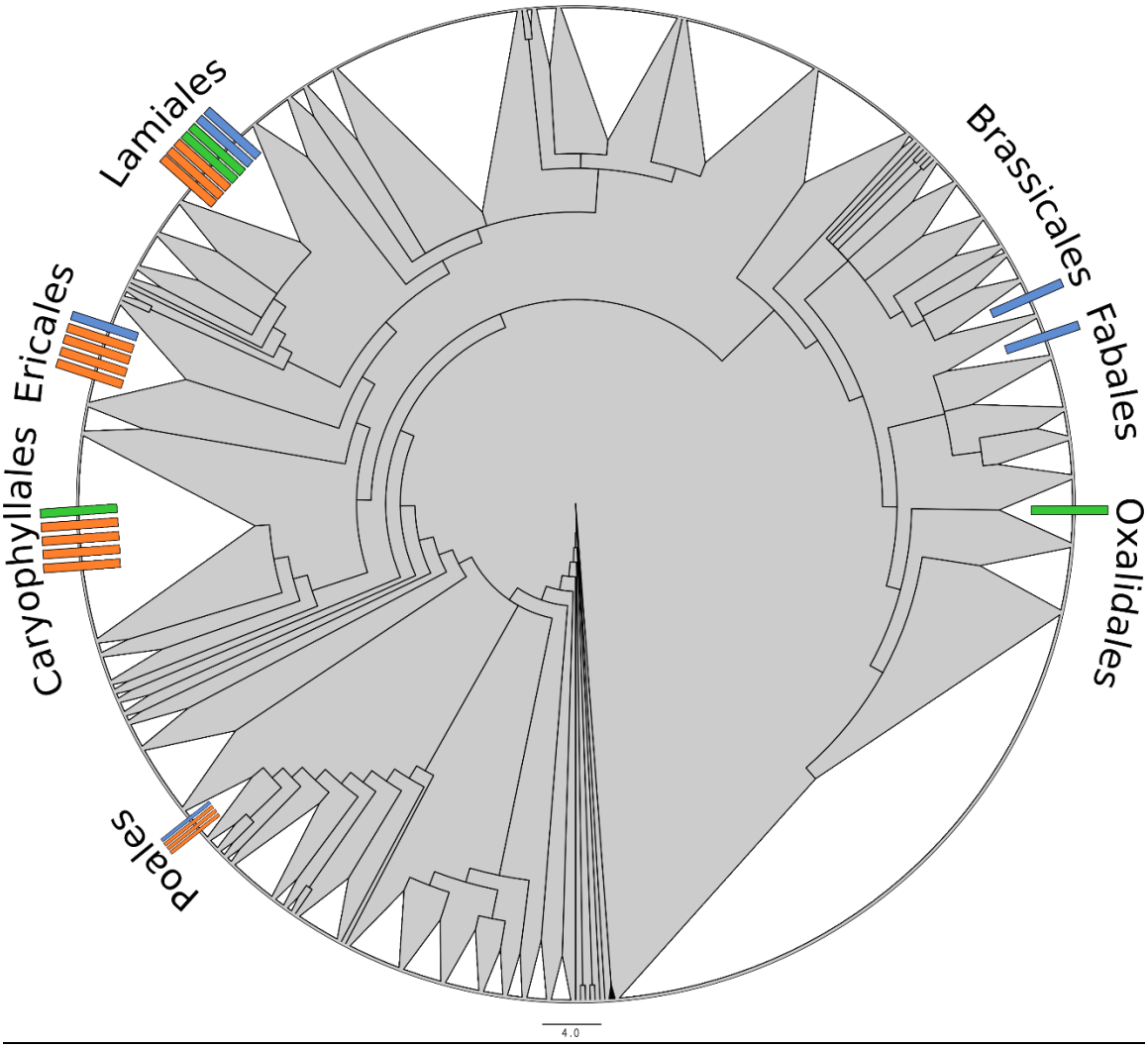

Figure 3: Radial phylogeny of all angiosperms, indicating the position of taxa relevant to this study. *White-filled triangles indicate monophyletic plant orders. Each bar indicates one genus. Blue indicates a typical/non-carnivorous control taxon included in this study; green indicates a carnivorous taxon included in this study; orange indicates a carnivorous genus (as listed in Givnish, 2015) for which no genome is available. Created from tree data found in Soltis et al.* (2011)*, visualized in FigTree* (Rambaut, 2009) *and manually edited in InkScape.*

Data processing

For taxa lacking GO annotation but having putative genes already identified (e.g., *Cephalotus follicularis*), FASTA-formatted amino acid sequence data was downloaded. The remaining samples (*Actinidia chinensis*, *Ocimum tenuiflorum*, *Drosera capensis*, and *Utricularia gibba*) lacked any usable annotation data. While ideally genes and gene functions are predicted

by in-depth transcriptomic studies, the training of species-specific gene identification models, and then confirmed by individual-gene experimental studies, this is simply unfeasible for studies of diverse sets of non-model taxa. Instead, predictions of genes had to be made on the simpler basis of reading frame detection. Unannotated genomes were downloaded as FASTA-formatted nucleotide sequence and processed with ORFFinder (Wheeler et al., 2003) using parameters: [-ml 450 –n false]. These parameters were selected to identify putative genes and extract the predicted amino acid sequence. While some error in gene prediction are still likely from this method, parameters were set with the hope of preventing truncated or erroneously-predicted genes from entering the pipeline, e.g. very short of less than 150 amino acids and those contained entirely within the reading frame of another longer gene. Amino acid sequence data was then analyzed via BLAST-P on the Ohio Supercomputer (Ohio Supercomputer Center, 1987) with the following parameters: [-db nr -task blastp-fast -seg yes -num_alignments 10 -max_hsps 2 -evalue 1e-3], searching against the non-redundant protein sequence database (Pruitt, Tatusova & Maglott, 2007). BLAST outputs were imported into Blast2GO (Conesa et al., 2005; Conesa & Götz, 2008) and matched to GO codes using the automated "Mapping" function. Exported mapping results were then processed via the custom "AnnotationConverter.pl" script, to convert data into a more accessible simplified text format. For taxa already accompanied by GO-coded gene annotations (*Arabidopsis thaliana*, *Boea hygrometrica, Glycine soja, Oryza sativa,* and *Genlisea aurea*), GenBank GBFF files were downloaded. The custom Perl script "GBFFConverter.pl" was used to extract genes with associated GO information as simplified text. Using the "Functionalizer.pl" Perl script, the resulting text data was then scanned for GO codes matching to the hypothesized carnivory-associated functions selected. Counts of carnivory-associated genes were weighted against total number of genes for which at least one

function could be assigned, with the resulting proportions (count of function, per thousand genes) used for subsequent statistical analyses. This process is summarized graphically in Figure 3. Putative genes that could not be assigned to any function, or that were assigned functions that could not be mapped to any GO codes, were not included in total gene counts or proportional weighting of data. By using a conservative E-value parameter in BLAST assignment of protein functions, we hoped to filter out low-certainty annotations, particularly those potentially arising from erroneously-predicted protein sequences. Following this process, a data normalization step was performed to correct for differences in tendency to detect certain functions in BLAST searches vs. from GenBank annotation data.

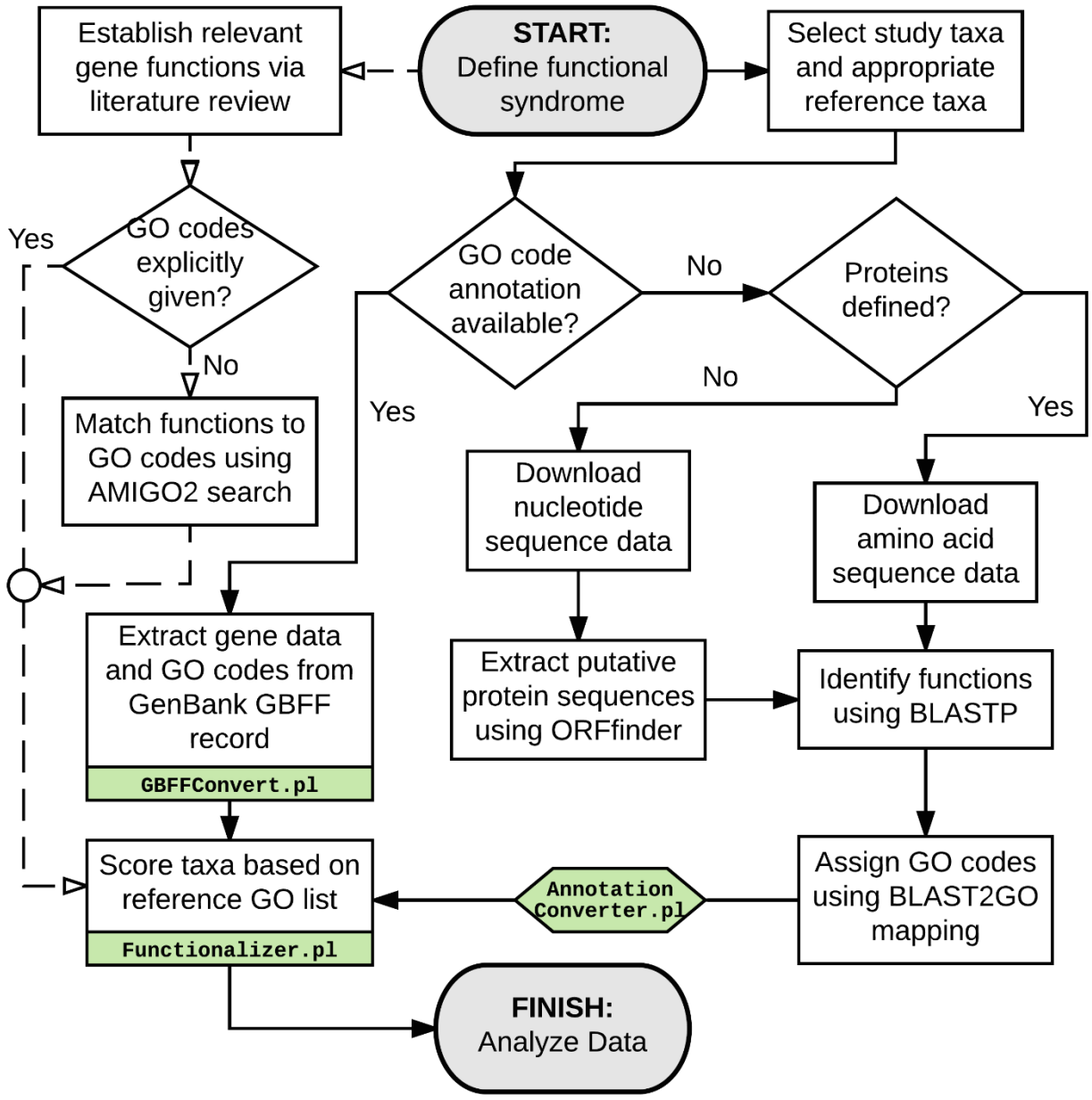

Figure 4: Flowchart detailing the preparation and processing steps to obtain gene function representation data used for subsequent statistical analyses. *Solid lines indicate processing of sampled taxa, while dashed lines indicate preparation of the reference functional set by which the taxa will be evaluated. Green boxes indicate stages utilizing custom data-processing scripts.*

To correct for differences in the likelihood of assigning a given function between samples accompanied by previous annotation and those coded using BLAST and mapping, *Arabidopsis thaliana* was analyzed by both methods, with the additional data set following the nucleotide sequence data preparation steps detailed above. The raw results of BLAST-annotated data were then multiplied by the quotient of the pre-annotated data results and the *A. thaliana* BLAST data results to produce corrected gene representation data. These data, along with pre-annotated samples that did not require correction, were used in all statistical tests; raw data were subjected to the same analyses, to ensure that the magnitude of changes in results would not be extreme (suggesting the need for more complex methods of error correction). The overall assessment of carnivory-associated function ("Total Carnivorous" vs. "None of the Above") was recalculated for each taxon from the adjusted values of each function and the total gene count ("Total"). Statistical significance was considered to have six levels ("NS", ".", "*", "**", "***", "****"); number of levels changed – either increasing or decreasing – were noted. The raw values used in these corrections are listed in Table 11.

Statistical analyses

Species were divided into "carnivorous" and "non-carnivorous" groups and analyzed on 25 criteria (24 carnivory-associated functions, plus the sum representation of all carnivory-associated functions in the genome) using a series of upper-tailed t-tests. To correct for multiple tests, Storey's correction, which uses a Bayesian approach to determine realistic false discovery rate (FDR) for the numerous tests involved in genome-wide studies (Storey, 2003; Storey & Tibshirani, 2003; Dabney, Storey & Warnes, 2010) was applied, with resulting q-values used to assess significance ($\alpha = 0.05$).

Carnivorous taxa were also tested individually, against reference normal distributions created by assessing the values seen in non-carnivorous taxa. Twenty-five reference distributions were used (24 functions + overall), each defined by the median and standard deviation value determined for that function in the non-carnivorous reference taxa. Statistical evaluations were conducted via a series of upper-tailed Z-tests, with Storey's correction then used within each series of tests (four sets of 25 tests) to account for repeat testing effects.

Results

A general assessment of all included genomes for genome size, total gene number, and number of unique Gene Ontology codes identified (a representation of diversity of functions encoded) showed largely overlapping ranges of values (Table 2). Both the largest and smallest genomes analyzed were to carnivorous plants: *Genlisea aurea* at 43.3 Mb and *Cephalotus follicularis* with 1.6 Gb (non-carnivorous plants ranged from 119.7 Mb to 1.5 Gb). Similar results were found for number of genes encoded, ranging from 17,685 in *G. aurea* to 89,073 in *Drosera capensis* (typical plants: 28,382 to 70,250). The largest number of unique GO codes identified was found in *Arabidopsis thaliana*; however, this results from the utilization of *A. thaliana* in the development of plant GO codes. The smallest number was found in *Oryza sativa* (1262), with the largest (after *A. thaliana*) found in *C. follicularis* (4664). Interestingly, when testing for relationships between these factors, no significant ($\alpha = 0.05$) associations were found between genome size and gene number ($p = 0.857$, $R^2 = 0.1202$), genome size and number of unique GOs ($p = 0.630$ $R^2 = 0.0909$), or gene number and number of unique GOs ($p = 0.901$, $R^2 = 0.1227$).

	Sequence (Mb)	# Genes	% Results	GO Hits	Unique GOs
A. chinensis	604.2	70,250	54.5%	182,315	3830
A. thaliana	119.7	48,350	56.3%	173,184	6503
B. hygrometrica	1521.3	47,778	23.2%	76,891	2916
G. soja	863.6	50,399	51.4%	76,044	1421
O. sativa	382.8	28,382	45.7%	35,445	1262
O. tenuiflorum	321.9	34,920	47.8%	76,891	2916
C. folicularis	1614.5	36,667	42.2%	80,567	4664
D. capensis	263.8	89,073	24.5%	113,593	3723
G. aurea	43.3	17,685	96.6%	79,194	4883
U. gibba	100.7	32,621	40.5%	64,529	3276

Table 2: General statistics of the plant genomes included in this study. *"Sequence (Mb)" indicates the available genome sequence in million base pairs. "# Genes" indicates the number of putative genes identified, either as indicated in GenBank documentation or detected via ORFfinder. "% Results" indicates the portion of genes that could be associated with at least one GO code. "GO Hits" indicates the total number of GOs matched to a gene across all genes. The number of unique codes present in this number is given as "Unique GOs".*

When comparing proportion of functionally identifiable genes that could be mapped to a carnivory associated function, there was little difference between carnivorous and non-carnivorous taxa. The percentage of genes mapping to a carnivory associated function in carnivorous taxa ranged from 3.1% to 5.0% of all function-assigned genes; in typical plants, this value ranged from 2.4% to 4.3%. In terms of which specific carnivory associated functions made up each plant's proportion, the representation of each function varied wildly from taxon to taxon (Figure 5; Table 12).

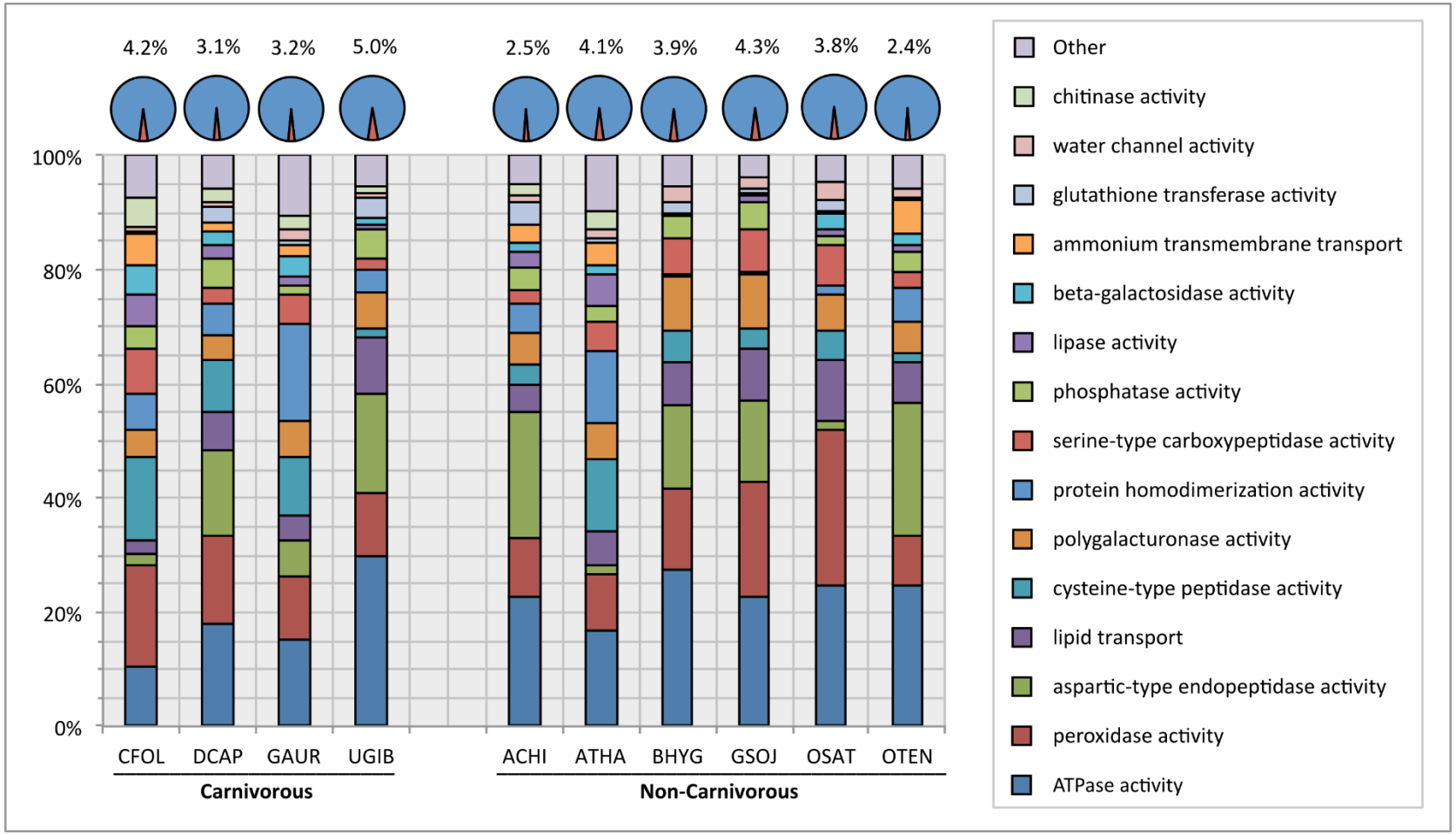

Figure 5: Chart of proportional representation of carnivorous functions vs. overall gene functions in all taxa sampled. *Pie-charts above indicate total proportion of all carnivorous functions combined (red & percentage) vs. all other genes for which at least one function could be identified (blue). Stacked bars below indicate the proportion ascribed to each carnivory-associated function within the total, sorted from (on average) most represented (bottom) to least represented (top). The final bar, "Other", combines the rarest nine functions, which each on average represent only 0.7% of the carnivory-associated functions detected. A complete numerical view of this data is available in Table 12.*

Statistical comparisons of the genomic representation of each carnivory associated function in carnivorous vs. typical plants yielded two significant ($\alpha = 0.05$) results: "Alternative oxidase activity" ($t = 3.14$, $p = 0.011$, $q = 0.047$) and "ATP:ADP antiporter activity" ($t = 4.00$, $p = 4.30E\text{-}03$, $q = 0.037$). A third function, "phospholipase activity" ($t = 2.79$, $p = 0.019$, $q = 0.053$), was detected as significant before correction, but retained only marginal significance ($\alpha = 0.10$) after accounting for multiple tests. Detailed results of these tests are presented in Table 3. For each test, statistical power reaches 50% ($\beta = 0.50$) at an effect size of 1.06 standard deviations and 95% ($\beta = 0.05$) at an effect size of 2.12 standard deviations ($\alpha = 0.05$).

	Z	p	q	Sig.
Actin	0.24	0.407	0.218	*NS*
AltOx	**3.14**	**0.011**	**0.047**	*
AspPep	0.01	0.496	0.241	*NS*
ATP	-0.24	0.590	0.241	*NS*
ATP_ADP	**4.00**	**4.30E-03**	**0.037**	*
BGal	1.90	0.062	0.127	*NS*
Chit	-1.82	0.944	0.324	*NS*
CinAlc	0.39	0.355	0.218	*NS*
CystPep	-0.32	0.619	0.241	*NS*
FrucBPA	0.66	0.266	0.217	*NS*
GlutTrans	0.29	0.391	0.218	*NS*
H2OChan	1.68	0.074	0.127	*NS*
HeatShock	0.32	0.377	0.218	*NS*
Lipase	0.53	0.309	0.218	*NS*
LipTrans	0.94	0.193	0.217	*NS*
NHTrans	0.64	0.278	0.217	*NS*
Perox	-0.14	0.552	0.241	*NS*
Phoslip	2.79	0.019	0.053	.
Phosp	0.77	0.240	0.217	*NS*
Polygal	-0.75	0.763	0.284	*NS*
ProtHomo	1.27	0.122	0.174	*NS*
RiboNuc	-1.10	0.841	0.300	*NS*
SerCarPep	-0.29	0.608	0.241	*NS*
ThioGluc	-0.18	0.570	0.241	*NS*
Total	0.62	0.278	0.217	*NS*

Table 3: Results of statistical analyses comparing non-carnivorous plants to carnivorous plants for each of 24 carnivory-associated functions, plus the total of all functions. *"t" indicates the test statistic of an upper-tailed Student's t-test. "p" indicates the p-value of this test. "q" indicates a corrected p-value accounting for multiple comparisons, using Storey's correction. Significance ("Sig.") is indicated by bolding and with "*" for q < 0.05, "**" for q < 0.01, and "*** for q < 0.001. A non-bolded "." indicates marginal values (q < 0.10), while "NS" indicates non-significance (q > 0.10).*

Testing of individual carnivorous taxa yielded a total of 13 significant ($\alpha = 0.05$) results and an additional five marginal ($\alpha = 0.10$) results, out of 100 total tests (4 species x 25 distributions). *Genlisea aurea* and *Drosera capensis* had very few functions that showed a signal of genomic overrepresentation. In *Genlisea aurea,* only a single function, "phospholipase activity" (Z = 2.76, p = 2.89E-03, q = 0.054) result reached marginal significance. *Drosera capensis* had one significant function: "alternative oxidase activity" (Z = 3.72, p = 1.01E-04, q = 2.45E-03). *Utricularia gibba* and *Cephalotus follicularis* were found to have a substantial portion of carnivory-associated functions showing strong signals of genomic overrepresentation. In *U. gibba*, five functions reached statistical significance: "alternative oxidase activity" (Z= 2.36, p = 9.17E-03, q = 0.025), "ammonium transmembrane transport" (Z = 3.58, p = 1.74E-04, q = 2.09E-03), "ATPase activity" (Z = 3.19, p = 7.19E-04, q = 4.31E-03), "cysteine-type peptidase activity" (Z = 2.00, p = 0.023, q = 0.046), "phosphatase activity" (Z = 2.65, p = 4.05E-03, q = 0.016), and "phospholipase activity" (Z = 2.30, p = 0.011, q = 0.025). An additional three test results were marginally significant: "aspartic-type peptidase activity" (Z = 1.75, p = 0.040, q = 0.060), "ATP:ADP antiporter activity" (Z = 1.54, p = 0.062, q = 0.083), and total proportion of carnivory-associated functions (Z = 1.79, p = 0.036, q = 0.060). In *C. follicularis*, seven functions reached significance: "alternative oxidase activity" (Z = 2.36, p = 9.10E-03, q = 0.030), "beta-galactosidase activity" (Z = 4.19, p = 1.40E-05, q = 1.99E-04), "glutathione transferase activity" (Z = 2.22, p = 0.13, q = 0.31), "lipase activity" (Z = 2.14, p = 0.016, q = 0.033), "lipid transport" (Z = 2.31, p = 0.010, q = 0.030), "phospholipase activity" (Z = 3.50, p = 2.28E-04, q = 0.002), and "water channel activity" (Z = 3.08, p = 1.04E-03, q = 4.96E-03). One additional function, "ATP:ADP antiporter activity" (Z = 1.91, p = 0.028, q = 0.050), was marginally significant. Detailed results of these tests are presented in Table 4. For each test,

statistical power reaches 50% ($\beta = 0.50$) at an effect size of 2.62 standard deviations and 95% ($\beta = 0.05$) at an effect size of 4.94 standard deviations ($\alpha = 0.05$).

	Genlisea aurea				*Drosera capensis*				*Utricularia gibba*				*Cephalotus follicularis*			
	Z	**p**	**q**	**Sig.**	**Z**	**p**	**q**	**Sig.**	**Z**	**p**	**q**	**Sig.**	**Z**	**p**	**q**	**Sig.**
Actin	0.25	0.403	0.626	*NS*	-0.69	0.755	0.861	*NS*	-0.06	0.525	0.351	*NS*	1.04	0.150	0.214	*NS*
AltOx	0.75	0.228	0.416	*NS*	**3.72**	**1.01E-04**	**2.45E-03**	**	**2.36**	**9.17E-03**	**0.025**	*	**2.36**	**9.10E-03**	**0.030**	*
AspPep	-0.74	0.771	0.674	*NS*	0.20	0.421	0.861	*NS*	1.75	0.040	0.060	.	-1.17	0.880	0.544	*NS*
ATP	-1.50	0.933	0.697	*NS*	-1.22	0.889	0.861	*NS*	**3.19**	**7.19E-04**	**4.31E-03**	**	-1.67	0.952	0.544	*NS*
ATP_ADP	1.52	0.064	0.238	*NS*	1.71	0.043	0.523	*NS*	1.54	0.062	0.083	.	1.91	0.028	0.050	.
BGal	1.92	0.028	0.172	*NS*	0.55	0.291	0.861	*NS*	0.55	0.291	0.296	*NS*	**4.19**	**1.40E-05**	**1.99E-04**	***
Chit	-0.35	0.637	0.674	*NS*	-1.19	0.884	0.861	*NS*	-0.70	0.757	0.392	*NS*	-1.03	0.847	0.544	*NS*
CinAlc	1.22	0.111	0.345	*NS*	-0.39	0.650	0.861	*NS*	0.31	0.379	0.311	*NS*	-0.29	0.616	0.482	*NS*
CystPep	-1.15	0.876	0.697	*NS*	-0.55	0.708	0.861	*NS*	**2.00**	**0.023**	**0.046**	*	-1.43	0.923	0.544	*NS*
FrucBPA	1.56	0.059	0.238	*NS*	-0.64	0.739	0.861	*NS*	0.75	0.227	0.273	*NS*	-0.02	0.508	0.454	*NS*
GlutTrans	-0.19	0.577	0.674	*NS*	-0.20	0.578	0.861	*NS*	-0.90	0.817	0.392	*NS*	**2.22**	**0.013**	**0.031**	*
H2OChan	0.74	0.229	0.416	*NS*	0.65	0.259	0.861	*NS*	0.48	0.314	0.296	*NS*	**3.08**	**1.04E-03**	**4.96E-03**	**
HeatShock	0.69	0.244	0.416	*NS*	-0.74	0.771	0.861	*NS*	-0.07	0.527	0.351	*NS*	0.83	0.204	0.224	*NS*
Lipase	-0.26	0.602	0.674	*NS*	-0.02	0.508	0.861	*NS*	-0.35	0.637	0.376	*NS*	**2.14**	**0.016**	**0.033**	*
LipTrans	0.69	0.245	0.416	*NS*	0.41	0.340	0.861	*NS*	-0.65	0.741	0.392	*NS*	**2.31**	**0.010**	**0.030**	*
NHTrans	-0.82	0.794	0.674	*NS*	0.98	0.163	0.861	*NS*	**3.58**	**1.74E-04**	**2.09E-03**	**	-0.86	0.806	0.544	*NS*
Perox	-0.59	0.722	0.674	*NS*	-0.24	0.596	0.861	*NS*	0.00	0.499	0.351	*NS*	0.57	0.284	0.270	*NS*
Phoslip	2.76	2.89E-03	0.054	.	0.31	0.379	0.861	*NS*	**2.30**	**0.011**	**0.025**	*	**3.50**	**2.28E-04**	**0.002**	**
Phosp	-1.35	0.912	0.697	*NS*	0.69	0.244	0.861	*NS*	**2.65**	**4.05E-03**	**0.016**	*	0.83	0.203	0.224	*NS*
Polygal	-0.46	0.678	0.674	*NS*	-1.13	0.870	0.861	*NS*	0.47	0.320	0.296	*NS*	-0.45	0.674	0.482	*NS*
ProtHomo	2.09	0.018	0.169	*NS*	0.11	0.458	0.861	*NS*	0.28	0.388	0.311	*NS*	0.62	0.269	0.270	*NS*
RiboNuc	-0.62	0.734	0.674	*NS*	-0.42	0.661	0.861	*NS*	-0.34	0.635	0.376	*NS*	-0.43	0.668	0.482	*NS*
SerCarPep	-0.32	0.626	0.674	*NS*	-0.92	0.821	0.861	*NS*	-0.87	0.808	0.392	*NS*	1.35	0.089	0.142	*NS*
ThioGluc	0.85	0.198	0.416	*NS*	-0.41	0.658	0.861	*NS*	-0.41	0.658	0.376	*NS*	-0.41	0.658	0.482	*NS*
Total	-0.38	0.649	0.674	*NS*	-0.56	0.711	0.861	*NS*	1.79	0.036	0.060	.	0.86	0.195	0.224	*NS*

Table 4: Results of statistical analyses comparing non-carnivorous plants to carnivorous plants in four sets, with each evaluating 24 carnivory-associated functions, plus the total of all functions. "Z" indicates the test-statistic of an upper-tailed Z-test (equal to number of standard deviations from the mean). "p" indicates the p-value of this test. "q" indicates a corrected p-value accounting for multiple comparisons, using Storey's correction. Significance ("Sig.") is indicated by bolding and with "*" for $q < 0.05$, "**" for $q < 0.01$, and "*** for $q < 0.001$. A non-bolded "." indicates marginal values ($q < 0.10$), while "NS" indicates non-significance ($q > 0.10$)

Some changes in results of the above analyses were observed when testing with the raw data sets. In analysis 1 (Table 13), a t-test comparison of carnivorous vs. non-carnivorous taxa as groups, *beta-galactosidase activity*, *phosphatase activity,* protein *homodimerization, thioglucosidase activity*, and *water channel activity* were noted as marginally significant when using uncorrected data; these values dropped below marginal significance ($q < 0.10$) when using corrected data. *Phospholipase* activity was identified as significant ($q < 0.05$) from uncorrected data, but decreased to marginal significance after correction. In analysis 2 (Table 14), changes were as follows: *Genlisea aurea*: no changes. *Drosera capensis*: *ATP:ADP antiporter activity, phosopholipase activity,* and *thioglucosidase activity* declined from significant ($q < 0.05$) to NS ($q > 0.10$), while *cysteine-type peptidase activity* declined from marginal to NS. *Utricularia gibba*: *ammonium transmembrane transport activity* and *ATPase activity* increased from NS to ** ($q < 0.01$), *cysteine-type peptidase activity,* and *phospholipase activity* increased from NS to significant, and *aspartic-type peptidase activity* and total carnivorous function increased from NS to marginal; however, *phosphatase activity* declined from ** to significant, and *cinnamyl-alcohol dehydrogenase activity, polygalactosidase activity,* and *protein homodimerization activity* declined from significant to NS. *Cephalotus follicularis*: *beta-galactosidase activity* increased from NS to *** ($q < 0.001$), *phospholipase activity* and *water channel activity* increased from NS to **, *glutathione transferase activity* increase from marginal to significant, and *lipase activity* and *lipid transferase activity* increased from NS to significant; *heat shock protein activity* decreased from **** ($q < 0.0001$) to NS, *ATP:ADP antiporter activity* declined from ** to marginal, *ATPase activity* and *protein homodimerization activity* decreased from ** to NS, *polygalactosidase activity* declined from significant to NS, and *fructose bisphosphate aldolase activity* declined from marginal to NS.

In total across all tests, 81% of results remained unchanged in designation (Table 5). For those that did change, a decrease in significance was more common (10%) than an increase (8%). Less than half of all changes (7% of total tests) were of more than a single significance level. Thus, the analyses presented in the main text are performed using the adjusted data.

	No Change	**Increase**	**Decrease**	**Change > 1**
Class	19 (76%)	0 (0%)	6 (24%)	0 (0%)
Individual	83 (83%)	10 (10%)	5 (5%)	9 (9%)
C. follicularis	17 (68%)	6 (24%)	2 (8%)	6 (24%)
D. capensis	21 (80%)	0 (0%)	4 (16%)	2 (8%)
G. aurea	25 (100%)	0 (0%)	0 (0%)	0 (0%)
U. gibba	20 (80%)	4 (16%)	0 (0%)	1 (4%)
Overall	102 (82%)	10 (8%)	12 (10%)	9 (7%)

Table 5: Effects of data adjustment on statistical significance detected in results. *Change was measured in significance levels, considering six levels: q > 0.10 (NS), q < 0.10 (.), q < 0.05 (*), q < 0.01 (**), q < 0.001 (***), q < 0.0001 (****). "Increase" shows cases where a result went up one or more significance levels; "Decrease" shows cases where a result went down one or more significance levels; "Change > 1" shows cases where a result went either up or down two or more significance levels. "Class" indicates the results based off the categorical correction seen in Table 13 vs. Table 3. "Individual" indicates the total of individual comparison results as shown by Table 14 vs. Table 4; species names show these comparisons for each species considered separately. "Overall" shows the total of all tests.*

Discussion

The analyses presented here were designed to identify similarities in function among carnivorous plants, and we found mixed support for our hypotheses. The null hypothesis ("H_0: Carnivorous plant genomes are not distinct from typical plants in functional terms") cannot be rejected for 11 of the 24 functions tested. The first alternate hypothesis ("H_1: All carnivorous plants contain a shared functional signal as a result of convergence") is given some support by

the results of statistical comparisons between carnivorous and typical plants overall, as it does appear that alternative oxidase activity and ATP:ADP anti-porter activity (as well as, potentially, phospholipase activity) may be commonly overrepresented in carnivorous taxa. Our results support the other alternative hypothesis ("H_2: Carnivorous plants are distinct in gene function from typical plants, but this difference varies from taxon to taxon"), as seen in nine (and one additional, marginally) of the 24 functions tested. In short, only a small number of functions appear to be consistently over-represented in taxa sharing the syndrome of plant carnivory; others, while from a predictable set, are over-represented on a taxon-to-taxon basis but in an unpredictable manner. A majority of functions, even if involved in the functional syndrome described, will likely not show a detectable signal, either due to high levels of variation within the control group or because other methods of up-regulation (transcriptional, translational, or structural) have been employed. In any case, our study suggests that plant carnivory can evolve using multiple independent metabolic pathways.

Overall effects

This study sought to detect a signal of genomic overrepresentation of functions researchers had previously determined were associated with carnivory in plants. The two functions consistently identified as significantly overrepresented were "alternative oxidase activity" and "ATP:ADP anti-porter activity". Alternative oxidase functions primarily in the mitochondria, as part of the electron transport chain. It is believed to function as a "protective" enzyme, to prevent over-oxidation in the mitochondria and can be activated in response to oxidative stress (Day & Wiskich, 1995). For this function to be of common importance to carnivorous taxa, there are three possible explanations: (i) carnivorous plants, due to their

digestive function, produce larger amounts of reactive oxygen species (consistent with Chia et al., 2004), requiring more alternative oxidase to counteract their negative effects; (ii) alternative oxidase is encoded in a modified form, having been co-opted to perform a different function outside the mitochondrion; or (iii) alternative oxidase is functioning as it would typically, but due to the similar habitat parameters of carnivorous plants, they require its effects more frequently. ATP:ADP anti-porter activity has two functions: It is involved in the maintenance of cellular electrical potential (due to an H+ gradient) in the presence of free fatty acids (Vianello, Petrussa & Macrì, 1994) and it allows exchange of cellular ATP for plastid or bacterial-symbiont ADP (Greub & Raoult, 2003). In the first case, it may be responsible for interacting with the cellular proton gradient if pH changes substantially during digestion; in the second, it may provide aid to the symbiotic bacteria that assist carnivorous plants in digestion. A third function, "phospholipase activity", was marginally significant. Phospholipases are involved in signaling interactions as well as in metabolism of fatty acids and in degrading cell membranes (Chapman, 1998). Carnivorous plants may possess an increased need for complex signaling pathways to regulate their digestive machinery, as well as a clear need to break down cell membranes to access the contents of insect cells.

Individual taxa

In individual analyses of each carnivorous taxon, alternative oxidase was found to be significant in three species and phospholipase in two species (plus one marginal). Interestingly, ATP:ADP anti-porter activity overall signal was driven by marginal results in two taxa. *Genlisea aurea* and *Drosera capensis* had no significant or marginal functions outside of this set (with one each). In stark contrast, *Utricularia gibba* and *Cephalotus follicularis* both had large numbers of

significantly overrepresented carnivory associated functions. Other than alternative oxidase (significant in both), phospholipase (significant in both), and ATP:ADP anti-porter activity (marginal in both), the two taxa did not overlap in any of their other nine (combined) over-represented functions.

Utricularia gibba uniquely possessed overrepresentation in *ATPase activity*, *cysteine-type peptidase activity*, *ammonium transmembrane transport*, *phosphatase activity*, and *aspartic-type endopeptidase activity*. Phosphatase, aspartic peptidase, and cysteine peptidase, as catabolic enzymes found localized to the digestive fluids of other carnivorous taxa (Schulze et al., 2012; Rottloff et al., 2016), most likely have roles in direct digestive function. *Ammonium transmembrane transport*, while required in some amount by all plants, may be more vital for *Utricularia*, which must extract the concentrated nitrogenous products of digestion from an aquatic environment. ATPase in plants is involved in regulation of endocytotic and secretory processes (Dettmer et al., 2006), which would logically be involved in both the release of digestive enzymes and the absorption of digested material. *Utricularia gibba* was the only taxon studied that had even marginal significance in the total genomic proportion of carnivorous functions. Also of note is the vast difference in portion of carnivory-associated functions between *U. gibba* and its close relative *Genlisea aurea*. While both taxa have characteristically-reduced genomes, *G. aurea* has approximately half the genome size and gene number of *U. gibba* (Table 2). It may be that in *Genlisea*, selective pressure strongly favored deletion of duplicated genes, with up-regulation or modification instead occurring at the transcriptional or translational stage.

In *Cephalotus follicularis*, *beta-galactosidase activity*, *water channel activity*, *glutathione transferase activity*, *lipase activity*, and *lipid transferase activity* were found to be uniquely

overrepresented. Lipase and beta-galactosidase (which breaks polysaccharide bonds) are likely to have direct involvement in digestion, having also been found in the digestive fluids of other carnivorous taxa (Schulze et al., 2012; Rotloff et al., 2016); lipid transferase would logically accompany lipase, either to localize lipid substrates or to move the products of their decomposition. As *C. follicularis* must transfer water to the interior of its pitchers for digestive functions to be possible at all, high levels of *water channel activity* is also a logical finding.

Non-significant functions

Conversely, 11 functions (*actin*, *chitinase activity*, *cinnamyl-alcohol dehydrogenase activity*, *fructose bisphosphate aldolase activity*, *heat shock protein activity*, *peroxidase activity*, *polygalactonuronase activity*, *protein homodimerization activity*, *ribonuclease activity*, *serine-type carboxypeptidase activity*, and *thioglucosidase activity*) showed no significant over-representation in any taxa sampled. However, due to the relatively low statistical power to detect low to moderate effect sizes with the tests performed, it is possible that these effects do exist but cannot be detected. Even if accepting these negative results as accurate, it is possible that these functions are preferentially utilized in other ways, such as increased transcription, increased protein translation, or increased protein efficiency due to changes in amino acid sequence. Any of these scenarios may also explain why certain functions are overrepresented in the genomes of some carnivorous taxa but not in others.

Conclusions

The findings of this study are consistent with expectations of evolutionary convergence. As distant taxa converge on a similar phenotype, predictable functional convergence occurs. This was seen in the cases where there predicted functions, gathered from past studies of carnivorous taxa, were determined to be significantly overrepresented in the taxa sampled. However, this effect was not seen in all functions predicted, nor were the functions showing significant overrepresentation consistent across all four taxa. It is likely that, while these taxa may often show strong signal in some of the functions predicted, the number of potential avenues by which to reach the same practical result is too great for any prediction to hold true in all cases.

The degree of molecular specificity required to meet an organism's needs can also be expected to play a role, with ability to predict a specific functional set increasing proportional to specificity of the convergent syndrome. In carnivorous plants, a wide range of morphologies (as evidenced by the taxa included in this study) have arisen to reach the same end. In other cases, there is little flexibility in how an organism can reach the needed outcome. For example, organisms that rely on the mimicry of pheromones, such as orchids that imitate bee sex and alarm pheromones, are far less likely to show variation in the functions required for the end-result (Stökl et al., 2005, 2007; Brodmann et al., 2009). Conversely, even broadly-defined, frequently re-derived evolutionary syndromes may still show repeated selection for specific functional codes. It has been shown that organisms experience substantial convergence of microbiome even for classes as broad as "carnivore" vs. "herbivore" (Muegge et al., 2011); it is reasonable to consider that this occurrence may be accompanied by host genome functional convergence as well. However, to detect a signal in these broader groups, where it may be

difficult to assemble a manageable list of target syndrome-associated functions, much more thorough sampling would likely be required.

Future Directions

This study is currently limited primarily by the lack of available genomic sequence data for carnivorous taxa, as well as the lack of thorough Gene Ontology annotation of plant taxa in general. This study's BLAST-based annotation methodology is currently impractical for substantially larger taxon sampling, and even in limited taxon sets, greater accuracy is desirable. As more annotated genomes, more consistently high-quality genome assemblies, and more accompanying transcriptomic data sets on which to train gene prediction models become available, it will be possible to more thoroughly assess this phenomenon. As more carnivorous plant taxa are sequenced and annotated (*Nepenthes* and *Dionaea* are expected, as well as *Sarracenia* by the authors), it also becomes possible to refine the reference GO set created for this study, e.g. using functions implicated in previous studies in at least 3 of 10 taxa. Another potential approach is to apply similar methods to a different functional syndrome. While results may differ based on the evolutionary idiosyncrasies of groups of organisms or from one specific syndrome to another, the same methods could be employed.

Chapter 3: *De novo* assembly, annotation, and analysis of *Sarracenia alata*, a carnivorous plant

Abstract

Sequencing and assembly of large plant genomes has remained difficult due to their large, repetitive genomes enriched in pseudogenes that likely result from recurrent genomic duplication events. Typical strategies often exceed resource use expectations but fall far short of their benchmarks, resulting in the vast majority of published plant genomes being highly fragmentary and largely unannotated. Our investigation explores assembly and annotation methods that leverage multiple sequence data types, parallel assemblies, and incremental merging to improve contiguity of the assembly of a large plant genome. It tests an annotation methodology focusing on the pre-processing of reference databases to eliminate unnecessary comparisons and reduce computation times. The genome of the carnivorous plant *Sarracenia alata* is assembled and annotated, and used to test the hypothesis that *S. alata* is entirely reliant on its microbiome and symbiotic organisms to digest prey.

The assembly produced for *S. alata* is of comparable quality to other plant genomes of similar size and complexity. While contiguity does not approach the desired chromosome-length quality, contig lengths (N50: 35,650) were sufficient for genome annotation. Annotation predicted 28,750 genes, of which 64.9% could be assigned functional codes; 36,979 pseudogenes were also detected. Genic sequence comprises 4.9% of the genome, while 3.2% is composed of pseudogenes, and the remaining portion is made up of repetitive elements; of this repetitive sequence, Long Terminal Repeats (LTRs) are the vast majority.

Our results indicate that preliminary filtering of the plant protein and *Sarracenia* RNA databases improved the efficiency of the MAKER-P annotations. However, incorporating additional sequence assembly stages did little to improve on the functions of the Canu and Pilon steps in the assembly. The gene annotations produced by this pipeline were sufficient to test the prediction that *S. alata* lacks genes that produce products with obvious carnivorous function, and a strong signal of overrepresentation of such genes was observed. *S. alata* exceeded all other carnivorous taxa tested in both its total carnivory gene proportion and individual functions found to be significant Thus, the hypothesis that *S. alata* is enzymatically inactive is rejected, warranting further study into the relative contributions of the plant and its microbiome in the digestion of prey.

Keywords

plant genome, assembly, annotation, *de novo*, carnivorous plant, Canu

Introduction

Sarracenia alata

The North American pale pitcher plant, *Sarracenia alata* Wood, is a Gulf Coast endemic with a patchy, disjunct distribution, found only in nutrient-poor wetland areas such as bogs and pine savannas (Zellmer et al., 2012). The plant's leaves have been thoroughly modified into tall, lidded pitchers, which possess extrafloral nectaries to lure their primary prey item, ants

(Bhattarai & Horner, 2009). Over time, prey is digested, allowing the plant to obtain nitrogen, phosphorus, and mineral resources which the environment lacks; however, despite its seemingly thorough adaptation to a carnivorous lifestyle, it has been argued that *Sarracenia* species possess no actual carnivorous ability (Anderson & Midgley, 2003). The plant lacks the macrocellular digestive enzyme glands seen in other (unrelated) carnivorous plant groups, *Nepenthes* and *Cephalotus* (Adams & Smith, 1977; Płachno et al., 2006), despite the presence of digestive enzymes within its fluid. Instead, it has been hypothesized that *Sarracenia* are wholly reliant on the activities of symbiotic microbes. The diversity of microbiota in *Sarracenia alata* fluid is well-documented, and their similarity to the gut microflora of animals has also been noted (Koopman et al., 2010; Koopman & Carstens, 2011).

Genome estimates and expectations

The genome size of *S. alata* is most likely similar to that of its congeners because there is no variation in chromosome number (N=26) in the genus (Rogers et al., 2010) and because the species are relatively closely related (Ellison et al., 2012). The most recent estimates of genome size in *Sarracenia,* based on flow cytometry analyses in *Sarracenia purpurea* and *S. psittacina* (Rogers et al., 2010), suggest that *S. alata's* genome should be ~3.6 Gbp. However, a signal of genome duplication has been reported in *Sarracenia* (Srivastava et al., 2011). Using the Blanc and Wolfe (2004) method, Srivastava et al. (2011) suggest that a whole genome duplication event may have occurred some 2 million years ago, at a time that may correspond to the origin of the clade (Ellison et al., 2012). While the genome is considered to be diploid, signs of polyploidy (e.g., multiple banding) remain (Rogers et al., 2010), leading some to suggest species in the genus are "partial polyploid[s]" (Stephens et al., 2015) or have a history of "paleopolyploidy"

(Stace et al., 1997), as seen in other Ericales (Shi, Huang & Barker, 2010). The genome of *S. alata* has likely been inflated by past duplication events, and may be enriched in repetitive elements and pseudogenes.

Functional genomics of plant carnivory

This investigation addresses the question of whether *S. alata* possesses the necessary genetic machinery to produce digestive enzymes using a whole-genome approach. Following genome assembly and annotation, identified genes were assigned Gene Ontology (GO) identifiers to quantitatively group genes on the basis of function (Ashburner et al., 2000b; Carbon et al., 2017). Using the approach of Chapter 2, functions identified *a priori* as likely to be associated with plant carnivory were identified, counted, and evaluated against other carnivorous taxa and non-carnivorous reference taxa. As genes associated with adaptation and radiation are likely to increase output and diversify function through duplication (Moore & Purugganan, 2005; Flagel & Wendel, 2009), a strong genome-scale signal of overrepresentation of functional categories associated with plant carnivory were taken as indicative of adaptation to an actively-carnivorous lifestyle.

Improved assembly and annotation

As the number of sequenced genomes grows, the possibilities for their use increase, allowing for a wider range of comparative functional genomic and phylogenomic studies to be conducted at both broader and finer scales, as well as increasing the feasibility of sequencing new genomes by providing usable references (Chain et al., 2003; Giribet, 2016).Unfortunately, the availability of thoroughly sequenced and annotated plant genomes has remained more limited than animal genomes, with only 88 annotated plant genomes listed versus 372 animal genomes (GenBank, NCBI). This is due to the complex, degenerate nature of large plant genomes, which

can possess high, mixed, or uncertain ploidy, low differentiation between genes and pseudo genes, and a greater diversity of plant repetitive elements and compared animal genomes (Schatz, Witkowski & McCombie, 2012). Even with access to powerful supercomputer resources, the time required to perform a thorough annotation of a plant genome without relevant pre-annotated reference samples can make the task intractable. This study seeks to test a modified assembly and annotation pipeline, with the hope of allowing individuals and small groups to tackle the bioinformatics of large plant genomes.

Methods

Extraction

Low molecular weight

Low molecular weight DNA was extracted using the QIAGEN DNeasy Plant DNA extraction kit, in five reactions. Five tissue sections of approximately 1 cm^2 were excised from a leaf collected from *S. alata* in a wild population (Abita Springs, Louisiana, USA, 2014). Samples were macerated via bead mill in 100 uL QIAGEN kit extraction buffer. Following the kit extraction protocol, DNA was quantified via Qubit before the desired mass was dried via vacuum centrifuge prior to shipping.

High molecular weight

The extraction of high molecular weight DNA for *S. alata* presented problems, likely due to the tough, waxy composition of the pitchers and the possible presence of secondary

compounds. A modified CTAB extraction (Wolfe, 2005) was attempted, taking into account published guidelines for carnivorous plant DNA extraction (Fleischmann & Heubl, 2009), but no usable DNA was obtained after multiple attempts. Instead, ultra-high-quality DNA extracted for use in BioNano optical sequence mapping was used, following processing to SMRT-appropriate fragment lengths.

Sequencing

Illumina

Short-read sequencing was performed using QIAGEN-extracted DNA on the Illumina HiSeq 2000 platform, producing 410 million paired-end reads (100 bp, 11.4x coverage). Quality-control assessment was performed using *FastQC* (Andrews, 2010)). Sequences possessed a per-read Phred score of 37.5 (>99.9%), with a minimum per-base score of 21(99.2%). Sequence data passed all relevant quality control tests (Figures 14-20).

PacBio SMRT

Long-read sequencing was performed on the PacBio Sequel platform, using size-selected DNA fragments from BioNano extraction and preparation. Total length of sequence obtained was 122.3 Gbp (34.0x coverage) in 9.1 million reads. The average read length was 6.3 kbp, while N50 was 21.3 kbp. A distribution of the fragments sequenced is presented in Figure 14.

BioNano optical mapping

Optical mapping of ultra-long sequence fragments was performed on the BioNano Irys platform (Staňková et al., 2016), using the BspQI “nickase” enzyme. Molecule data were filtered to a length of 150 kbp and a marker density of 5 labeled sites per 100 kbp. The usable data obtained amounted to 1.1 million molecules, for the coverage-equivalent of 237.7 gigabase of

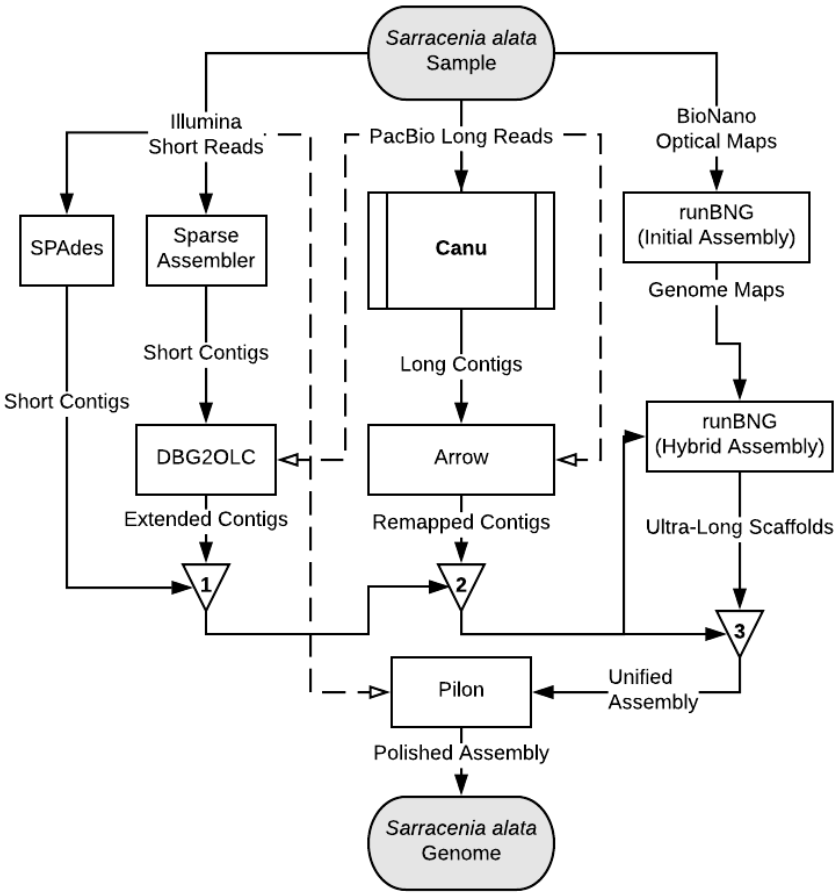

Figure 6: Genome assembly pipeline. *Starting with extracted DNA of Sarracenia alata, three NGS libraries are prepared: short Illumina reads, long PacBio SMRT reads, and ultra-long BioNano optical sequence maps. Sequences of the three types are assembled de novo in parallel (Illumina by two methods), and then passed forward for hybrid assembly or elongation via realignment. Assemblies of parallel methodologies were then merged via QuickMerge (indicated as triangle joiners). Polishing and error correction using high-accuracy Illumina reads was performed, producing the final assembly. Dotted lines indicate reads re-used outside their original assembly, for hybrid assembly, improvement, or quality control.*

sequence (66.0x coverage). The average molecule had a length covering 216.1 kbp, with 14.7 labeled sites (6.8/100 kbp).

Genome Assembly

An assembly pipeline was designed to best leverage the strengths of PacBio, Illumina, and BioNano data sets (Fig. 6). First, each set was assembled *de novo*, followed by hybrid assembly and/or assembly-merging with other data types, and a final polishing stage to increase individual-nucleotide accuracy.

Initial assembly

The *de novo* long-read assembler *Canu* (Koren et al., 2017), a branch of *Celera Assembler* (Myers et al., 2000), serves as the core of this pipeline. Taking raw PacBio reads as input, *Canu* utilizes highly-parallelized job scheduling (with some stages broken into over 13,000 individual jobs) and iterative trimming and filtering stages, to produce a high-quality primary assembly.

In order to create the necessary starting-point for downstream hybrid assembly via *DBG2OLC* (Ye et al., 2015), Illumina data were assembled *de novo* in two stages; first using *SPAdes* (Bankevich et al., 2012) with kmer sizes 21, 33, 55, 67, and 79, and again with *SparseAssembler* (Ye et al., 2011) using a maximum kmer size of 75. Contigs were assembled using *runBNG* (Yuan et al., 2017), a commandline wrapper for the proprietary BioNano assembler, in *"denovo"* mode.

Assembly extension & polishing

Arrow was used to realign PacBio SMRT reads to the PacBio assembly, to build off the initial high-quality assembly with reads that may have been excluded by *canu*. *DBG2OLC* was used to produce a hybrid assembly, building PacBio long reads onto the *SparseAssembler* Illumina assembly. The resulting *DBG2OLC* assembly and previous *SPAdes* assembly were merged (#1) reciprocally (two merges, one with each as a reference and other as query) with *QuickMerge*, and results concatenated. These merged assemblies were then merged (#2) to the resulting Arrow assembly, which was used as the basis for *runBNG*'s BioNano hybrid assembly. The hybrid BioNano/Arrow assembly, consisting of a small set of long contigs, was recombined to the overall sequence assembly in a final merge step (#3).

Pilon (Walker et al., 2014) was used for polishing of the final merged assembly. Illumina reads were aligned to contigs to identify possible errors. As the per-base quality of single Illumina reads is much higher than that of PacBio sequence, this is highly desirable both for correcting assembly issues (primarily single bases, but also sequence repeats or inversions) and identifying heterozygous loci for later use. Sequence assembly statistics were generated using "FastaStats", a self-contained custom Perl script which allows quick assessment of FASTA sequence files.

Genome Annotation

Input preparation

The number of parallel CPUs and size of reference databases are the primary factors that influence the time required for *de novo* genome annotation. In order to reduce number of extraneous downstream comparisons, custom reference sets were used wherever possible (Figure 22). For general plant proteins, a reference set was created of all angiosperm sequences on RefSeq (Pruitt, Tatusova & Maglott, 2007) and UniProt (Bairoch et al., 2005). Then, a set of all potential *Sarracenia alata* protein sequences was produced using ORFfinder (Sayers et al., 2009), with a minimum length of 150 amino acids, excluding wholly-overlapping reading frames. This set of 670,607 translated ORFs (extremely liberal, expected to be >90% false-positives) was used as a reference to filter the compiled plant protein database 6.6 million sequences. Protein blast (E-value = 10e-7) produced 1,697,548 matches with identity $\geq$65%, a database reduction of 74.2%. Sequences with headers matching the keyword "[T/t]ranspos*" (8,159) were also filtered. For reference mRNAs, the assembled transcriptomes of congenerics *S. purpurea* and *S. psittacina* (Srivastava et al., 2011) were obtained and combined, totaling 55,506

transcripts. These were then compared to the *S. alata* genome via blastn (E-value = 10e-8), producing 50,581 matches with identity ≥85%, a database reduction of 8.9%.

A reference transcriptase library provided with MAKER (Cantarel et al., 2008) was used to determine transposons on a protein basis; this consisted of 24,916 protein sequences and was not reduced. A nucleotide-based custom TE library was also constructed, following the guidelines of "MAKER Custom Repeat Library, Basic and Advanced" (Campbell et al., 2014). This process consisted of repeat extraction from the *S. alata* genome using RepeatModeler with subsequent filtering of likely misidentified genes via ProtExcluder.

MAKER-P annotation

Annotation was performed using *MAKER-P* (Campbell et al., 2014), an improvement on the *MAKER* (Cantarel et al., 2008) package incorporating improved parallelization and specialized scripts for overcoming the complexities of plant genomes. The pipeline was analyzed using the aforementioned reference libraries. Within the pipeline, gene identifications and functional inferences were performed via Augustus using a *Zea mays*-based HMM, and Exonerate. Repeats were identified using RepeatMasker (Tarailo-Graovac & Chen, 2009) with an *Arabidopsis thaliana* model. Non-coding regulatory tRNAs were identified using *tRNAscan-SE* (Lowe & Eddy, 1996). Functional predictions were made based on ESTs and protein homology; single-exon ESTs were excluded. Expected max intron size was set at 10 kbp, with a max non-split contig size of 100 kbp and a minimum contig size of 20 kbp (73.9% of sequence). *MAKER-P* was implemented on the Ohio Supercomputer Center's OWENS cluster (The Ohio Supercomputer Center, 1987), using 2,268 CPUs (28 cores x 81 nodes).

Evaluation of Carnivorous Function

A modification of the approach described in Chapter 2 was utilized to evaluate the role of *Sarracenia alata* genes in producing digestive enzymes. *S. alata* genes, as identified via the MAKER-P pipeline, were assigned function via *blastx* search against the NCBI non-redundant protein database, "*nr"*, followed by Gene Ontology assignment via *Blast2GO*'s "Mapping" function (Conesa et al., 2005; Conesa & Götz, 2008). Codes were then processed and analyzed via the scripts and pipeline presented in Figure 3. The functions of presumed adaptive importance tested are as follows: *actin filament, alternative oxidase activity, aspartic-type peptidase activity, ATPase activity, ATP:ADP antiporter activity, beta-galactosidase activity, chitinase activity, cinnamyl-alcohol dehydrogenase activity, cysteine-type peptidase activity, fructose bisphosphate aldolase activity, glutathione transferase activity, water channel activity, heat shock protein activity, lipase activity, lipid transferase activity, nitrogen transport activity, peroxidase activity, phospholipase activity, polygalactosidase activity, protein homodimerization activity, ribonuclease activity, serine carboxypeptidase activity,* and *thioglucosidase activity.* Functions were scored by genomic representation as a proportion of all genes assigned GO-based identifiers. Individual taxa (*Sarracenia alata,* as well as four other carnivorous plants: *Cephalotus follicularis, Drosera capensis, Genlisea aurea,* and *Utricularia* gibba) were tested against a reference distribution of non-carnivorous plants via Z-test; an overall test of carnivorous taxa as a whole vs. reference plant taxa was performed via t-test. Statistical error due to multiple testing was accounted for using Storey's q-value correction (Storey, 2003; Storey & Tibshirani, 2003; Dabney, Storey & Warnes, 2010).

Results

Assembly Statistics

The results of the pipeline shown in Figure 1 – initial assembly via parallel *de novo* methods, improvements, and final assembly – are summarized below. Full assembly metrics for each stage are available in Table 15.

Initial assembly

Four *de novo* assemblies were performed, two using Illumina sequence data (SPAdes and SparseAssembler), one using PacBio long-read data (Canu), and one using BioNano sequence map data (runBNG). In Illumina assembly, SPAdes appeared superior to SparseAssembler, which produced only a 252 megabase assembly (6.75% coverage) versus SPAdes's 1.54 gigabase assembly (42.9%). SPAdes assembly contiguity was much higher, with an average length of 1015.9 versus 257.9 (3.9x length) and an N50 of 2132.0 versus 245.0 (8.7x length). SPAdes assembly also captured nine contigs greater than 100 kbp in length and one greater than 250 kbp, while SparseAssembler produced a maximum contig length of only 46.1 kbp.

Canu *de novo* assembly of PacBio sequence data was expected to serve as the core of the assembly pipeline, and it successfully produced an assembly exceeding the ability of the short-read assemblers tested. Contigs totaled 3.16 Gbp (87.8% coverage) with an N50 of 35.6 kbp. Over 2,600 contigs exceeded 100 kbp in length, with 3.1 Mbp the maximum length assembled.

BioNano optical map data assembled via runBNG's "denovo" function produced an assembly covering an equivalent of 2.15 Gbp of sequence (59.7%); however, contiguity was below the expectations of BioNano assembly with the genome split into 7,387 maps, equivalent to nearly five hundred for each of *S. alata*'s 13 chromosomes and two plastids. An N50 of only

295 kbp limits the utility of this data in bridging sequencing gaps and long tracts of repetitive elements.

Assembly improvement

DBG2OLC hybrid assembly substantially improved the SparseAssembler result, with 21.0% genome coverage (a 3.1x increase) and an N50 of 10,411 (42.5x increase). This assembly possessed longer average contigs than the SPAdes *de novo* assembly; however, it possessed only half the total coverage and assembled no contigs above 100 kbp.

Arrow alignment extension increased the total assembly length by only 0.1%. Mean length increased by 11.8 bp, while N50 increased by 26.5 bp. As some sequences grew in length, the count of contigs >100kbp increased by seven. The merging of the SPAdes and DBG2OLC assemblies with this assembly via QuickMerge produced only 14 sequence changes, resulting in little improvement but suggesting that the PacBio assembly was relatively accurate.

Hybrid assembly using BioNano contigs as scaffolds produced 760 Mbp of sequence in 306 contigs. This accounts for only 2.1% of the expected genome size, but produces only long contigs, with an N50 of 223.1 kbp and >99% of contigs larger than 100 kbp in length. However, subsequent reintegration of this assembly with the Canu/Arrow assembly had no effect, as the same long contigs were present within the PacBio-only assembly.

Pilon polishing with Illumina sequences resulted in a total of 19.6 million changes (Table 6), averaging one per 162.6 bp of assembly. Of these changes, 12.1 million were changes of a single base (61.7%), while 1.9 million were the addition or removal of a base (9.7%), and 0.48 million (2.4%) were multi-base sequence changes, including MSATs, repeats, and inversions.

The remaining changes were due to the identification of variable sites coded as IUPAC two-base degeneracy codes, where the Illumina reference sample was heterozygous – 5.3 million in all.

	Number	Percentage
Total Changes	19.6 million	-
BP/Change	162.6 bp	-
Base Change	12.1 million	61.7%
Base Addition	1.3 million	6.6%
Base Removal	621 thousand	3.2%
Sequence Change	2.7 thousand	0.0%
Sequence Addition	264 thousand	1.3%
Sequence Removal	215 thousand	1.1%
Variable Sites	5.3 million	27.0%
BP/Variable Site	5.9 thousand	-

Table 6: Alignment sequence changes due to Pilon final polish pass. Of 19.6 million changes, the vast majority were single-base errors, with less than 2.5% consisting of multi-base error corrections. 5.3 million changes were due to heterozygous sites in the reference sample, equating to one SNP per 5.9 thousand bases of sequence.

Final statistics

The final draft assembly of the genome of *Sarracenia alata* totals 3.16 billion base pairs, accounting for 87.8% of the expected genome size based on prior estimates (i.e., Rogers et al., 2010). Final coverage and contiguity is comparable to other carnivorous plant genomes, such as *Cephalotus follicularis* (Australian pitcher) and *Dionaea muscipula* (Venus's flytrap), and also similar to *S. alata*'s closest relative with a genome sequence, *Actinidia chinensis* (kiwifruit). A detailed comparison of these four genomes by a variety of metrics is presented in Table 7.

Coverage	*S. alata*	*A. chinensis*	*D. muscipula*	*C. follicularis*
Total Sequence	3.16 Gb	0.604 Gb	1.44 Gb	1.61 Gb
Est. Genome Size	3.60 Gb	2.60 Gb	2.79 Gb	2.12 Gb
Genome Coverage	87.8%	23.2%	51.6%	76.0%
Contiguity				
Contigs	122,036	26,721	86,832	16,307
Mean Length	25,910.1	22,611.9	-	99,007.6
N50	35,649.5	58,852.0	34,655.0	287,484.5
Length Quartiles				
Maximum	3,099,149	423,496	1,117,843	2,219,130
Q3	30,990	28,425	-	124,000
Median	17,619	7,933	-	30,357
Q1	1,150	2,597	-	3,136
Minimum	1,000*	200*	1,000*	968
# Large Assemblies				
>100kb	2,640	1,106	-	4,767
>250kb	73	29	-	1,992
>500kb	7	0	-	577
>1Mb	2	0	1	58

Table 7: Assembly statistics of Sarracenia alata draft genome & comparison taxa. *Metrics given show total sequence assembled (bp), percent of genome covered. assembly contiguity, length distribution (bp), and number of large assembled contigs. Minimum sequence lengths marked with asterisks indicate that all sequences smaller than this size were filtered from the assembly. Actinidia chinensis statistics generated from Huang et al., 2013 assembly via "FastaStats"; Cephalotus follicularis statistics generated from Fukushima, Fang, et al., 2017 assembly via "FastaStats"; Dionaea muscipula assembly statistics from Hackl, 2015, where available.*

Genome Annotation

Genes

From annotation results based on MAKER-P gene predictions, *Sarracenia alata* was found to possess 28,750 genes predicted to be protein-coding. Of these, mRNA transcripts were directly predicted for 27,792, encompassing 66,068 exons. Genic regions accounted for 73.6 million base pairs, 2.01% of the assembled genome sequence; exons accounted for 21.1 million base pairs, or 0.66% of the assembled genome. Of the 27,792 genes for which protein-coding

mRNAs were predicted, protein BLAST results were generated for 25,133 (90.4%); 17,956 (64.6%) could be assigned annotations containing Gene Ontology identifiers. The most commonly-detected GOs are shown in Figure 2; counts of all GOs detected are given in Table 16.

Non-coding RNA & pseudogenes

Elements coding for functional but non-translated RNAs were also identified. Sequence coding for rRNA, the structural component of ribosomes, was found at 609 sites, comprising 297.5 kilobase (>0.01%). tRNA, which can serve a regulatory or directly functional role (Park & Kim, 2018), was identified at 961 sites, comprising 72.9 kilobase (>0.01%).

Pseudogenes, meeting certain predictive criteria for genes but scoring very low by functional coding metrics, were substantial in number but comprised a very small portion of genome sequence. MAKER-P predicted 36,979 probable pseudogenes, comprising 53.0 million base pairs (1.47%). The ratio of pseudogenes to genes was found to be 1.29:1.0, while the ratio of DNA sequence associated with pseudogenes versus genes was 0.72:1.0.

Repetitive elements

Repetitive elements comprised the vast majority of *Sarracenia alata*'s genome – 2.9 gigabase, accounting for 91.7% of assembly sequence (Figure 2). Long Terminal Repeats (LTRs) were found to be the dominant class by far, comprising 1.3 gigabase (40.3%), followed by DNA transposons (77.8 megabase, 2.46%) and Long Interspersed Nuclear Elements (LINEs) (34.4 megabase, 1.09%). A thorough breakdown of the repetitive elements in the *Sarracenia*

genome to finer levels of classification can be found in Table 17.

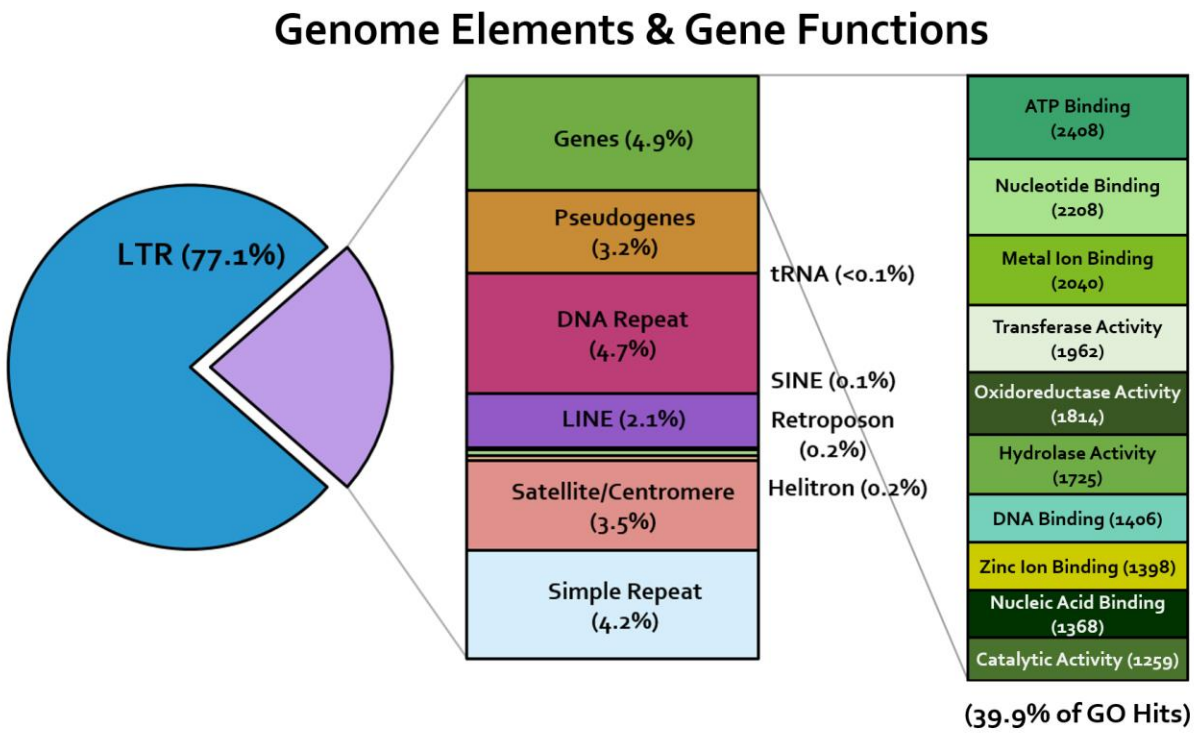

Figure 7: Genome composition of Sarracenia alata, as determined by MAKER-P annotation. *Pie chart (left) shows the portion of the genome constituted by Long Terminal Repeats (LTR), which account for the vast majority of sequence (blue). The remaining sequence (purple) is broken down as a stacked bar (middle). "Genes" comprise the portion of the genome predicted to be functional exonic sequence. "Pseudogenes" include gene-like sequence not likely to be transcribed. All remaining sequence is attributed to eight general classifications of repetitive or degenerate nucleotide sequence. Genes are further differentiated by function (bar, right) based on Gene Ontology. Molecular functions with over one thousand sequence hits are shown; the represented functions comprise 39.9% of all gene annotations, with the remaining 60.1% divided over 1,956 additional terms. A complete list of all Gene Ontology codes and their representation can be found in Table 16; a list of all repetitive and transposable elements including finer levels of classification is presented in Table 17.*

Evaluation of Carnivorous Function

Of the genes identified and annotated by MAKER-P and successfully assigned GO codes by Blast2GO, 5.4% matched to the reference list, identifying them as encoding carnivory-associated functions. Reference non-carnivorous plant taxa had previously been found to devote an average of 3.5% of genes to functions associated with plant carnivory while other carnivorous

taxa were found to devote 3.9% on average, a statistically non-significant differentiation (Chapter 2). Unlike other carnivorous taxa, the total level of representation in *Sarracenia* is significantly above the reference range (Z = 2.24; a = 0.05; p = 0.013; q = 0.30). This pattern is illustrated in Figure 8, which presents both the reference and overall carnivorous ranges as well as values for all carnivorous taxa previously tested.

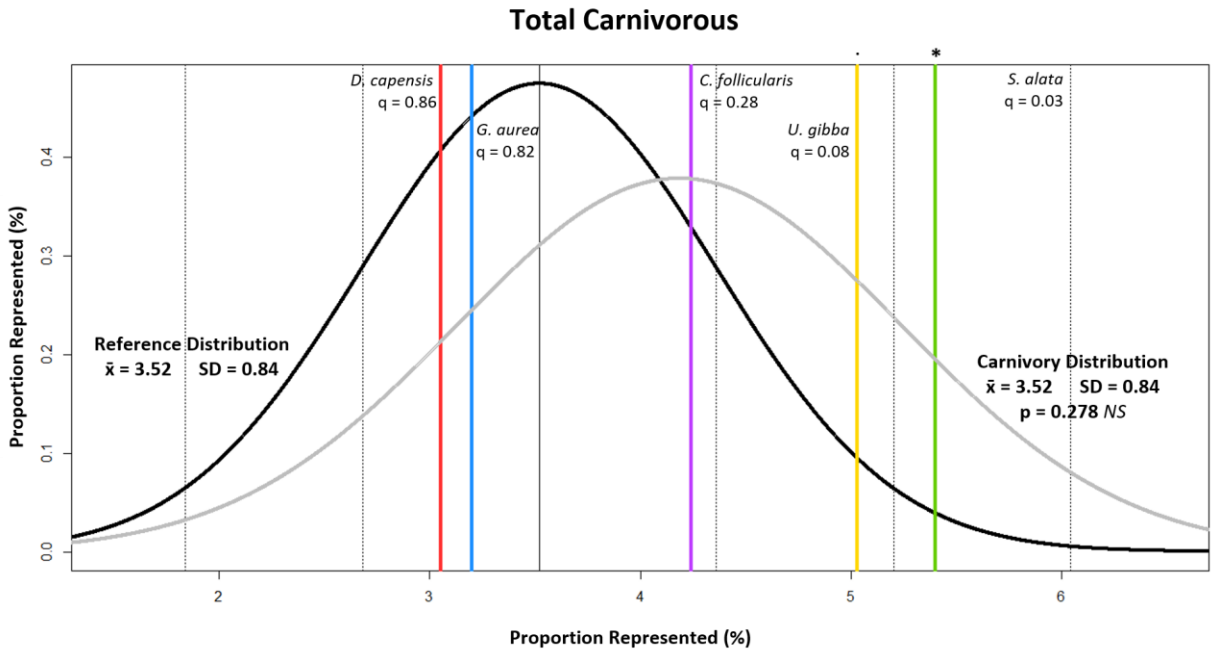

Figure 8: Total representation of carnivory-associated functions in carnivorous plant taxa vs. reference taxa. *Reference distribution (black curve) is constructed from data gathered from representative non-carnivorous plant taxa, while carnivory distribution (grey curve) uses the combined values of all carnivorous taxa shown. Colored lines indicate carnivorous taxa, while solid black line indicates reference mean and dotted grey lines indicate standard deviations from reference. Tests of individual carnivorous taxa are made against the reference distribution; q-values are presented when account for multiple testing using Storey's q.*

This pattern holds true for individual functions as well. For 24 tests of carnivory gene representation, carnivorous taxa averaged 3.5 functions (14.6%) falling significantly above the reference range; *Sarracenia alata* possessed six functions falling significantly above the reference range (25.0%). Of these functions, four (*alternative oxidase activity, ATPase activity, phosphatase activity,* and

phospholipase activity) had been identified as significant previously in at least one other taxon, whereas two (*ATP:ADP antiporter activity* and *thioglucosidase activity*) have been ascribed statistical significance for the first time in *S. alata.* Statistical information about the tests performed is presented in Table 3.

Function	Z	p	q	Sig.
Actin	0.00	0.500	0.490	*NS*
AltOx	**2.69**	**0.004**	**0.012**	*
AspPep	0.20	0.419	0.466	*NS*
ATP	**7.02**	**1.1E-12**	**9.5E-12**	****
ATP_ADP	**2.29**	**0.011**	**0.030**	*
BGal	1.67	0.047	0.098	.
Chit	-0.46	0.677	0.553	*NS*
CinAlc	-0.61	0.729	0.553	*NS*
CystPep	-0.61	0.728	0.553	*NS*
FrucBPA	0.89	0.185	0.281	*NS*
GlutTrans	1.49	0.068	0.126	*NS*
H2OChan	0.30	0.380	0.453	*NS*
HeatShock	1.41	0.079	0.132	*NS*
Lipase	0.33	0.370	0.453	*NS*
LipTrans	-0.24	0.596	0.523	*NS*
NHTrans	0.00	0.499	0.490	*NS*
Perox	-0.09	0.536	0.497	*NS*
Phoslip	**3.50**	**2.3E-04**	**1.3E-03**	**
Phosp	**2.89**	**0.002**	**0.008**	**
Polygal	-0.93	0.823	0.572	*NS*
ProtHomo	0.65	0.258	0.358	*NS*
RiboNuc	-0.91	0.819	0.572	*NS*
SerCarPep	-1.11	0.867	0.578	*NS*
ThioGluc	**7.94**	**9.9E-16**	**1.7E-14**	****
Total	**2.24**	**0.013**	**0.030**	*

Table 8: Overrepresentation of carnivory-associated Gene Ontology codes in Sarracenia alata genome. Values generated in R using the methods of Chapter 2: a Z-test was performed against a reference range of non-carnivorous plant taxa, with p-values adjusted for multiple testing using Storey's q-value. "Total" indicates the overall proportion of the genome devoted to any carnivory-associated functions. Significance codes are as follows: $q > 0.10$ (NS); $q < 0.10$ (.); $q < 0.05$ (*); $q < 0.01$ (**); $q < 0.001$ (***); $q < 0.0001$ (****). Functions found to have statistically significant overrepresentation of at least the $q < 0.05$ (*) level are bolded. Values less than 0.0015 are presented in scientific notation.

Discussion

Genome Assembly

The methods used in this study were able to produce a genome assembly comparable or superior to those performed in similar organisms; however, nearly all success is attributable to the Canu assembler. Additional parallel assembly steps with subsequent merging had negligible effect on the coverage or contiguity of the *S. alata* genome. BioNano optical mapping data was also unable to join any genome fragments and produced no detectable end effect. This is likely due to BioNano assembly's dependence on long anchor-reads with thorough marker-site coverage; though the method may be able to greatly improve the contiguity of chromosomes split into dozens of fragments, it is unable to assemble chromosomes from thousands. Additional BioNano sequencing coverage could potentially improve on these issues, but at the assembly's current level of completeness and contiguity, additional PacBio SMRT sequencing would be prioritized.

While merging with Illumina-based *de novo* assemblies did not improve the primary assembly of long-read data, the short-read sequence library still proved to be a useful resource in the final polishing of the genome. Pilon's corrections, one per 162 bp, were the most substantial improvement to the Canu assembly. As this step requires only a modest amount of Illumina sequencing to produce a large number of corrections, it is likely to be a worthwhile investment in any PacBio-based sequencing effort. Furthermore, Pilon identified 5.3 million variable sites; however, these are based on the sequencing of only a single individual, representing individual heterozygosity rather than species diversity. Collecting a smaller amount of sequence for a larger range of individuals could quickly and easily determine tens of millions of informative sites for the species.

Genome annotation

The MAKER (Cantarel et al., 2008) methodology is a time-tested approach to automated genome annotation, improved by parallelization and support scripting in the MAKER-P package. However, while MAKER-P is specifically designed for plants and boasts reasonable assembly times for even large plant genomes (Campbell et al., 2014), the large-and-growing size of general plant reference sequence databases keeps this methodology out of practical reach of most users. While the standard MAKER-P procedure did not approach completion after >200,000 CPU-hours, our added database pre-processing step greatly increased its speed, allowing the completion of a satisfactory genome annotation in 145,000 CPU-hours (64 realtime hours).

The results of this annotation are within the expected range for genes and repetitive elements; however, the number of genes identified is towards the smaller end of the range for flowering plants. As no close relatives have been sequenced to a level that would allow for annotation, it is impossible to determine if this is representative of reality, or if annotation efforts were less successful than desired. In the latter case, it would still be difficult to determine if this is an effect of the pre-filtering approach used or simply a difficulty of annotating this particular plant genome, in the absence of specific resources for the purpose.

Plant carnivory

Repeating the analyses previously used to assess the genomic representation of functions associated with plant carnivory, *S. alata* was found to possess a genome significantly adapted to suit a carnivorous lifestyle. In terms of overall proportion of genes devoted to these functions, *S. alata* exceeded all other taxa tested. Furthermore, it was found to possess a signal of increased abundance of more individual carnivory-associated functions than any other taxa tested.

Considering the continued argument that *Sarracenia* is largely or wholly reliant on microbiota and its unusual lack of digestive glands, this is a surprising result.

There are several potential explanations for this phenomenon. First, it is possible that this result is in error, due to a quirk of the annotation pipeline; however, no aspect of the MAKER-P pipeline itself nor of the filtration steps applied should create bias favoring these specific classes. In fact, as carnivorous taxa are poorly represented in the plant protein database, a bias towards underrepresentation seems more likely. Second, *S. alata* may require such substantial overrepresentation of genes encoding these functions *because* of its lack of glandular elements. If *Sarracenia* has evolved a less efficient means of enzyme secretion than other pitcher genera, it may account for this via some form of overproduction. One past (Plummer & Kethley, 1964) had suggested that *Sarracenia* species perform digestion intracellularly, uptaking intact peptides and completing digestion in the cytoplasm. Finally, *S. alata* may have transitioned to a "post-carnivorous" lifestyle, previously possessing high enzymatic secretory activity but later deactivating expression due to increased reliance on symbioses. In this case, depending on the evolutionary time passed, genes could still be present and detectable within the genome, able to code for functions whose expression is suppressed. This has been seen in other pitcher plant taxa; *Nepenthes lowii, N. macrophylla,* and *N. rajah* have formed symbioses with tree shrews (Clarke, Moran & Chin, 2010), while *N. rafflesiana (*forma *elongata)* has formed a similar interaction with bats (Grafe et al., 2011a). These taxa have adapted to gather nutrients from mammalian feces deposited into their pitchers, gaining the majority of their nitrogen from these sources while reducing the presence of characteristics used in insect capture and digestion.

Conclusions

While the use of parallel assembly and merging steps did not noticeably improve assembly, the use of PacBio sequence with *Canu* and Illumina sequence with Pilon did produce an assembly comparable to many other plant genome assemblies and sufficient for the purposes of annotation and gene prediction. Improvement of this assembly may be possible at a future time but would almost certainly require the addition of additional sequence data. In particular, BioNano data is expected to be increasingly useful with increasing PacBio assembly N50 or the addition of other longer-read sequence data such as Dovetail (Moll et al., 2017).

Pre-filtering of reference data sets before MAKER-P annotation greatly increased the speed at which the genome of *S. alata* could be completed. Based on current results, we recommend this method for future projects where large genomes may be limited by the availability of compute resources or time; however, further verification is needed to ascertain the extent of error or loss in annotations, which could be tested using other taxa with pre-existing high-confidence genome annotation.

The surprising findings in the *S. alata* genome, indicating functional adaptation to plant carnivory even beyond the extent of other carnivorous taxa, warrants further investigation. For example, transcriptomic analyses could leverage the new genomic resources presented here and be used to disentangle the role of the host plant from its symbionts. With the ability to determine both the sequence origin and level of expression of carnivory functions, it will be possible to determine whether *S. alata* has in fact transitioned to a symbiont-based "post-carnivorous" strategy, or if it is still active in digestion by unknown means.

The annotated genome of *S. alata* is now available on NCBI GenBank. In addition, raw sequence reads of Illumina, PacBio, and BioNano data have been made available via the NCBI

Sequence Read Archive, for use in future studies should further assembly be attempted. Finally, custom scripts used in this study, including those for the generation of assembly statistics ("FastaStats.pl"), summarization of element annotations ("SummarizeAnnotation.pl") and the pre-filtering of MAKER-P reference libraries ("RedundancyFilter.pl", "KeywordFilter.pl", "BlastFilter.pl"), are available on GitHub (https://github.com/GWheelerEB/SarraceniaGenome).

Chapter 4: Unraveling the mystery of *Sarracenia alata*'s plant carnivory using meta-transcriptomics

Abstract

In addition to phenotypic adaptations for prey attraction and capture, plant carnivory requires enzymes for the digestion of prey into assimilable nutrients. In *Sarracenia*, there has been considerable debate as to whether the requisite enzymes are supplied by the plant and derived from homologous enzymes in other Angiosperms or products of the large and diverse microbiome that inhabit the plants pitcher fluid. In order to address this question, metatranscriptomic data were collected from *Sarracenia alata* pitcher fluid and assembled. The *S. alata* genome sequence and transcriptomes of *S. purpurea* and *S. psittacina* were used as references to determine the source of digestive enzymes; sequences not mapping to *Sarracenia* were separated out and assigned taxonomy via BLAST Transcripts were then assigned to functions via BLAST and subsequent mapping.

Results indicate that the digestive enzymes are produced by the microbiome rather than the plant genome. While 10.3% of microbial genes are associated with carnivory functions, this was true of only 1.0% of plant genes; furthermore, none of these genes mapped to known enzymatic activity. Of host plant genes that mapped to a carnivory-associated function, nearly all transcripts were associated with a single gene, encoding a sodium and nitrogen transporter. In addition to this transport function, we uncovered evidence that *S. alata* is actively regulating the composition of its microbiome. Over 40% of all transcripts produced were mapped as probable

anti-microbial peptides (AMPs), short proteins known to impact microbial assemblages in the guts of many animals.

Keywords

Sarracenia alata, metatranscriptome, plant carnivory, gene function, AMPs

Introduction

Givnish (2015) defines a carnivorous plant as having the ability to "absorb nutrients from dead bodies adjacent to its surfaces, obtain some advantage in growth or reproduction, and have unequivocal adaptations for active prey attraction, capture, and digestion". *Sarracenia* and *Roridula* (Ericales) are explicitly included in this categorization because they have numerous adaptations for the luring and trapping of prey, but it is unclear the extent to which they rely on mutualistic interactions with symbiotic microorganisms. Previous studies in *Sarracenia* have determined that bacteria (Plummer & Jackson, 1963; Anderson & Midgley, 2003; Luciano & Newell, 2017b), and potentially other organisms (Canter et al., 2018), are involved in prey digestion. Studies of the process of signal transduction on triggering digestive enzyme production have also been conducted, revealing that pitchers are enzymatically active shortly after opening and when prey are present, but cease enzyme production after prolonged prey absence (Gallie & Chang, 1997). These studies lack a quantitative evaluation of the contribution of the plant and microbiome to digestive function, in part due to the many-to-one relationship between nucleotide sequence and protein, which have complicated proteomic work (Gotelli, Ellison & Ballif, 2012), and in part due to the lack of a reference genome for *Sarracenia*.

A reference genome for *Sarracenia alata* was assembled, annotated, and analyzed in Chapter 2. The fraction of *S. alata*'s genes devoted to carnivory-associated functions was tested against a reference range of non-carnivorous plants; of 24 functions tested, it was found that six were overrepresented at statistically significant levels, the highest number identified in a carnivorous plant to date. An overall signal of carnivory, taken as the total fraction of genes associated with a carnivory-associated function was also detected, a finding not observed in other carnivorous taxa. These results are highly suggestive of *S. alata* taking an active role in the digestion of prey, a finding which is at odds with the genus's lack of digestive glands (Adams & Smith, 1977; Płachno et al., 2006) and studies in related species suggesting a reliance on microbiota.

This study uses high-throughput sequencing to characterize and evaluate the meta-transcriptome contained within *Sarracenia alata*'s digestive fluid. This presents two notable advantages over the use of DNA: a transcriptome can be used as both sequence and expression level data, allowing the inference of the genotype as well as a proxy for phenotype; and. RNA has a shorter environmental half-life than DNA, increasing the likelihood of sequence that is being actively used and thus actively produced, reducing the prevalence of incidental sequence. This approach has been successfully used for diversity assessment of marine microbes (Gifford et al., 2011) and high-altitude slime molds (Kamono et al., 2013), and has also been used for functional assessment, for example in the human gut (Franzosa et al., 2014). RNA was isolated direction from the fluid in order to quantify the enzymes used in digestion and to increase the likelihood of identifying *S. alata* compounds that originate from the cell walls that line the

pitcher, as such compounds are assumed to be more likely to contribute to plant carnivory functions.

Materials & Methods

Sample Collection

Samples of *Sarracenia alata* pitcher fluid were collected from wild plants at three field sites in July, 2016, a point in the growing season where microbial diversity is expected to be high (Koopman et al., 2010). Fluid was immediately mixed with RNALater RNA storage solution in a 4:1 solution:sample ratio for shipping to The Ohio State University. Samples were quantified via Qubit to determine the presence of detectable nucleotides. For fluid samples from two of the three populations were found to possess detectable levels of RNA. Samples of these populations were combined to obtain sufficient nucleotide mass for library preparation.

Library preparation and sequencing

Combined pitcher fluid samples were processed and prepared by RTL Genomics. Samples were filtered, but due to the low concentrations overall, rRNA was not removed. Sequencing on Illumina HiSeq platform (150 bp, paired-end) produced 167.2 million reads. Per-sequence quality phred scores averaged 33.4 (99.95%). Sequence reads were filtered for contamination by known RNA-Seq adapters. Reads were also filtered for lengths of >36 guanine residues, as poor reverse-sequence quality in some cells resulted in “dark reads” misinterpreted as long poly-G regions. After trimming and filtering, 21.4 Gbp of sequence remained (average clipped read length: 127.9 bp), with 61.3% of reads pairable.

Transcriptome assembly

Illumina reads were assembled using Trinity (Grabherr et al., 2011). For the purposes of downstream analyses, Trinity contig identifiers were used to identify distinct genes (coded as g#) and isoforms (i#).

Meta-transcriptome characterization

Assembled contigs were assigned function and taxonomy via BLAST. To determine the probable functions of proteins encoded in transcripts, blastx was performed against the non-redundant protein database "nr" (NCBI) with an E-value of 1E-9, using an identity cutoff of 0.95. Hits were then mapped in Blast2GO (Conesa et al., 2005; Conesa & Götz, 2008) to add functional descriptors and assign Gene Ontology (GO) identifiers to each gene. The longest isoform of a gene was used to determine its function, with smaller isoforms assigned the same descriptors and GO identifiers.

As specific microbes are expected to have differing levels of utility to the plant, another possibility considered was the involvement of anti-microbial peptides (AMPs) in regulating the population levels of different bacterial species and groups. These very short (12-50 amino acids) proteins have high-specificity anti-microbial activity and are known to play a role in the guts of other organisms (Ostaff, Stange & Wehkamp, 2013; Cullen et al., 2015) but are not identified in more general BLAST database searches. The complete APD3 (Anti-microbial Peptide Database ver. 3) (Wang, Li & Wang, 2016) was downloaded, and all sequences for which GO codes could not be assigned were searched against it via blastx. An identity score of >75% was used to mark genes as potentially coding for AMPs.

To identify transcripts originating from the *Sarracenia alata* pitcher itself among those originating from its inquiline microbiota, the draft genome of *S. alata* (Chapter 3) and reference transcriptomes of *Sarracenia purpurea* and *Sarracenia psittacina* (Srivastava et al., 2011) were used as references. A BLAST database was prepared from each, and the *Sarracenia alata* fluid metatranscriptome was compared to each via blastn. To identify a sequence as originating from *S. alata*, either of two conditions had to be met: a sequence was identified as present in at least two of the three references with ≥95% identity in each OR it was identified in at least one reference with ≥99% identity. For genes possessing multiple isoforms, a match for any isoform was considered a match at the gene level. It should be noted that in a previous study, approximately 40% of contigs were shared between *S. purpurea* and *S. psittacina* (Srivastava et al., 2011), making this a conservative approach.

Identification of other taxa was performed via blastn of all sequences not assigned to *Sarracenia alata*, using the "nt" nucleotide database (NCBI) as a reference. The blastn search was performed using an E-value of 1E-9 and an identity score cutoff of 0.90. The software package BLASTGrabber (Neumann et al., 2014) was used to convert BLAST identifiers into hierarchical taxonomy.

To determine expression levels for each gene and isoform, filtered Illumina read libraries were aligned to the Trinity assembly contigs using *bowtie2* (Langmead & Salzberg, 2012) and indexed with *samtools* (Li et al., 2009)). *eXpress* (Roberts & Pachter, 2013) was used to generate transcript counts from the indexed data, in relative units of transcripts/million.

Genes were marked as carnivory-associated on the basis of their Gene Ontology mappings. A list of 36 functions previously identified as carnivory-related (Chapter 2) was used as a starting point, with this list expanded to 288 functions by including codes one hierarchical

level up (parents) or down (children). Genes matching to the original 36 functions were identified as “narrow-sense”, while those only matching to the expanded criteria were marked as “broad-sense”.

Results

Assembly statistics

Trinity assembly produced 36,699 contigs, comprising 10.6 megabase of nucleotide sequence. Contigs ranged in length from 201 to 9136 bp, with an average length of 290.1 bp and an N50 of 268 bp. Each contig constitutes a unique isoform, corresponding to 30,482 inferred genes with an average 1.2 isoforms per gene. Subsequent alignment with Bowtie2 was able to match 78.0% of total sequence fragments to the resulting assembly.

Relative contribution

BLAST analysis revealed that 1249 (3.4%) of total assembled contigs produced high-certainty matches with the genome of *Sarracenia alata*. As the average *Sarracenia*-contributed gene possessed 4.1 isoforms, this translates to 302 total unique genes. For the remaining 35,420 contigs, an average of 1.2 isoforms were found per gene; as prokaryotic organisms do not perform differential splicing (Roy & Gilbert, 2006), 0.2 isoforms/gene can be attributed to the contributions of either eukaryotic organisms present in the pitcher fluid, or to error.

Sarracenia alata genes, despite their low number, were found to produce 67.7% of transcripts (677,240/million). The average *S. alata* gene contributed 2,242 transcripts/million, while bacterial genes produced only 10.7 transcripts/million.

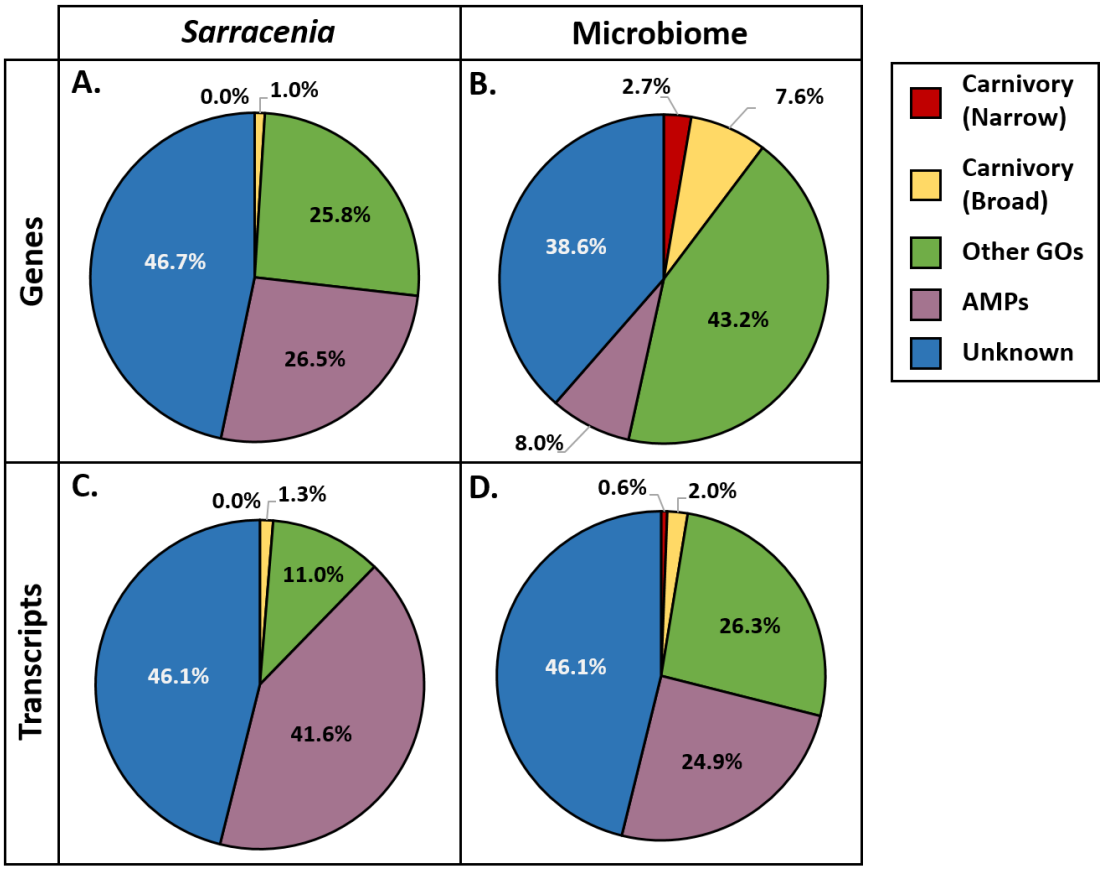

Figure 9: Representation of Sarracenia alata genes (A) and transcripts (C) versus microbial genes (B) and transcripts (D). *Narrow-sense carnivory shows the portion that maps to a carnivory-associated function as listed in Chapter 2, while broad-sense carnivory shows additional hits using an expanded list (+/- one GO hierarchical level). "Other GOs" indicate sequences for which Gene Ontology-coded function(s) could be confidently assigned but which did not match any carnivory-associated functions. "AMPs" indicates the portion of sequences that mapped to reference samples from the anti-microbial peptide database "Unknown" contains all sequences fitting into none of the aforementioned categories.*

When assigning functions to genes and transcripts (Figure 9), no narrow-sense carnivory genes were found in *Sarracenia alata*, and only 1% of genes were devoted to carnivory functions in the broad sense. Roughly a quarter were assigned to other specific functions with no known carnivory-related role, while an additional 26.5% mapped to anti-microbial peptides (AMPs). The remaining 46.7% remained unidentifiable. In the microbiota, a much larger portion of carnivory-associated functional genes were identified, with 2.7% matching to narrow-sense

criteria and an additional 7.6% to broad-sense. Other functions composed 43.2% of genes, AMPs composed only 8.0%, and the remaining 38.6% remained unknown. At the level of transcription, representation of carnivory-associated functions was consistent with the gene level. AMPs were over-transcribed, making up 41.6% of transcripts. In the microbiota, a quarter of all transcripts were associated with AMPs; only 2.6% of transcripts mapped to a carnivory function, making these genes under-transcribed relative to their overall representation.

Carnivory-associated functions

For the set of functions associated with plant carnivory (Figure 10, broad-sense functions collapsed into their narrow-sense term), the most transcribed functions by far were *ammonium transmembrane transport* and *sodium ion transmembrane transport*. These functions mapped primarily to *Sarracenia alata* transcripts, while *S. alata* did not possess genes for any other functions. *ATPase activity* was the most-transcribed catalytic enzyme, with 13 total found to be transcribed at levels above 90 transcripts/million. The remaining functions represented at this level were two stress-response functions (*heat-shock protein* and *protein homodimerization activity*) and one additional transport function (*lipid transferase activity*). The remaining 17 functions together accounted for 229.7 transcripts/million.

Taxon assemblage

The micro-ecosystem inside *Sarracenia alata*'s pitchers was found to be dominated by bacterial strains. Based on contig representation, proteobacteria were most abundant, with gammaproteobacteria (containing enterobacteria) dominating within this group. Firmicute bacteria were also abundant, followed by bacterioidetes.

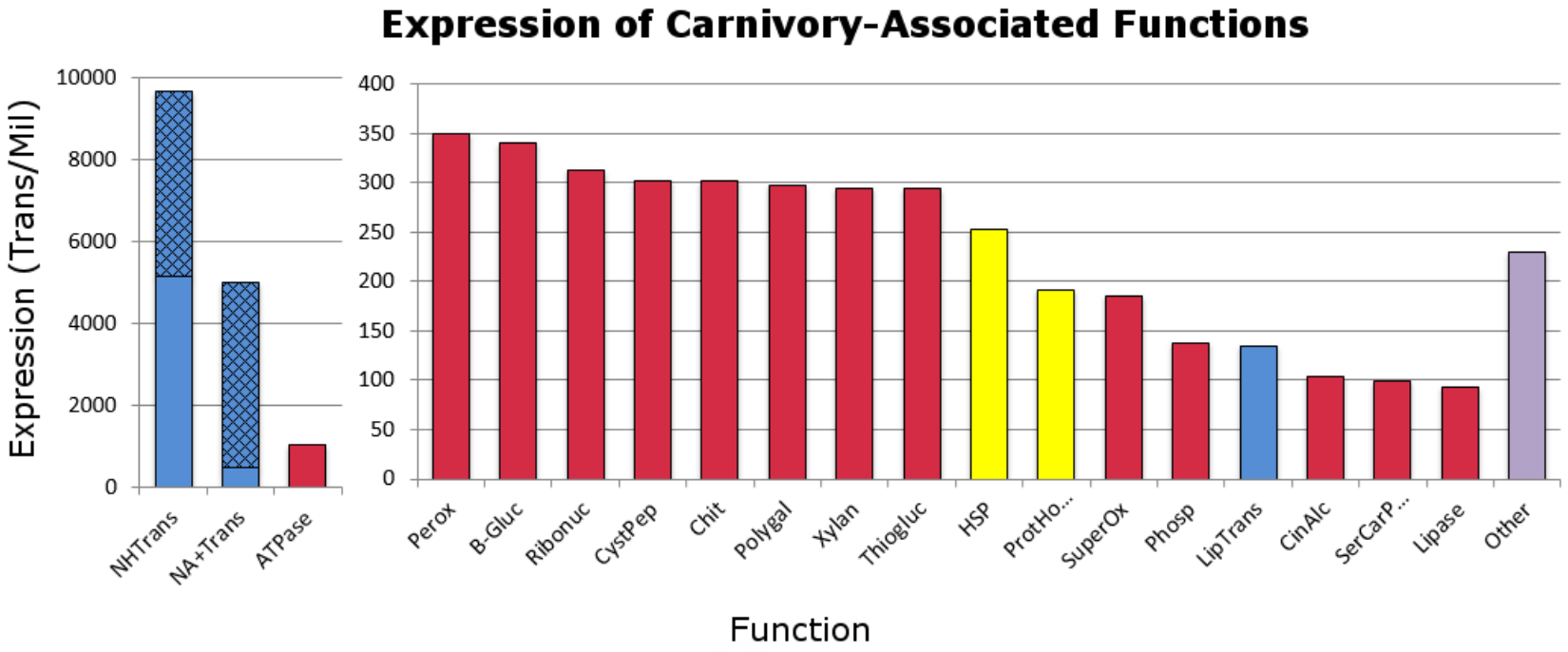

Figure 10: Expression of all carnivory-associated (broad-sense) functions as shown by total transcription of all matching genes in units of transcripts per million. *Functions at expression levels higher than one transcript per thousand are shown in the left sub-chart, functions between 90 and 1000 transcripts/million are shown in the right sub-chart. Seventeen functions scoring below 90/mil are combined into the column "Other". Cross-hatched column portions indicate contribution by Sarracenia alata, while non-cross-hatched bars are microbial contribution. Blue indicates transport functions, red indicates enzymatic catabolysis, yellow indicates stress response/protein folding, and lilac shows miscellaneous/other. Numerical representation of expression and a key for function abbreviations can be found in Table 18.*

Discussion

The finding of few unique *Sarracenia alata* genes at enormously high transcriptional levels presents an interesting situation within this system. The finding of no carnivory-associated enzyme production genes originating from *Sarracenia alata* is consistent with the null hypothesis that all digestive enzyme activity is actually a product of inquiline microbes. The microbes were found to devote a substantial portion of their genes (10.3% total) to functions previously associated with plant carnivory, but while these comprised >99.9% of unique carnivory-associated genes, they represented only slightly more than half of all carnivory-associated transcripts. A substantial portion (45.4%) of carnivory-associated transcripts were found to map to a single *S. alata* sodium- and ammonium transport gene. Within the microbiome, transcription was divided among transportation functions and a wide range of enzymatic activities.

In addition to its more conventional carnivory-associated functions, *Sarracenia alata* likely performs a substantial management role in determining the function of its microbiome. Over a quarter of all transcribed *S. alata* genes were identified as potentially encoding AMPs. Furthermore, these genes were over-represented at the transcriptional level, comprising nearly 42% of all *S. alata* transcripts. When surveying the *S. alata* predicted proteome and proteomes of other carnivorous plants (*Cephalotus follicularis* and *Genlisea aurea*) for AMPs, the findings for *S. alata* were in line with other taxa; however, it was found that the proteomes of non-carnivorous taxa tested (*Arabidopsis thaliana, Actinidia chinensis,* and *Glycine max*) were significantly more enriched in AMPs ($\alpha = 0.05$, $p = 0.0045$, Figure 11).

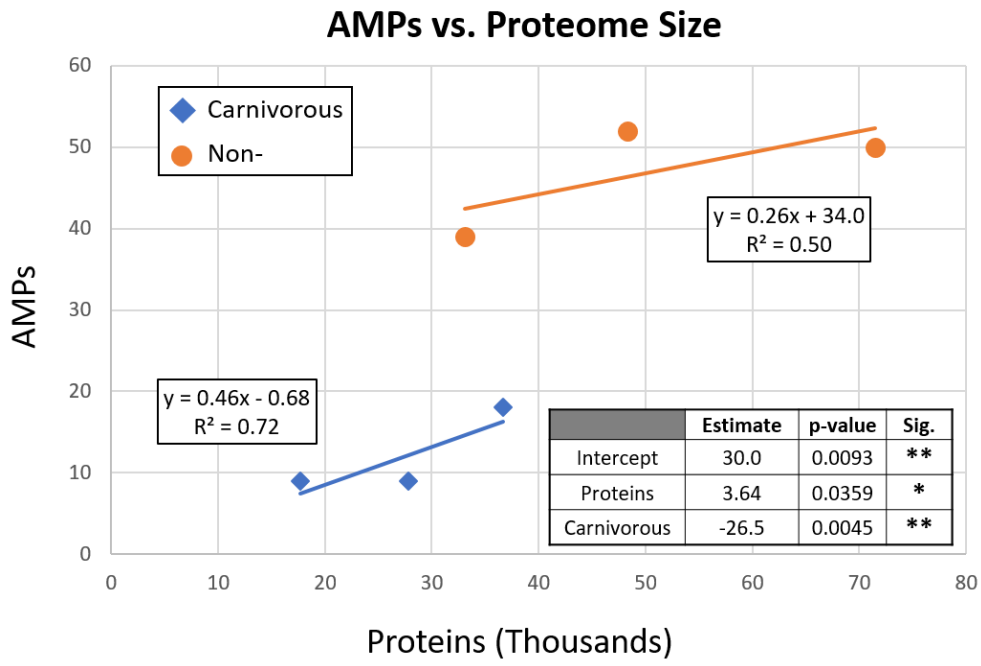

Figure 11: Anti-microbial peptides (AMPs) detected in predicted proteomes, by proteome size. *Point-markers show the protein and AMP counts of individual taxa, while lines show linear models of AMP representation for carnivorous (blue) and non-carnivorous (orange) taxa. Inset boxes (lower-left, top-right) show linear model and coefficient of determination (R^2) for each taxon set. Inset table (bottom-right) shows parameters of linear regression (AMPs ~ Proteins + Carnivory). Significance codes: *: $p < 0.05$; **: $p < 0.01$, ***: $p < 0.001$. Carnivorous taxa included are:* Cephalotus follicularis, Genlisea aurea*, and* Sarracenia alata*; non-carnivorous taxa are:* Actinidia chinensis, Glycine max*, and* Arabidopsis thaliana*.*

While these findings are consistent with the hypothesis that *S. alata* is reliant on microbial symbionts for digestion, it is at odds with the results of Chapter 3, showing that *S. alata* devotes an unusually large portion of its genes to carnivorous functions including enzyme production. The question remains as to why these genes are present in large numbers in the genome if there is no indication of their expression, despite the large number of host-plant transcripts sequenced (over two-thirds of all transcripts). One possibility is that these functions are simply not expressed at the point in the growing season when samples were collected, or expression was otherwise not activated by a required stimulus (e.g., a critical mass of prey). Alternatively, these functions may be expressed at a location in the plant where transcripts would be unlikely to appear in the fluid. We had hypothesized that most expression

related to carnivory functions would be in the cells of the pitcher's inner layer, but if enzymes were produced elsewhere in the plant and then transported into the pitcher lumen, or if intact proteins were transported out as suggested by Plummer & Kethley (1964), this may not be the case. Finally, it is possible that *S. alata* has transitioned to a "post-carnivory" lifestyle, as seen in some *Nepenthes* species which have switched to a reliance on symbioses over their own enzymatic production (Clarke, Moran & Chin, 2010; Grafe et al., 2011b).

Characterization of taxon composition of the *Sarracenia* fluid shows dominance by prokaryotic organisms. The fluid's bacterial assemblage is broadly consistent with the findings of Koopman et al. (2010), with proteobacteria (primarily enterobacteria) being most common (Table 9).

	Contigs	**Transcripts**
Bacteria	10,618	170,422.5
Terrabacteria Group	2,514	52,979.8
Firmicutes	1,690	14,339.5
Actinobacteria	578	21,844.4
Proteobacteria	4,123	45,899.4
Alpha-	963	15,382.6
Beta-	650	18,570.3
Gamma-	2,375	10,288.8
Delta-	92	1,139.6
FCB Group	2,943	41,456.3
Bacterioidetes	866	10,981.2
Chlorobi	1	0.0
Cyanobacteria	77	7,365.3

Table 9: Representation of bacterial taxa in pitcher fluid as determined by contigs and transcription level. *Transcription is shown in units of transcripts per million. More specific taxonomic groups are listed under their parent group, with increasing indentation.*

Conclusions

While this study supports the previous consensus that *Sarracenia* digestive enzymes are entirely the product of microbial activity, it also suggests that the role of the host plant in this interaction has been underestimated. In addition to specialization for rapid uptake of the products of digestion, *S. alata* is also likely to play a regulatory and organizational role in the composition of its microbial assemblage. This suggests a more complex interplay of symbiotic organisms than incidental microbes simply happening into the pitcher environment. *Sarracenia alata* has evolved to co-opt the microbiome of the ants it traps, taking advantage of bacterial strains pre-adapted to digest arthropod prey material. The plant then employs high levels of AMP activity to exclude deleterious bacterial species and maximize the representation of those most advantageous in making its needed resources available. This elegant solution allows *S. alata* to always possess the microbes it requires even without any sort of vertical transmission or tight co-adaptive symbiosis, as the microbiome is effectively "packaged" with the prey.

Chapter 5: Summary & Conclusions

In this study, I sought to investigate the mechanisms by which novel features and strategies can evolve, using carnivorous plants as a study system. In this polyphyletic group of organisms, highly similar phenotypes have arisen in response to similar evolutionary pressures; however, due to the lack of recent common ancestors between carnivorous lineages, I did not expect that matching changes would occur in a specific set of orthologous genes. Instead, I hypothesized that different genes across the species' genomes would be utilized to produce the same effects. In particular, if certain functions are particularly important to the evolution of a particular syndrome of traits, I expected these functions to show increased representation in the genome. In the case of carnivorous plant taxa, the functions expected to be of particular importance were those involved in active digestion of proteins, lipids, chitins, and other insect components, as well as those involved in the transport of nutrients, ions, and water.

Using four carnivorous plant taxa, compared to a distribution of typical-plant gene representation, I statistically demonstrated that increased representation of the expected functional categories does occur in some, but not all cases. While two of the taxa tested had post-correction significant effects in over a quarter of the expected functions, another had only one function found to be significantly overrepresented, and the in the fourth I found no overrepresentation whatsoever. Interestingly, while some functions were consistently overrepresented (e.g., *alternative oxidase activity* in 75% of taxa; *phospholipase activity* in 50% of taxa), most were unique to a particular taxon, with no overlap. From these findings, I conclude that, while my initial hypothesis is correct in some cases, more frequently taxa have evolved similar strategies and phenotypes by leveraging the use of different gene functions. Despite the unpredictability of the exact functions a taxon will specialize in, it was possible to predict a set of these functions, with taxa undergoing significant effects in a subset of these. For taxa that had little or no

gene overrepresentation in the predicted functions, it is possible that specialization has only occurred via functions outside the predicted set; however, it is also likely that changes have occurred through copy-neutral routes that would not have been detected by the methods used. Examples of this are functional up-regulation through increased transcription or translation. Instead of evolutionary specification of duplicated genes, functional diversity may instead arise from post-translational changes to proteins or differential RNA splicing.

This method of analysis, relying on convergence of function instead of shared gene ancestry, is expandable to a wide range of questions and systems. It provides a tool for quantitative comparisons in situations where most conclusions are qualitative and drawn *post hoc.* The primary limitation of the method is that, as functions of interest must be specified *a priori* by Gene Ontology identifier, some literature review or pilot research is required before any tests can be conducted. The method would also be expected to function more reliably in situations where a syndrome can be defined by a small number of functions, with statistical power lost in multiple-testing correction when functional sets are large. The most obvious use case for this methodology is in testing the similarity of adaptive strategies of distantly related organisms occupying similar niche space; however, it could also prove useful for better understanding the adaptation of pesticide or medication resistance in pathogenic organisms, or of the origins of similar cancers.

After studying general effects across carnivorous plant taxa as a group, I focused onto *Sarracenia alata* to test the validity of my previous conclusions in a specific case, as well as to test the conclusions of past studies. This required the assembly and annotation of this species's large and complex genome, a substantial undertaking. While I was unable to obtain chromosome-length scaffolds as I had hoped, a draft reference genome was produced that rivals the completeness and contiguity of many plant assemblies. Annotation of this genome was consistent with the expectations of large plant genomes – most genomic material was occupied by a diverse array of repetitive and transposable elements, with more pseudogenes than genes. This is also consistent with a history of genome duplication and

subsequent reduction, leaving behind degenerate sequence as non-adaptive redundant genes are subfunctionalized.

Findings from analyses of *S. alata* genes annotated to carnivory-associated functions were far more surprising. As *Sarracenia* species have been shown to rely heavily on microbial symbioses and lack discernible secretory structures for digestive enzymes, I had hypothesized that *S. alata* would possess little to no signal of increased representation of carnivory functions. The results were quite the opposite, with *S. alata* possessing a greater portion of carnivory-associated genes than any other taxon tested. Five categories of digestive enzyme activity were statistically elevated, though functions related to transport of water and nutrients were not. This result is at odds with the common assertion that *Sarracenia* species are mostly or entirely passive in their carnivorous lifestyle, instead suggesting that the host plant is as involved as any other carnivorous taxon, if not more.

My genomic research in *S. alata* provided several resources that will be of value in future studies. First is the genome itself, as well as its accompanying annotations. While the number of available plant genomes continues to grow, most of these lack annotations. Taxonomic representation is also very uneven, with only one other species (*Actinidia chinensis*) sequenced from *Sarracenia'*s section of Ericales. The *S. alata* reference genome sequence should prove useful to carnivorous plant researchers in particular, but also to those studying Ericales, or constructing large data sets where thorough phylogenetic coverage is important. Methodologically, my adaptations to the standard MAKER-P annotation pipeline may prove valuable to other researchers studying large, complex genomes without unlimited computational resources. My sequencing pipeline, unfortunately, did not deliver meaningful improvement through merging parallel assemblies; however, the lessons learned here can inform future studies. When performing *de novo* assembly of large plant genomes, PacBio SMRT sequencing should be the primary focus, with a smaller amount of Illumina sequence collected for subsequent polishing. At least in the case of this study, a pipeline using only Canu and Pilon would have produced nearly identical results, with moderate savings of resources and substantial reduction of effort.

The availability of an assembled *Sarracenia alata* genome has already opened previously-unavailable avenues of research. In particular, the meta-transcriptomic portion of this study would not have been possible without this reference, which allowed *S. alata* transcripts to be sorted from microbial transcripts in the pitcher fluid. While analysis of the annotated genes of *S. alata* was highly suggestive of adaptation for plant carnivory, it could not answer whether or not these genes are expressed, nor could it determine the relative contribution of microbial symbionts. By pairing expression data with genomic data, it was possible to sort transcripts and fill in the gaps in this knowledge.

While roughly two-thirds of transcripts were found to originate from the *S. alata* host plant, no transcripts were found to encode any of the enzymatic functions that had been detected in genomic functional analyses; however, a majority of the carnivory functions tested were expressed by the pitchers' symbiotic organisms. Of functions expected to be associated with plant carnivory, only those related to transport had any activity at all, which was mostly devoted to a single highly-expressed gene. Instead of my expectation of high enzymatic expression based on the findings of Chapter 3's genomic analyses, the most highly-expressed genes were a subset that had not been assigned any function. Though similar sequences have been identified in other plant genomes, most had not been assigned any meaningful identifiers, labeled only as predicted or hypothetical proteins.

I hypothesized that these mystery transcripts were encoding anti-microbial peptides (AMPs), an eclectic category of small proteins defined by their anti-microbial activity. AMPs had been implicated in determining digestive microbiome in animals, as well as in immune function in plants; however, they had never been investigated in any carnivorous plant taxa. A BLAST-search against an AMP-specific reference database produced matches for a large number of the high-expression unknown transcipts, supporting this hypothesis.

These findings, all taken together, paint an exciting picture of the evolution of *Sarracenia alata*, and carnivorous plants in general. First, while some aspects of the evolution of plant carnivory are predictable, the specific evolutionary trajectories of individual lineages vary widely. In *Sarracenia alata,*

this appears to have taken the form of increased enzymatic activity through duplication of adaptive gene copies (or preferential retention following whole-genome duplication). Second, a lineage may not retain the same strategy through its history. Present-day *Sarracenia* have not been observed to possess digestive glands, and no expression of these gene types from pitcher wall cells could be detected. Instead, findings are consistent with a switch to reliance on microbial symbioses. This type of change is not unique to *Sarracenia alata*, as a number of *Nepenthes* taxa have well-documented adaption to symbioses with vertebrates. Finally, there may be core aspects of carnivorous plant function that remain undocumented. For example, while evidence from *S. alata* strongly suggests that AMPs play a major role in its digestive biology, there is no evidence of any investigation of these peptides in other carnivorous plant taxa. As other lineages are studied, it seems probable that other surprising findings will emerge.

Future research

While this study will likely be the end of my research in *Sarracenia*, a number of questions are ripe for future investigation. First, as the transcriptomic analysis performed using mixed RNA transcripts from within the pitcher fluid, it is not known if conflicting information is present in the RNA from other parts of the plant. It is possible that the plant does possess its own digestive secretions, but produced via an avenue other than the macrocellular glands seen in other pitcher plant lineages. Second, can the anti-microbial peptides predicted from transcripts be isolated from *S. alata*'s digestive fluids? If so, what is their specific action, and what does this tell us about *S. alata*'s relationship with different microbial groups. Third, can the assembled reference genome of *S. alata* be used as a basis for assembly of other *Sarracenia* genomes? Do these species' genomes differ in their functional make-up, and if so, what can we learn about niche specialization within the genus? Fourth, can additional genomic data be used to clarify the phylogeny of the *Sarraceniaceae*, which has remained uncertain for decades? Using a whole-genome approach, will it finally be possible to determine the status of the many uncertain *Sarracenia* taxa?

Resources

For use in future studies, such as the examples listed above or any others, a number of resources are available from my research. First is the reference genome of *Sarracenia alata,* available on NCBI GenBank. Next are the annotations including Gene Ontology coding, which are available accompanying the genome. Raw read data, including genomic Illumina, PacBio, and BioNano sequence, is available on the NCBI Sequence Read Archive. All scripts used can be found on GitHub. Carnivorous plant carnivory gene count data tables are also provided with these scripts.

References

Adamec L. 2002. Leaf absorption of mineral nutrients in carnivorous plants stimulates root nutrient uptake. *New Phytologist* 155:89–100. DOI: 10.1046/j.1469-8137.2002.00441.x.

Adams RM., Smith GW. 1977. An SEM survey of the five carnivorous pitcher plant genera. *American Journal of Botany* 64:265–272. DOI: 10.2307/2441969.

Adams KL., Wendel JF. 2005. Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology* 8:135–141. DOI: 10.1016/J.PBI.2005.01.001.

Ahn WS., Kim K-W., Bae SM., Yoon JH., Lee JM., Namkoong SE., Kim JH., Kim CK., Lee YJ., Kim Y-W. 2003. Targeted cellular process profiling approach for uterine leiomyoma using cDNA microarray, proteomics and gene ontology analysis. *International journal of experimental pathology* 84:267–79. DOI: 10.1111/J.0959-9673.2003.00362.X.

Altenhoff AM., Studer RA., Robinson-Rechavi M., Dessimoz C. 2012. Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. *PLoS Computational Biology* 8:e1002514. DOI: 10.1371/journal.pcbi.1002514.

An C Il., Fukusaki EI., Kobayashi A. 2002. Aspartic proteinases are expressed in pitchers of the carnivorous plant Nepenthes alata Blanco. *Planta* 214:661–667. DOI: 10.1007/s004250100665.

Anderson B., Midgley JJ. 2003. Digestive mutualism, an alternate pathway in plant carnivory. *Oikos* 102:221–224. DOI: 10.1034/j.1600-0706.2003.12478.x.

Andrews S. 2010.FastQC: A quality control tool for high throughput sequence data. DOI: citeulike-article-id:11583827.

Ashburner M., Ball CA., Blake JA., Botstein D., Butler H., Cherry JM., Davis AP., Dolinski K., Dwight SS., Eppig JT., Harris MA., Hill DP., Issel-Tarver L., Kasarskis A., Lewis S., Matese JC., Richardson JE., Ringwald M., Rubin GM., Sherlock G. 2000a. Gene ontology: Tool for the unification of biology. *Nature Genetics* 25:25–29. DOI: 10.1038/75556.

Ashburner M., Ball CA., Blake JA., Botstein D., Butler H., Cherry JM., Davis AP., Dolinski K., Dwight SS., Eppig JT., Harris MA., Hill DP., Issel-Tarver L., Kasarskis A., Lewis S., Matese JC., Richardson JE., Ringwald M., Rubin GM., Sherlock G. 2000b. Gene Ontology: tool for the unification of biology. *Nature Genetics* 25:25–29. DOI: 10.1038/75556.

Bairoch A., Apweiler R., Wu CH., Barker WC., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin MJ., Natale DA., O'Donovan C., Redaschi N., Yeh LSL. 2005. The Universal Protein Resource (UniProt). *Nucleic Acids Research* 33. DOI: 10.1093/nar/gki070.

Balsa-Canto E., Henriques D., Gabor A., Banga JR. 2016. AMIGO2, a toolbox for dynamic modeling, optimization and control in systems biology. *Bioinformatics (Oxford, England)*:1–2. DOI: 10.1093/bioinformatics/btw411.

Bankevich A., Nurk S., Antipov D., Gurevich AA., Dvorkin M., Kulikov AS., Lesin VM., Nikolenko SI., Pham S., Prjibelski AD., Pyshkin A V., Sirotkin A V., Vyahhi N., Tesler G., Alekseyev MA., Pevzner PA. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* 19:455–477. DOI: 10.1089/cmb.2012.0021.

Barker NG., Williamson GB. 1988. Effects of a winter fire on Sarracenia alata and S. psittacina. *American Journal of Botany* 75:138–143. DOI: 10.2307/2443912.

Bemm F., Becker D., Larisch C., Kreuzer I., Escalante-Perez M., Schulze WX., Ankenbrand M., Van De Weyer AL., Krol E., Al-Rasheid KA., Mithï¿½fer A., Weber AP., Schultz J., Hedrich R. 2016. Venus flytrap carnivorous lifestyle builds on herbivore defense strategies. *Genome Research* 26:812–825. DOI: 10.1101/gr.202200.115.

Bhattarai GP., Horner JD. 2009. The Importance of Pitcher Size in Prey Capture in the Carnivorous Plant, Sarracenia alata Wood (Sarraceniaceae). *The American Midland Naturalist* 161:264–272. DOI: 10.1674/0003-0031-161.2.264.

Blanc G., Wolfe KH. 2004. Functional Divergence of Duplicated Genes Formed by Polyploidy during Arabidopsis Evolution. *THE PLANT CELL ONLINE* 16:1679–1691. DOI: 10.1105/tpc.021410.

Bloom Arnold J., Chapin Stuart F., Mooney Harold A. 1985. Resource limitation in plants- An economic analogy. *Annual review of ecology and systematics* 16:363–392. DOI: 10.1146/annurev.ecolsys.16.1.363.

Böhm J., Scherzer S., Krol E., Kreuzer I., Von Meyer K., Lorey C., Mueller TD., Shabala L., Monte I., Solano R., Al-Rasheid KAS., Rennenberg H., Shabala S., Neher E., Hedrich R. 2016. The venus flytrap dionaea muscipula counts prey-induced action potentials to induce sodium uptake. *Current Biology* 26:286–295. DOI: 10.1016/j.cub.2015.11.057.

Bork P., Sander C., Valencia A. 1993. Convergent evolution of similar enzymatic function on different protein folds: The hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein Science* 2:31–40. DOI: 10.1002/pro.5560020104.

Brodmann J., Twele R., Francke W., Yi-bo L., Xi-qiang S., Ayasse M. 2009. Orchid Mimics Honey Bee Alarm Pheromone in Order to Attract Hornets for Pollination. *Current Biology*

19:1368–1372. DOI: 10.1016/j.cub.2009.06.067.

Butler JL., Gotelli NJ., Ellison AM. 2008. Linking the brown and green: Nutrient transformation and fate in the Sarracenia microecosystem. *Ecology* 89:898–904. DOI: 10.1890/07-1314.1.

Butts CT., Bierma JC., Martin RW. 2016. Novel proteases from the genome of the carnivorous plant Drosera capensis: Structural prediction and comparative analysis. *Proteins: Structure, Function and Bioinformatics* 84:1517–1533. DOI: 10.1002/prot.25095.

Campbell MS., Holt C., Moore B., Yandell M. 2014. Genome Annotation and Curation Using MAKER and MAKER-P. *Current Protocols in Bioinformatics* 2014:4.11.1-4.11.39. DOI: 10.1002/0471250953.bi0411s48.

Cantarel BL., Korf I., Robb SMC., Parra G., Ross E., Moore B., Holt C., Alvarado AS., Yandell M. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* 18:188–196. DOI: 10.1101/gr.6743907.

Canter EJ., Cuellar-Gempeler C., Pastore AI., Miller TE., Mason OU. 2018. Predator identity more than predator richness structures aquatic microbial assemblages in *Sarracenia purpurea* leaves. *Ecology* 99:652–660. DOI: 10.1002/ecy.2128.

Caravieri FA., Ferreira AJ., Ferreira A., Clivati D., de Miranda VFO., Araújo WL. 2014. Bacterial community associated with traps of the carnivorous plants Utricularia hydrocarpa and Genlisea filiformis. *Aquatic Botany* 116:8–12. DOI: 10.1016/j.aquabot.2013.12.008.

Carbon S., Dietze H., Lewis SE., Mungall CJ., Munoz-Torres MC., Basu S., Chisholm RL., Dodson RJ., Fey P., Thomas PD., Mi H., Muruganujan A., Huang X., Poudel S., Hu JC., Aleksander SA., McIntosh BK., Renfro DP., Siegele DA., Antonazzo G., Attrill H., Brown

NH., Marygold SJ., Mc-Quilton P., Ponting L., Millburn GH., Rey AJ., Stefancsik R., Tweedie S., Falls K., Schroeder AJ., Courtot M., Osumi-Sutherland D., Parkinson H., Roncaglia P., Lovering RC., Foulger RE., Huntley RP., Denny P., Campbell NH., Kramarz B., Patel S., Buxton JL., Umrao Z., Deng AT., Alrohaif H., Mitchell K., Ratnaraj F., Omer W., Rodríguez-López M., C. Chibucos M., Giglio M., Nadendla S., Duesbury MJ., Koch M., Meldal BHM., Melidoni A., Porras P., Orchard S., Shrivastava A., Chang HY., Finn RD., Fraser M., Mitchell AL., Nuka G., Potter S., Rawlings ND., Richardson L., Sangrador-Vegas A., Young SY., Blake JA., Christie KR., Dolan ME., Drabkin HJ., Hill DP., Ni L., Sitnikov D., Harris MA., Hayles J., Oliver SG., Rutherford K., Wood V., Bahler J., Lock A., De Pons J., Dwinell M., Shimoyama M., Laulederkind S., Hayman GT., Tutaj M., Wang SJ., D'Eustachio P., Matthews L., Balhoff JP., Balakrishnan R., Binkley G., Cherry JM., Costanzo MC., Engel SR., Miyasato SR., Nash RS., Simison M., Skrzypek MS., Weng S., Wong ED., Feuermann M., Gaudet P., Berardini TZ., Li D., Muller B., Reiser L., Huala E., Argasinska J., Arighi C., Auchincloss A., Axelsen K., Argoud-Puy G., Bateman A., Bely B., Blatter MC., Bonilla C., Bougueleret L., Boutet E., Breuza L., Bridge A., Britto R., Hye-A-Bye H., Casals C., Cibrian-Uhalte E., Coudert E., Cusin I., Duek-Roggli P., Estreicher A., Famiglietti L., Gane P., Garmiri P., Georghiou G., Gos A., Gruaz-Gumowski N., Hatton-Ellis E., Hinz U., Holmes A., Hulo C., Jungo F., Keller G., Laiho K., Lemercier P., Lieberherr D., Mac- Dougall A., Magrane M., Martin MJ., Masson P., Natale DA., O'Donovan C., Pedruzzi I., Pichler K., Poggioli D., Poux S., Rivoire C., Roechert B., Sawford T., Schneider M., Speretta E., Shypitsyna A., Stutz A., Sundaram S., Tognolli M., Wu C., Xenarios I., Yeh LS., Chan J., Gao S., Howe K., Kishore R., Lee R., Li Y., Lomax J., Muller HM., Raciti D., Van Auken K., Berriman M., Stein, Paul Kersey L., W. Sternberg

P., Howe D., Westerfield M. 2017. Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium. *Nucleic Acids Research* 45:D331–D338. DOI: 10.1093/nar/gkw1108.

Carstens BC., Satler JD. 2013. The carnivorous plant described as Sarracenia alata contains two cryptic species. *Biological Journal of the Linnean Society* 109:737–746. DOI: 10.1111/bij.12093.

Center OS. 1987. Ohio Supercomputer Center.

Chain P., Lamerdin J., Larimer F., Regala W., Lao V., Land M., Hauser L., Hooper A., Klotz M., Norton J., Sayavedra-Soto L., Arciero D., Hommes N., Whittaker M., Arp D. 2003. Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph Nitrosomonas europaea. *Journal of Bacteriology* 185:2759–2773. DOI: 10.1128/JB.185.9.2759-2773.2003.

Chapman KD. 1998. Phospholipase activity during plant growth and development and in response to environmental stress. *Trends in Plant Science* 3:419–426. DOI: 10.1016/S1360-1385(98)01326-0.

Clarke C., Moran JA., Chin L. 2010. Mutualism between tree shrews and pitcher plants: Perspectives and avenues for future research. *Plant Signaling and Behavior* 5:1187–1189. DOI: 10.4161/psb.5.10.12807.

Conesa A., Götz S. 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics* 2008. DOI: 10.1155/2008/619832.

Conesa A., Götz S., García-Gómez JM., Terol J., Talón M., Robles M. 2005. Blast2GO: A

universal annotation and visualization tool in functional genomics research. Application note. *Bioinformatics* 21:3674–3676. DOI: 10.1093/bioinformatics/bti610.

Cullen TW., Schofield WB., Barry NA., Putnam EE., Rundell EA., Trent MS., Degnan PH., Booth CJ., Yu H., Goodman AL. 2015. Antimicrobial peptide resistance mediates resilience of prominent gut commensals during inflammation. *Science* 347:170–175. DOI: 10.1126/science.1260580.

Dabney A., Storey JD., Warnes GR. 2010. Q-value estimation for false discovery rate control.

Darwin C., Darwin F. 1889. *Insectivorous Plants*.

Day D a., Wiskich JT. 1995. Regulation of alternative oxidase activity in higher plants. *Journal of bioenergetics and biomembranes*.

Doolittle RF. 1994. Convergent evolution: the need to be explicit. *Trends in Biochemical Sciences* 19:15–18. DOI: 10.1016/0968-0004(94)90167-8.

Ellison AM. 2006. Nutrient limitation and stoichiometry of carnivorous plants. In: *Plant Biology*. 740–747. DOI: 10.1055/s-2006-923956.

Ellison AM., Butler ED., Hicks EJ., Naczi RFC., Calie PJ., Bell CD., Davis CC. 2012. Phylogeny and biogeography of the carnivorous plant family Sarraceniaceae. *PLoS ONE* 7. DOI: 10.1371/journal.pone.0039291.

Ellison AM., Farnsworth EJ. 2005. The cost of carnivory for Darlingtonia californica (Sarraceniaceae): Evidence from relationships among leaf traits. *American Journal of Botany* 92:1085–1093. DOI: 10.3732/ajb.92.7.1085.

Ellison AM., Gotelli NJ. 2001. Evolutionary ecology of carnivorous plants. *Trends in Ecology*

and Evolution 16:623–629. DOI: 10.1016/S0169-5347(01)02269-8.

Ellison AM., Gotelli NJ., Brewer JS., Cochran-Stafira DL., Kneitel JM., Miller TE., Worley AC., Zamora R. 2003. The evolutionary ecology of carnivorous plants. In: *Advances in Ecological Research Vol 33*. 1–74. DOI: 10.1016/S0065-2504(03)33009-0.

Flagel LE., Wendel JF. 2009. Gene duplication and evolutionary novelty in plants. *New Phytologist* 183:557–564. DOI: 10.1111/j.1469-8137.2009.02923.x.

Fleischmann A., Heubl G. 2009. Overcoming DNA extraction problems from carnivorous plants. *Anales del Jardín Botánico de Madrid* 66:209–215. DOI: 10.3989/ajbm.2198.

Foote AD., Liu Y., Thomas GWC., Vinař T., Alföldi J., Deng J., Dugan S., van Elk CE., Hunter ME., Joshi V., Khan Z., Kovar C., Lee SL., Lindblad-Toh K., Mancia A., Nielsen R., Qin X., Qu J., Raney BJ., Vijay N., Wolf JBW., Hahn MW., Muzny DM., Worley KC., Gilbert MTP., Gibbs RA. 2015. Convergent evolution of the genomes of marine mammals. *Nature Genetics* 47:272–275. DOI: 10.1038/ng.3198.

Franzosa EA., Morgan XC., Segata N., Waldron L., Reyes J., Earl AM., Giannoukos G., Boylan MR., Ciulla D., Gevers D., Izard J., Garrett WS., Chan AT., Huttenhower C. 2014. Relating the metatranscriptome and metagenome of the human gut. *Proceedings of the National Academy of Sciences of the United States of America* 111:E2329-38. DOI: 10.1073/pnas.1319284111.

Fukushima K., Fang X., Alvarez-Ponce D., Cai H., Carretero-Paulet L., Chen C., Chang T-H., Farr KM., Fujita T., Hiwatashi Y., Hoshi Y., Imai T., Kasahara M., Librado P., Mao L., Mori H., Nishiyama T., Nozawa M., Pálfalvi G., Pollard ST., Rozas J., Sánchez-Gracia A., Sankoff D., Shibata TF., Shigenobu S., Sumikawa N., Uzawa T., Xie M., Zheng C., Pollock

DD., Albert VA., Li S., Hasebe M. 2017. Genome of the pitcher plant Cephalotus reveals genetic changes associated with carnivory. *Nature Ecology & Evolution* 1:0059. DOI: 10.1038/s41559-016-0059.

Gallie DR., Chang SC. 1997. Signal transduction in the carnivorous plant Sarracenia purpurea. Regulation of secretory hydrolase expression during development and in response to resources. *Plant physiology* 115:1461–71. DOI: 10.1104/PP.115.4.1461.

Gifford SM., Sharma S., Rinta-Kanto JM., Moran MA. 2011. Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *The ISME journal* 5:461–72. DOI: 10.1038/ismej.2010.141.

Giribet G. 2016. New animal phylogeny: future challenges for animal phylogeny in the age of phylogenomics. *Organisms Diversity and Evolution* 16:419–426. DOI: 10.1007/s13127-015-0236-4.

Givnish TJ. 2015. New evidence on the origin of carnivorous plants. *Proceedings of the National Academy of Sciences* 112:10–11. DOI: 10.1073/pnas.1422278112.

Givnish TJ., Burkhardt EL., Happel RE., Weintraub JD. 1984. Carnivory in the Bromeliad Brocchinia reducta, with a Cost/Benefit Model for the General Restriction of Carnivorous Plants to Sunny, Moist, Nutrient-Poor Habitats. *The American Naturalist* 124:479–497. DOI: 10.1086/284289.

Gotelli NJ., Ellison AM., Ballif BA. 2012. Environmental proteomics, biodiversity statistics and food-web structure. *Trends in Ecology & Evolution* 27:436–442. DOI: 10.1016/J.TREE.2012.03.001.

Grabherr MG., Haas BJ., Yassour M., Levin JZ., Thompson DA., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., Di Palma F., Birren BW., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29:644–652. DOI: 10.1038/nbt.1883.

Grafe TU., Schoner CR., Kerth G., Junaidi A., Schoner MG. 2011a. A novel resource-service mutualism between bats and pitcher plants. *Biology Letters* 7:436–439. DOI: 10.1098/rsbl.2010.1141.

Grafe TU., Schöner CR., Kerth G., Junaidi A., Schöner MG. 2011b. A novel resource-service mutualism between bats and pitcher plants. *Biology letters* 7:436–9. DOI: 10.1098/rsbl.2010.1141.

Greub G., Raoult D. 2003. History of the ADP/ATP-translocase-encoding gene, a parasitism gene transferred from a Chlamydiales ancestor to plants 1 billion years ago. *Applied and environmental microbiology* 69:5530–5. DOI: 10.1128/AEM.69.9.5530-5535.2003.

Gupta RS. 2000. The natural evolutionary relationships among prokaryotes. *Critical reviews in microbiology*. DOI: 10.1080/10408410091154219.

Hackl T. Evaluation of assembly strategies for a complex genome – Development of novel approaches and bioinformatics solutions.

Hatano N., Hamada T. 2008. Proteome analysis of pitcher fluid of the carnivorous plant *Nepenthes alata*. *Journal of Proteome Research* 7:809–816. DOI: 10.1021/pr700566d.

Hess S., Frahm J-P., Theisen I. 2005. Evidence of zoophagy in a second liverwort species,

Pleurozia purpurea. *The Bryologist* 108:212–218. DOI: 10.1639/6.

Holmans P., Green EK., Pahwa JS., Ferreira MAR., Purcell SM., Sklar P., Owen MJ., O'Donovan MC., Craddock N. 2009. Gene Ontology Analysis of GWA Study Data Sets Provides Insights into the Biology of Bipolar Disorder. *American Journal of Human Genetics* 85:13–24. DOI: 10.1016/j.ajhg.2009.05.011.

Huang S., Ding J., Deng D., Tang W., Sun H., Liu D., Zhang L., Niu X., Zhang X., Meng M., Yu J., Liu J., Han Y., Shi W., Zhang D., Cao S., Wei Z., Cui Y., Xia Y., Zeng H., Bao K., Lin L., Min Y., Zhang H., Miao M., Tang X., Zhu Y., Sui Y., Li G., Sun H., Yue J., Sun J., Liu F., Zhou L., Lei L., Zheng X., Liu M., Huang L., Song J., Xu C., Li J., Ye K., Zhong S., Lu BR., He G., Xiao F., Wang HL., Zheng H., Fei Z., Liu Y. 2013. Draft genome of the kiwifruit Actinidia chinensis. *Nature Communications* 4. DOI: 10.1038/ncomms3640.

Ibarra-Laclette E., Albert VA., Pérez-Torres CA., Zamudio-Hernández F., Ortega-Estrada M de J., Herrera-Estrella A., Herrera-Estrella L. 2011. Transcriptomics and molecular evolutionary rate analysis of the bladderwort (Utricularia), a carnivorous plant with a minimal genome. *BMC Plant Biology* 11:101. DOI: 10.1186/1471-2229-11-101.

Jensen RA. 1976. Enzyme Recruitment in Evolution of New Function. *Annual Review of Microbiology* 30:409–425. DOI: 10.1146/annurev.mi.30.100176.002205.

Kamono A., Meyer M., Cavalier-Smith T., Fukui M., Fiore-Donno AM. 2013. Exploring slime mould diversity in high-altitude forests and grasslands by environmental RNA analysis. *FEMS Microbiology Ecology* 84:98–109. DOI: 10.1111/1574-6941.12042.

Kim MY., Lee S., Van K., Kim T-H., Jeong S-C., Choi I-Y., Kim D-S., Lee Y-S., Park D., Ma J., Kim W-Y., Kim B-C., Park S., Lee K-A., Kim DH., Kim KH., Shin JH., Jang YE., Kim

KD., Liu WX., Chaisan T., Kang YJ., Lee Y-H., Kim K-H., Moon J-K., Schmutz J., Jackson SA., Bhak J., Lee S-H. 2010. Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome. *Proceedings of the National Academy of Sciences* 107:22032–22037. DOI: 10.1073/pnas.1009526107.

Koopman MM., Carstens BC. 2011. The Microbial Phyllo geography of the Carnivorous Plant Sarracenia alata. *Microbial Ecology* 61:750–758. DOI: 10.1007/s00248-011-9832-9.

Koopman MM., Fuselier DM., Hird S., Carstens BC. 2010. The carnivorous pale pitcher plant harbors diverse, distinct, and time-dependent bacterial communities. *Applied and Environmental Microbiology* 76:1851–1860. DOI: 10.1128/AEM.02440-09.

Koren S., Walenz BP., Berlin K., Miller JR., Bergman NH., Phillippy AM. 2017. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Research* 27:722–736. DOI: 10.1101/gr.215087.116.

Lan T., Renner T., Ibarra-Laclette E., Farr KM., Chang T-H., Cervantes-Pérez SA., Zheng C., Sankoff D., Tang H., Purbojati RW., Putra A., Drautz-Moses DI., Schuster SC., Herrera-Estrella L., Albert VA. 2017. Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome. *Proceedings of the National Academy of Sciences* 114:E4435–E4441. DOI: 10.1073/pnas.1702072114.

Langmead B., Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9:357–359. DOI: 10.1038/nmeth.1923.

Law R., Dieckmann U. 1998. Symbiosis through exploitation and the merger of lineages in evolution. *Proceedings of the Royal Society B: Biological Sciences* 265:1245–1253. DOI: 10.1098/rspb.1998.0426.

Leushkin E V., Sutormin R a., Nabieva ER., Penin A a., Kondrashov AS., Logacheva MD. 2013. The miniature genome of a carnivorous plant Genlisea aurea contains a low number of genes and short non-coding sequences. *BMC genomics* 14:476. DOI: 10.1186/1471-2164-14-476.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. DOI: 10.1093/bioinformatics/btp352.

Lloyd FE. 1934. Is Roridula a carnivorous plant? *Canadian Journal of Research*:780–786.

Losos JB. 2011. CONVERGENCE, ADAPTATION, AND CONSTRAINT. *Evolution* 65:1827–1840. DOI: 10.2307/41240779.

Lowe TM., Eddy SR. 1996. TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* 25:955–964. DOI: 10.1093/nar/25.5.0955.

Luciano CS., Newell SJ. 2017a. Effects of prey, pitcher age, and microbes on acid phosphatase activity in fluid from pitchers of Sarracenia purpurea (Sarraceniaceae). *PLoS ONE* 12. DOI: 10.1371/journal.pone.0181252.

Luciano CS., Newell SJ. 2017b. Effects of prey, pitcher age, and microbes on acid phosphatase activity in fluid from pitchers of Sarracenia purpurea (Sarraceniaceae). *PLOS ONE* 12:e0181252. DOI: 10.1371/journal.pone.0181252.

McLennan DA. 2008. The Concept of Co-option: Why Evolution Often Looks Miraculous. *Evolution: Education and Outreach* 1:247–258. DOI: 10.1007/s12052-008-0053-8.

Midgley JJ., Stock WD. 1998. Natural abundance of d15N confirms insectivorous habit of Roridula gorgonias, despite it having no proteolytic enzymes. *Annals of Botany* 82:387–388. DOI: 10.1006/anbo.1998.0684.

Mithöfer A. 2011. Carnivorous pitcher plants: Insights in an old topic. *Phytochemistry* 72:1678–1682. DOI: 10.1016/j.phytochem.2010.11.024.

Moll KM., Zhou P., Ramaraj T., Fajardo D., Devitt NP., Sadowsky MJ., Stupar RM., Tiffin P., Miller JR., Young ND., Silverstein KAT., Mudge J. 2017. Strategies for optimizing BioNano and Dovetail explored through a second reference quality assembly for the legume model, Medicago truncatula. *BMC Genomics* 18. DOI: 10.1186/s12864-017-3971-4.

Monson RK. 2003. Gene duplication, neofunctionalization, and the evolution of C4 photosynthesis. *International Journal of Plant Sciencesournal of plant sciences* 164:S43–S54. DOI: Doi 10.1086/368400.

Moore RC., Purugganan MD. 2005. The evolutionary dynamics of plant duplicate genes. *Current Opinion in Plant Biology* 8:122–128. DOI: 10.1016/j.pbi.2004.12.001.

Muegge BD., Kuczynski J., Knights D., Clemente JC., González A., Fontana L., Henrissat B., Knight R., Gordon JI. 2011. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science (New York, N.Y.)* 332:970–4. DOI: 10.1126/science.1198719.

Myers EW., Sutton GG., Delcher AL., Dew IM., Fasulo DP., Flanigan MJ., Kravitz SA., Mobarry CM., Reinert KHJ., Remington KA., Anson EL., Bolanos RA., Chou HH., Jordan CM., Halpern AL., Lonardi S., Beasley EM., Brandon RC., Chen L., Dunn PJ., Lai Z., Liang Y., Nusskern DR., Zhan M., Zhang Q., Zheng X., Rubin GM., Adams MD., Venter

JC. 2000. A whole-genome assembly of Drosophila. *Science* 287:2196–2204. DOI: 10.1126/science.287.5461.2196.

Neumann RS., Kumar S., Haverkamp THA., Shalchian-Tabrizi K. 2014. BLASTGrabber: A bioinformatic tool for visualization, analysis and sequence selection of massive BLAST data. *BMC Bioinformatics* 15. DOI: 10.1186/1471-2105-15-128.

Ohyanagi H. 2006. The Rice Annotation Project Database (RAP-DB): hub for Oryza sativa ssp. japonica genome information. *Nucleic Acids Research* 34:D741–D744. DOI: 10.1093/nar/gkj094.

Ostaff MJ., Stange EF., Wehkamp J. 2013. Antimicrobial peptides and gut microbiota in homeostasis and pathology. *EMBO Molecular Medicine* 5:1465–1483. DOI: 10.1002/emmm.201201773.

Owen TPJ., Lennon KA., Santo MJ., Anderson AN. 1999. Pathways for Nutrient Transport in the Pitchers of the Carnivorous Plant Nepenthes alata. *Annals of Botany* 84:459–466. DOI: 10.1006/anbo.1998.0944.

Palenik B., Haselkorn R. 1992. Multiple evolutionary origins of prochlorophytes, the chlorophyll b-containing prokaryotes. *Nature* 355:265–267. DOI: 10.1038/355265a0.

Płachno BJ., Adamec L., Lichtscheidl IK., Peroutka M., Adlassnig W., Vrba J. 2006. Fluorescence Labelling of Phosphatase Activity in Digestive Glands of Carnivorous Plants. *Plant Biology* 8:813–820. DOI: 10.1055/s-2006-924177.

Plummer G., Jackson T. 1963. Bacterial activities within the sarcophagus of the insectivorous plant, Sarracenia flava. *American Midland Naturalist* 69:462–469. DOI: 10.2307/2422922.

Plummer G., Kethley J. 1964. Foliar absorption of amino acids, peptides, and other nutrients by the pitcher plant, Sarracenia flava. *Botanical Gazette* 125:245–260. DOI: 10.1086/336280.

Primack RB. 1987. Relationships Among Flowers, Fruits, and Seeds. *Ann. Rev. Ecol. Sysl* 18:409–30.

Pruitt KD., Tatusova T., Maglott DR. 2007. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* 35. DOI: 10.1093/nar/gkl842.

Qi X., Li M-W., Xie M., Liu X., Ni M., Shao G., Song C., Kay-Yuen Yim A., Tao Y., Wong F-L., Isobe S., Wong C-F., Wong K-S., Xu C., Li C., Wang Y., Guan R., Sun F., Fan G., Xiao Z., Zhou F., Phang T-H., Liu X., Tong S-W., Chan T-F., Yiu S-M., Tabata S., Wang J., Xu X., Lam H-M. 2014. Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. *Nature Communications* 5. DOI: 10.1038/ncomms5340.

Qiu Q., Zhang G., Ma T., Qian W., Wang J., Ye Z., Cao C., Hu Q., Kim J., Larkin DM., Auvil L., Capitanu B., Ma J., Lewin HA., Qian X., Lang Y., Zhou R., Wang L., Wang K., Xia J., Liao S., Pan S., Lu X., Hou H., Wang Y., Zang X., Yin Y., Ma H., Zhang J., Wang Z., Zhang Y., Zhang D., Yonezawa T., Hasegawa M., Zhong Y., Liu W., Zhang Y., Huang Z., Zhang S., Long R., Yang H., Wang J., Lenstra JA., Cooper DN., Wu Y., Wang J., Shi P., Wang J., Liu J. 2012. The yak genome and adaptation to life at high altitude. *Nature Genetics* 44:946–949. DOI: 10.1038/ng.2343.

Rambaut A. 2009. FigTree, a graphical viewer of phylogenetic trees. *Institute of Evolutionary Biology University of Edinburgh*.

Reich PB., Wright IJ., Cavender-Bares J., Craine JM., Oleksyn J., Westoby M., Walters MB.

2003. The Evolution of Plant Functional Variation: Traits, Spectra, and Strategies. *International Journal of Plant Sciences* 164:S143–S164. DOI: 10.1086/374368.

Renner T., Specht CD. 2012. Molecular and functional evolution of class i chitinases for plant carnivory in the caryophyllales. *Molecular Biology and Evolution* 29:2971–2985. DOI: 10.1093/molbev/mss106.

Reynolds HL., Packer A., Bever JD., Clay K. 2003. Grassroots ecology: Plant-microbe-soil interactions as drivers of plant community structure and dynamics. *Ecology* 84:2281–2291. DOI: 10.1890/02-0298.

Roberts A., Pachter L. 2013. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods* 10:71–73. DOI: 10.1038/nmeth.2251.

Rogers WL., Cruse-Sanders JM., Determann R., Malmberg RL. 2010. Development and characterization of microsatellite markers in Sarracenia L. (pitcher plant) species. *Conservation Genetics Resources* 2:75–79. DOI: 10.1007/s12686-009-9165-x.

Rottloff S., Miguel S., Biteau F., Nisse E., Hammann P., Kuhn L., Chicher J., Bazile V., Gaume L., Mignard B., Hehn A., Bourgaud F. 2016. Proteome analysis of digestive fluids in Nepenthes pitchers. *Annals of botany* 117:479–495. DOI: 10.1093/aob/mcw001.

Roy SW., Gilbert W. 2006. The evolution of spliceosomal introns: Patterns, puzzles and progress. *Nature Reviews Genetics* 7:211–221. DOI: 10.1038/nrg1807.

Sangar V., Blankenberg DJ., Altman N., Lesk AM. 2007. Quantitative sequence-function relationships in proteins based on gene ontology. *BMC Bioinformatics* 8:294. DOI: 10.1186/1471-2105-8-294.

Sayers EW., Barrett T., Benson DA., Bryant SH., Canese K., Chetvernin. V., Church DM., Dicuccio M., Edgar R., Federhen S., Feolo M., Geer LY., Helmberg W., Kapustin Y., Landsman D., Lipman DJ., Madden TL., Maglott DR., Miller V., Mizrachi I., Ostell J., Pruitt KD., Schuler GD., Sequeira E., Sherry ST., Shumway M., Sirotkin K., Souvorov A., Starchenko G., Tatusova TA., Wagner L., Yaschenko E., Ye J. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 37. DOI: 10.1093/nar/gkn741.

Scamardella JM. 1999. Not plants or animals: A brief history of the origin of kingdoms protozoa, protista and protoctista. *International Microbiology* 2:207–216. DOI: 10.2436/im.v2i4.9219.

Schatz MC., Witkowski J., McCombie WR. 2012. Current challenges in de novo plant genome sequencing and assembly. *Genome Biology* 13:243. DOI: 10.1186/gb-2012-13-4-243.

Scherzer S., Krol E., Kreuzer I., Kruse J., Karl F., Von Rüden M., Escalante-Perez M., Müller T., Rennenberg H., Al-Rasheid KAS., Neher E., Hedrich R. 2013. The dionaea muscipula ammonium channel DmAMT1 provides NH4+ uptake associated with venus flytrap's prey digestion. *Current Biology* 23:1649–1657. DOI: 10.1016/j.cub.2013.07.028.

Schulze WX., Sanggaard KW., Kreuzer I., Knudsen AD., Bemm F., Thøgersen IB., Bräutigam A., Thomsen LR., Schliesky S., Dyrlund TF., Escalante-Perez M., Becker D., Schultz J., Karring H., Weber A., Højrup P., Hedrich R., Enghild JJ. 2012. The Protein Composition of the Digestive Fluid from the Venus Flytrap Sheds Light on Prey Digestion Mechanisms. *Molecular & Cellular Proteomics* 11:1306–1319. DOI: 10.1074/mcp.M112.021006.

Seibold I., Helbig AJ. 1995. Evolutionary history of New and Old World vultures inferred from

nucleotide sequences of the mitochondrial cytochrome b gene. *Philosophical Transactions of the Royal Society of London B* 350:163–178. DOI: 10.2307/56332.

Sémon M., Wolfe KH. 2007. Consequences of genome duplication. *Current Opinion in Genetics and Development* 17:505–512. DOI: 10.1016/j.gde.2007.09.007.

Seth S., Chakravorty D., Dubey VK., Patra S. 2014. An insight into plant lipase research - Challenges encountered. *Protein Expression and Purification* 95:13–21. DOI: 10.1016/j.pep.2013.11.006.

Shi T., Huang H., Barker MS. 2010. Ancient genome duplications during the evolution of kiwifruit (Actinidia) and related Ericales. *Annals of Botany* 106:497–504. DOI: 10.1093/aob/mcq129.

Shubin N., Tabin C., Carroll S. 2009. Deep homology and the origins of evolutionary novelty. *Nature* 457:818–823. DOI: 10.1038/nature07891.

Soltis DE., Albert VA., Leebens-Mack J., Bell CD., Paterson AH., Zheng C., Sankoff D., Depamphilis CW., Wall PK., Soltis PS. 2009. Polyploidy and angiosperm diversification. *American journal of botany* 96:336–48. DOI: 10.3732/ajb.0800079.

Soltis DE., Smith SA., Cellinese N., Wurdack KJ., Tank DC., Brockington SF., Refulio-Rodriguez NF., Walker JB., Moore MJ., Carlsward BS., Bell CD., Latvis M., Crawley S., Black C., Diouf D., Xi Z., Rushworth CA., Gitzendanner MA., Sytsma KJ., Qiu Y-L., Hilu KW., Davis CC., Sanderson MJ., Beaman RS., Olmstead RG., Judd WS., Donoghue MJ., Soltis PS. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *American journal of botany* 98:704–30. DOI: 10.3732/ajb.1000404.

Srivastava A., Rogers WL., Breton CM., Cai L., Malmberg RL. 2011. Transcriptome analysis of Sarracenia, an insectivorous plant. *DNA Research* 18:253–261. DOI: 10.1093/dnares/dsr014.

Stace HM., Chapman AR., Lemson KL., Powell JM. 1997. Cytoevolution, phylogeny and taxonomy in epacridaceae. *Annals of Botany* 79:283–290. DOI: 10.1006/anbo.1996.0333.

Stacey `Gary., Koh S., Granger C., Becker JM. 2002. Peptide transport in plants. *Trends in Plant Science* 7:257–263. DOI: 10.1016/S1360-1385(02)02249-5.

Staňková H., Hastie AR., Chan S., Vrána J., Tulpová Z., Kubaláková M., Visendi P., Hayashi S., Luo M., Batley J., Edwards D., Doležel J., Šimková H. 2016. BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. *Plant biotechnology journal* 14:1523–1531. DOI: 10.1111/pbi.12513.

Stephens JD., Rogers WL., Heyduk K., Cruse-Sanders JM., Determann RO., Glenn TC., Malmberg RL. 2015. Resolving phylogenetic relationships of the recently radiated carnivorous plant genus Sarracenia using target enrichment. *Molecular Phylogenetics and Evolution* 85:76–87. DOI: 10.1016/j.ympev.2015.01.015.

Stökl J., Paulus H., Dafni A., Schulz C., Francke W., Ayasse M. 2005. Pollinator attracting odour signals in sexually deceptive orchids of the Ophrys fusca group. *Plant Systematics and Evolution* 254:105–120. DOI: 10.1007/s00606-005-0330-8.

Stökl J., Twele R., Erdmann DH., Francke W., Ayasse M. 2007. Comparison of the flower scent of the sexually deceptive orchid Ophrys iricolor and the female sex pheromone of its pollinator Andrena morio. *Chemoecology* 17:231–233. DOI: 10.1007/s00049-007-0383-y.

Storey JD. 2003. The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics* 31:2013–2035. DOI: 10.1214/aos/1074290335.

Storey JD., Tibshirani R. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* 100:9440–9445. DOI: 10.1073/PNAS.1530509100.

Storz JF. 2016. Causes of molecular convergence and parallelism in protein evolution. *Nature Reviews Genetics* 17:239–250. DOI: 10.1038/nrg.2016.11.

Swarbreck D., Wilks C., Lamesch P., Berardini TZ., Garcia-Hernandez M., Foerster H., Li D., Meyer T., Muller R., Ploetz L., Radenbaugh A., Singh S., Swing V., Tissier C., Zhang P., Huala E. 2008. The Arabidopsis Information Resource (TAIR): Gene structure and function annotation. *Nucleic Acids Research* 36. DOI: 10.1093/nar/gkm965.

Tarailo-Graovac M., Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*. DOI: 10.1002/0471250953.bi0410s25.

Thomas PD., Wood V., Mungall CJ., Lewis SE., Blake JA. 2012. On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: A short report. *PLoS Computational Biology* 8. DOI: 10.1371/journal.pcbi.1002386.

Upadhyay AK., Chacko AR., Gandhimathi A., Ghosh P., Harini K., Joseph AP., Joshi AG., Karpe SD., Kaushik S., Kuravadi N., Lingu CS., Mahita J., Malarini R., Malhotra S., Malini M., Mathew OK., Mutt E., Naika M., Nitish S., Pasha SN., Raghavender US., Rajamani A., Shilpa S., Shingate PN., Singh HR., Sukhwal A., Sunitha MS., Sumathi M., Ramaswamy S., Gowda M., Sowdhamini R. 2015. Genome sequencing of herb Tulsi (Ocimum tenuiflorum) unravels key genes behind its strong medicinal properties. *BMC Plant Biology*

15:212. DOI: 10.1186/s12870-015-0562-x.

Vandenbussche M., Theissen G., Van de Peer Y., Gerats T. 2003. Structural diversification and neo-functionalization during floral MADS-box gene evolution by C-terminal frameshift mutations. *Nucleic Acids Research* 31:4401–4409. DOI: 10.1093/nar/gkg642.

Vezzi A., Campanaro S., D'Angelo M., Simonato F., Vitulo N., Lauro FM., Cestaro A., Malacrida G., Simionati B., Cannata N., Romualdi C., Bartlett DH., Valle G. 2005. Life at depth: Photobacterium profundum genome sequence and expression analysis. *Science (New York, N.Y.)* 307:1459–61. DOI: 10.1126/science.1103341.

Vianello A., Petrussa E., Macrì F. 1994. ATP/ADP antiporter is involved in uncoupling of plant mitochondria induced by low concentrations of palmitate. *FEBS Letters* 347:239–242. DOI: 10.1016/0014-5793(94)00540-0.

Walker BJ., Abeel T., Shea T., Priest M., Abouelliel A., Sakthikumar S., Cuomo CA., Zeng Q., Wortman J., Young SK., Earl AM. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* 9. DOI: 10.1371/journal.pone.0112963.

Wang G., Li X., Wang Z. 2016. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Research* 44:D1087–D1093. DOI: 10.1093/nar/gkv1278.

Wheeler GL., Carstens BC. 2018. Evaluating the adaptive evolutionary convergence of carnivorous plant taxa through functional genomics. *PeerJ* 6:e4322. DOI: 10.7717/peerj.4322.

Wheeler DL., Church DM., Federhen S., Lash AE., Madden TL., Pontius JU., Schuler GD.,

Schriml LM., Sequeira E., Tatusova TA., Wagner L. 2003. Database resources of the national center for biotechnology. *Nucleic Acids Research* 31:28–33. DOI: 10.1093/nar/gkg033.

Wheelwright NT., Orians GH. 1982. Seed Dispersal by Animals: Contrasts with Pollen Dispersal, Problems of Terminology, and Constraints on Coevolution. *The American Naturalist* 119:402–413. DOI: 10.1086/283918.

Williams L., Miller A. 2001. Transporters Responsible for the Uptake and Partitioning of Nitrogenous Solutes. *Annu Rev Plant Physiol Plant Mol Biol* 52:659–688. DOI: 10.1146/annurev.arplant.52.1.659.

Wolfe AD. 2005. ISSR techniques for evolutionary biology. *Methods in Enzymology* 395:134–144. DOI: 10.1016/S0076-6879(05)95009-X.

Xiao L., Yang G., Zhang L., Yang X., Zhao S., Ji Z., Zhou Q., Hu M., Wang Y., Chen M., Xu Y., Jin H., Xiao X., Hu G., Bao F., Hu Y., Wan P., Li L., Deng X., Kuang T., Xiang C., Zhu J-K., Oliver MJ., He Y. 2015. The resurrection genome of *Boea hygrometrica* : A blueprint for survival of dehydration. *Proceedings of the National Academy of Sciences* 112:5833–5837. DOI: 10.1073/pnas.1505811112.

Yang Z., Wafula EK., Honaas LA., Zhang H., Das M., Fernandez-Aparicio M., Huang K., Bandaranayake PCG., Wu B., Der JP., Clarke CR., Ralph PE., Landherr L., Altman NS., Timko MP., Yoder JI., Westwood JH., DePamphilis CW. 2015. Comparative transcriptome analyses reveal core parasitism genes and suggest gene duplication and repurposing as sources of structural novelty. *Molecular Biology and Evolution* 32:767–790. DOI: 10.1093/molbev/msu343.

Ye C., Hill C., Koren S., Ruan J., Zhanshan., Ma., Yorke J a., Zimin A. 2015. DBG2OLC: Efficient Assembly of Large Genomes Using the Compressed Overlap Graph. *arXiv*:1410.2801v2.

Ye C., Ma ZS., Cannon CH., Pop M., Yu DW. 2011. SparseAssembler: de novo Assembly with the Sparse de Bruijn Graph. *Arxiv preprint arXiv11062603*.

Yuan Y., Bayer PE., Lee HT., Edwards D. 2017. RunBNG: A software package for BioNano genomic analysis on the command line. *Bioinformatics* 33:3107–3109. DOI: 10.1093/bioinformatics/btx366.

Zellmer AJ., Hanes MM., Hird SM., Carstens BC. 2012. Deep phylogeographic structure and environmental differentiation in the carnivorous plant sarracenia alata. *Systematic Biology* 61:763–777. DOI: 10.1093/sysbio/sys048.

Appendix A: Chapter 2 Supplemental Materials

Chapter 2 Supplemental Tables

Table 10: List of protein functions identified in past studies of carnivorous plants, as published, and the taxon in which they were identified.

Function	Identified In	Term Assigned
Acid Chitinase	*Nepenthes* pitcher fluid [1]	*chitinase activity*
Actin	*Dionaea muscipula* secretion [2]	*actin filament*
ADP/ATP Carrier	*Dionaea muscipula* secretion [2]	*ATP:ADP antiporter activity*
alpha-Galactosidase	*Nepenthes* pitcher fluid [1]	*alpha-galactosidase activity*
Alternative Oxidase 1A	*Utricularia gibba* trap [6]	*alternative oxidase activity*
AMT1	*Dionaea muscipula* [3]	*ammonium transmembrane transport*
Aspartyl Protease	*Dionaea muscipula* secretion [2]; *Nepenthes* pitcher fluid [1, 7]	*aspartic-type endopeptidase activity*
ATP Synthase	*Dionaea muscipula* secretion [2]	*ATPase activity*
ATPase	*Dionaea muscipula* secretion [2]	*ATPase activity*
beta-1,3-Glucanase	*Nepenthes* pitcher fluid [1]	*beta-glucanase activity*
beta-Galactosidase	*Nepenthes* pitcher fluid [1]	*beta-galactosidase activity*
Cationic peroxidase	*Nepenthes* pitcher fluid [1]	*peroxidase activity*
Chitinase	*Dionaea muscipula* secretion [2]; *Nepenthes* pitcher fluid [1]	*chitinase activity*
Cinnamyl Alcohol Dehydrogenase	*Dionaea muscipula* secretion [2]	*cinnamyl-alcohol dehydrogenase activity*
Cysteine Protease	*Dionaea muscipula* secretion [2]; *Utricularia gibba* trap [6]	*cysteine-type peptidase activity*
Elongation Factor	*Dionaea muscipula* secretion [2]	No Match
Embryogenesis Protein	*Dionaea muscipula* secretion [2]	No Match
Endonuclease	*Dionaea muscipula* secretion [2]	*endonuclease complex*
Formate Dehydrogenase	*Dionaea muscipula* secretion [2]	*formate dehydrogenase complex*
Fructose-bisphosphate Aldolase	*Dionaea muscipula* secretion [2]	*fructose-bisphosphate aldolase activity*
G3P Dehydrogenase	*Dionaea muscipula* secretion [2]	No Match
G-Factor Binding Protein	*Dionaea* muscipula secretion [2]	No Match
Glucanase	*Dionaea* muscipula secretion [2]; *Nepenthes* pitcher fluid [1]	No Match
Glucosidase	*Nepenthes* pitcher fluid [1]	*glucosidase complex*
Glutathione Transferase	*Dionaea muscipula* secretion [2]	*glutathione transferase activity*
GPI-Anchored Protein Precursor	*Dionaea muscipula* secretion [2]	No Match
G-Protein Suppressor	*Dionaea muscipula* secretion [2]	
Heat Shock Protein	*Dionaea muscipula* secretion [2]	heat shock protein activity
Histone Protein	*Dionaea muscipula* secretion [2]	No Match
HKT1 Sodium Channel	*Dionaea muscipula* traps [5]	*sodium ion transmembrane transporter activity*
Lipase	*Dionaea muscipula* secretion [2]; *Nepenthes* pitcher fluid [1]	*lipase activity*
Lipid Transfer Protein	*Dionaea muscipula* secretion [2]; *Utricularia gibba* trap [6]; *Nepenthes* pitcher fluid [1]	*lipid transport*
Methylammonium Transmembrane Channel	*Utricularia* gibba shoot [6]	*methylammonium channel activity*
Nucleotide phosphodiesterase	*Dionaea muscipula* secretion [2]; *Nepenthes* pitcher fluid [1]	*cyclic-nucleotide phosphodiesterase activity*
Osmotin-like Protein	*Dionaea muscipula* secretion [2]	*water channel activity*
Pathogenesis-related Protein	*Dionaea muscipula* secretion [2]	No Match
Peroxidase	*Dionaea muscipula* secretion [2]; *Utricularia gibba* shoot [6]; *Nepenthes* pitcher fluid [1]	*peroxidase activity*
Phosphatase	*Dionaea muscipula* secretion [2]	*phosphatase activity*
Phospholipase	*Dionaea muscipula* secretion [2]	*phospholipase activity*
Plasma membrane water channel	*Utricularia gibba* shoot [6]	*water channel activity*

Function	Identified In	Term Assigned
Polygalacturonase	*Utricularia gibba* shoot [6]	*polygalacturonase activity*
Polygalacturonase Inhibitor	*Dionaea muscipula* secretion [2]	*polygalacturonase inhibitor activity*
Protein homodimerization	*Utricularia gibba* shoot [6]	*protein homodimerization activity*
Protein phosphatase	*Nepenthes* pitcher fluid [1]	*phosphatase activity*
Protodermal Factor	*Dionaea muscipula* secretion [2]	No Match
Ribonuclease	*Dionaea muscipula* secretion [2]; *Utricularia gibba* trap [6]	*ribonuclease activity*

Function	Identified In	Term Assigned
Serine Carboxypeptidase	*Dionaea muscipula* secretion [2]; *Utricularia gibba* trap [6]; *Nepenthes* pitcher fluid [1]	*serine-type carboxypeptidase activity*
Stigma-specific Protein	*Dionaea muscipula* secretion [2]	No Match
Superoxide Dismutase	*Dionaea muscipula* secretion [2]	*superoxide dismutase activity*
Symplast	*Nepenthes* pitcher glands [4]	*symplast*
Thioglucosidase	*Dionaea muscipula* secretion [2]	*thioglucosidase activity*
Thiol Protease	*Utricularia* gibba trap [6]	No Match
Ubiquitin Extension Protein	*Dionaea muscipula* secretion [2]	No Match
Xylosidase	*Nepenthes* pitcher fluid [1]	*xylanase activity*

Table 11: List of protein functions identified in past studies of carnivorous plants, as published, and the taxon in which they were identified. *Where zones of the plant are given (eg., trap, fluid, or secretion), analysis was localized to that specific zone or accounted for differential expression. Where only the taxon is specified, analysis considered the whole plant. Information from: Rottloff et al., 2016 [1], Schulze et al., 2012 [2] Scherzer et al., 2013 [3], Owen et al., 1999 [4], Böhm et al., 2016 [5], Ibarra-Laclette et al., 2011 [6], and An, Fukusaki & Kobayashi, 2002 [7].*

Table 12: Calculation of adjustment parameters to correct for differential detection of functions between GenBank-annotated and BLAST-annotated samples.

Function	GenBank #	GenBank %*10	BLAST #	BLAST %*10	Adjustment
actin filament	11	0.40	4	0.19	2.15
alpha-galactosidase activity	4	0.15	0	0	∅
alternative oxidase activity	6	0.22	4	0.19	1.17
ammonium transmembrane transport	8	0.29	7	0.33	0.89
aspartic-type endopeptidase activity	17	0.62	93	4.37	0.14
ATP:ADP antiporter activity	5	0.18	6	0.28	0.65
ATPase activity	188	6.91	369	17.34	0.40
beta-galactosidase activity	19	0.70	12	0.56	1.24
beta-glucanase activity	0	0	0	0	∅
chitinase activity	15	0.55	29	1.36	0.40
cinnamyl-alcohol dehydrogenase activity	17	0.62	11	0.52	1.21
cyclic-nucleotide phosphodiesterase activity	1	0.04	0	0	∅
cysteine-type peptidase activity	64	2.35	122	5.73	0.41
endonuclease complex	0	0	0	0	∅
formate dehydrogenase complex	0	0	0	0	∅
fructose-bisphosphate aldolase activity	9	0.33	7	0.33	1.01
glucosidase complex	0	0	0	0	∅
glutathione transferase activity	46	1.69	22	1.03	1.64

Function	GenBank #	GenBank %*10	BLAST #	BLAST %*10	Adjustment
lipase activity	63	2.32	22	1.03	2.24
lipid transport	146	5.37	57	2.68	2.00
methylammonium channel activity	0	0	0	0	∅
peroxidase activity	113	4.15	73	3.43	1.21
phosphatase activity	31	1.14	42	1.97	0.58
phospholipase activity	11	0.40	6	0.28	1.43
polygalacturonase activity	71	2.61	70	3.29	0.79
polygalacturonase inhibitor activity	2	0.07	2	0.09	0.78
protein homodimerization activity	141	5.18	139	6.53	0.79
ribonuclease activity	25	0.92	25	1.17	0.78
serine-type carboxypeptidase activity	55	2.02	36	1.69	1.19
sodium ion transmembrane transporter activity	0	0	0	0	∅
superoxide dismutase activity	5	0.18	0	0	∅
symplast	0	0	0	0	∅
thioglucosidase activity	8	0.29	1	0.05	6.26
water channel activity	39	1.43	14	0.66	2.18
xylanase activity	0	0	0	0	∅
heat shock protein activity	16	0.59	9	0.42	1.39
None of the Above	26091	959.02	20126	945.86	-
Total Carnivorous	1115	40.98	1152	54.14	-
Total	27206	-	21278	-	1.28

Table 13: Calculation of adjustment parameters to correct for differential detection of functions between GenBank-annotated and BLAST-annotated samples. *The "Adjustment" column indicates the value that BLAST proportions are multiplied by to match the value distribution expected of their GenBank counterparts. Nullset symbols indicate functions for which one or both methods detected zero instances, preventing adjustment calculations. Hyphens indicate values that are calculated individually for each sample, using the sum of other post-adjustment values.*

	CFOL	DCAP	GAUR	UGIB	ACHI	ATHA	BHYG	GSOJ	OSAT	OTEN	Averag
ATPase activity	10.50	17.79	15.09%	29.79	22.76	16.73	27.59	22.82	24.54	24.66	21.23%
peroxidase activity	17.70	15.65	11.32%	11.08	10.32	10.05	14.25	19.88	27.59	8.83%	14.67%
aspartic-type endopeptidase activity	2.07%	14.92	6.37%	17.22	21.88	1.51%	14.48	14.35	1.42%	23.09	11.73%
lipid transport	2.44%	6.79%	4.25%	10.03	4.72%	5.69%	7.59%	9.18%	10.75	7.22%	6.87%
cysteine-type peptidase activity	14.34	9.04%	10.14%	1.80%	3.91%	12.99	5.29%	3.48%	5.07%	1.51%	6.76%
polygalacturonase activity	4.83%	4.17%	6.37%	6.14%	5.22%	6.32%	9.43%	9.36%	6.29%	5.59%	6.37%
protein homodimerization activity	6.28%	5.61%	16.98%	4.04%	5.46%	12.54	0.69%	0.62%	1.42%	5.99%	5.96%
serine-type carboxypeptidase activity	8.01%	3.05%	4.95%	1.95%	2.33%	4.89%	6.21%	7.40%	7.10%	2.71%	4.86%
phosphatase activity	3.78%	5.03%	1.65%	4.94%	3.80%	2.76%	3.91%	4.63%	1.83%	3.49%	3.58%
lipase activity	5.80%	2.36%	1.65%	0.90%	2.76%	5.60%	0.23%	1.34%	1.22%	1.13%	2.30%
ammonium transmembrane transport	5.48%	1.72%	1.65%	0.00%	3.36%	4.09%	0.00%	0.09%	0.20%	5.76%	2.24%
beta-galactosidase activity	4.90%	2.23%	3.77%	1.35%	1.40%	1.69%	0.23%	0.18%	2.84%	2.18%	2.08%
glutathione transferase activity	0.68%	2.82%	0.94%	3.29%	3.95%	0.71%	1.84%	1.07%	1.83%	0.67%	1.78%
water channel activity	0.74%	0.79%	1.89%	0.90%	1.16%	1.33%	2.99%	1.96%	3.45%	1.42%	1.66%
chitinase activity	4.98%	2.29%	2.36%	1.20%	2.01%	3.47%	0.00%	0.00%	0.00%	0.00%	1.63%
Other	**7.45%**	**5.73%**	**10.61**	**5.39%**	**4.96%**	**9.61%**	**5.29%**	**3.65%**	**4.46%**	**5.76%**	6.29%
heat shock protein activity	1.69%	0.84%	2.12%	0.90%	1.57%	1.42%	2.30%	1.25%	0.00%	1.75%	1.38%
fructose-bisphosphate aldolase activity	0.92%	0.91%	2.12%	1.05%	1.14%	0.80%	1.38%	1.43%	1.22%	0.51%	1.15%
ribonuclease activity	0.83%	1.18%	0.94%	0.75%	0.96%	2.22%	1.61%	0.62%	1.42%	0.98%	1.15%
phospholipase activity	1.53%	0.43%	1.65%	0.90%	0.00%	0.98%	0.00%	0.00%	0.00%	0.36%	0.58%
actin filament	0.98%	0.00%	0.71%	0.30%	0.22%	0.98%	0.00%	0.00%	1.42%	0.00%	0.46%
alternative oxidase activity	0.18%	0.18%	1.42%	0.45%	0.25%	1.51%	0.00%	0.00%	0.00%	0.91%	0.49%
ATP:ADP antiporter activity	0.60%	0.78%	0.71%	0.45%	0.33%	0.44%	0.00%	0.36%	0.41%	0.66%	0.47%
cinnamyl-alcohol dehydrogenase	0.71%	1.41%	0.47%	0.60%	0.48%	0.53%	0.00%	0.00%	0.00%	0.59%	0.48%
thioglucosidase activity	0.00%	0.00%	0.47%	0.00%	0.00%	0.71%	0.00%	0.00%	0.00%	0.00%	0.12%

Table 14: Representation of each carnivory-associated function proportion to the total of all carnivory-associated functions, as depicted graphically in Figure 5. *Functions are ordered by their representation on average (given in far-right column), from most common to most rare. "Other" indicates the total of the nine rarest carnivory-associated functions, listed with indentation. [Abbreviation key: CFOL – Cephalotus follicularis; DCAP – Drosera capensis; GAUR – Genlisea aurea; UGIB – Utricularia gibba; ACHI – Actinidia chinensis; ATHA – Arabidopsis thaliana; BHYG – Boea hygrometrica; GSOJ – Glycine soja; OSAT – Oryza sativa; OTEN – Ocimum tenuiflorum].*

	t	p	q	Sig.
Actin	-0.17	0.565	0.149	*NS*
AltOx	**3.34**	**0.007**	**0.033**	*
AspPep	-0.29	0.609	0.152	*NS*
ATP	0.05	0.479	0.144	*NS*
ATP_ADP	**2.91**	**0.013**	**0.033**	*
BGal	1.94	0.054	0.068	.
Chit	-2.27	0.972	0.194	*NS*
CinAlc	0.40	0.351	0.144	*NS*
CystPep	0.08	0.469	0.144	*NS*
FrucBPA	0.65	0.269	0.135	*NS*
GlutTrans	0.10	0.462	0.144	*NS*
H2OChan	1.42	0.098	0.082	.
HeatShock	0.14	0.448	0.144	*NS*
Lipase	-0.12	0.546	0.149	*NS*
LipTrans	0.23	0.413	0.144	*NS*
NHTrans	0.67	0.270	0.135	*NS*
Perox	-0.39	0.647	0.154	*NS*
Phoslip	**2.57**	**0.022**	**0.037**	*
Phosp	1.25	0.142	0.089	.
Polygal	-0.67	0.739	0.161	*NS*
ProtHomo	1.47	0.091	0.082	.
RiboNuc	-0.89	0.796	0.166	*NS*
SerCarPep	-0.54	0.696	0.158	*NS*
ThioGluc	1.34	0.129	0.089	.
Total	0.03	0.489	0.144	*NS*

Table 15: Results of statistical analyses comparing non-carnivorous plants to carnivorous plants for each of 24 carnivory-associated functions, plus the total of all functions. *Equivalent to main text Table 3, but using unadjusted data. "t" indicates the test statistic of an upper-tailed Student's t-test. "p" indicates the p-value of this test. "q" indicates a corrected p-value accounting for multiple comparisons, using Storey's correction. Significance ("Sig.") is indicated by bolding and with "*" for q < 0.05, "**" for q < 0.01, and "*** for q < 0.001. A non-bolded "." indicates marginal values (q < 0.10), while "NS" indicates non-significance (q > 0.10).*

	Genlisea aurea				*Drosera capensis*				*Utricularia gibba*				*Cephalotus follicularis*			
	Z	**p**	**q**	**Sig**	**Z**	**p**	**q**	**Sig**	**Z**	**p**	**q**	**Sig**	**Z**	**p**	**q**	**Sig.**
Actin	0.26	0.396	0.66	*NS*	-	0.746	0.554	*NS*	-	0.517	0.152	*NS*	0.13	0.448	0.189	*NS*
AltOx	0.85	0.198	0.47	*NS*	**3.24**	**5.98E-**	**9.56E-**	**	2.53	**0.006**	**0.012**	*	**2.04**	**0.021**	**0.023**	*
AspPep	-	0.760	0.79	*NS*	0.91	0.181	0.353	*NS*	-	0.635	0.164	*NS*	-	0.686	0.199	*NS*
ATP	-	0.978	0.79	*NS*	0.85	0.199	0.353	*NS*	1.31	0.094	0.079	.	0.05	0.482	0.189	*NS*
ATP_AD	1.03	0.150	0.47	*NS*	2.80	**2.55E-**	**0.016**	*	1.05	0.148	0.106	*NS*	**3.06**	**0.001**	**0.002**	**
BGal	1.97	0.024	0.16	*NS*	0.28	0.390	0.554	*NS*	0.62	0.269	0.142	*NS*	**3.19**	**7.19E-**	**2.15E-**	**
Chit	-	0.860	0.79	*NS*	-	0.865	0.554	*NS*	-	0.946	0.189	*NS*	-	0.682	0.199	*NS*
CinAlc	1.26	0.104	0.41	*NS*	-	0.653	0.554	*NS*	0.34	0.366	0.152	*NS*	-	0.625	0.198	*NS*
CystPep	-	0.996	0.79	*NS*	2.13	0.017	0.053	.	2.16	0.015	0.015	*	-	0.880	0.235	*NS*
FrucBPA	1.56	0.059	0.29	*NS*	-	0.741	0.554	*NS*	0.75	0.227	0.142	*NS*	-	0.512	0.189	*NS*
GlutTrans	0.00	0.500	0.76	*NS*	-	0.623	0.554	*NS*	-	0.789	0.164	*NS*	1.36	0.087	0.083	.
H2OChan	0.83	0.203	0.47	*NS*	0.08	0.470	0.554	*NS*	0.57	0.284	0.142	*NS*	1.20	0.114	0.095	.
HeatShoc	0.78	0.217	0.47	*NS*	-	0.792	0.554	*NS*	0.06	0.477	0.152	*NS*	0.26	0.398	0.189	*NS*
Lipase	-	0.556	0.79	*NS*	-	0.651	0.554	*NS*	-	0.591	0.164	*NS*	0.54	0.295	0.164	*NS*
LipTrans	0.71	0.238	0.47	*NS*	-	0.616	0.554	*NS*	-	0.709	0.164	*NS*	0.60	0.274	0.164	*NS*
NHTrans	-	0.792	0.79	*NS*	1.12	0.130	0.298	*NS*	3.18	**7.40E-**	**3.70E-**	**	-	0.773	0.215	*NS*
Perox	-	0.700	0.79	*NS*	-	0.667	0.554	*NS*	0.04	0.483	0.152	*NS*	0.21	0.416	0.189	*NS*
Phoslip	2.78	2.69E-	0.05	.	0.09	0.464	0.554	*NS*	2.33	**9.95E-**	**0.012**	*	**2.32**	**1.03E-**	**0.014**	*
Phosp	-	0.975	0.79	*NS*	2.76	0.003	0.016	*	2.40	**0.008**	**0.012**	*	**3.01**	**1.29E-**	**2.15E-**	**
Polygal	-	0.740	0.79	*NS*	-	0.859	0.554	*NS*	0.42	0.337	0.152	*NS*	-	0.539	0.189	*NS*
ProtHomo	2.03	0.021	0.16	*NS*	0.28	0.390	0.554	*NS*	0.22	0.413	0.152	*NS*	0.92	0.179	0.132	*NS*
RiboNuc	-	0.777	0.79	*NS*	-	0.557	0.554	*NS*	-	0.678	0.164	*NS*	-	0.568	0.189	*NS*
SerCarPe	-	0.609	0.79	*NS*	-	0.838	0.554	*NS*	-	0.790	0.164	*NS*	0.83	0.203	0.136	*NS*
ThioGluc	0.40	0.345	0.62	*NS*	2.38	8.63E-	0.035	*	-	0.771	0.164	*NS*	**4.63**	**1.82E-**	**1.21E-**	***
Total	-	0.889	0.79	*NS*	1.25	0.106	0.283	*NS*	0.02	0.490	0.152	*NS*	0.02	0.491	0.189	*NS*

Table 16: Results of statistical analyses comparing non-carnivorous plants to carnivorous plants for each of 24 carnivory-associated functions, plus the total of all functions. *Equivalent to main text Table 4 but using unadjusted data. "t" indicates the test statistic of an upper-tailed Student's t-test. "p" indicates the p-value of this test. "q" indicates a corrected p-value accounting for multiple comparisons, using Storey's correction. Significance ("Sig.") is indicated by bolding and with "*" for q < 0.05, "**" for q < 0.01, and "*** for q < 0.001. A non-bolded "." indicates marginal values (q < 0.10), while "NS" indicates non-significance (q > 0.10).*

Chapter 2 Supplemental Figures

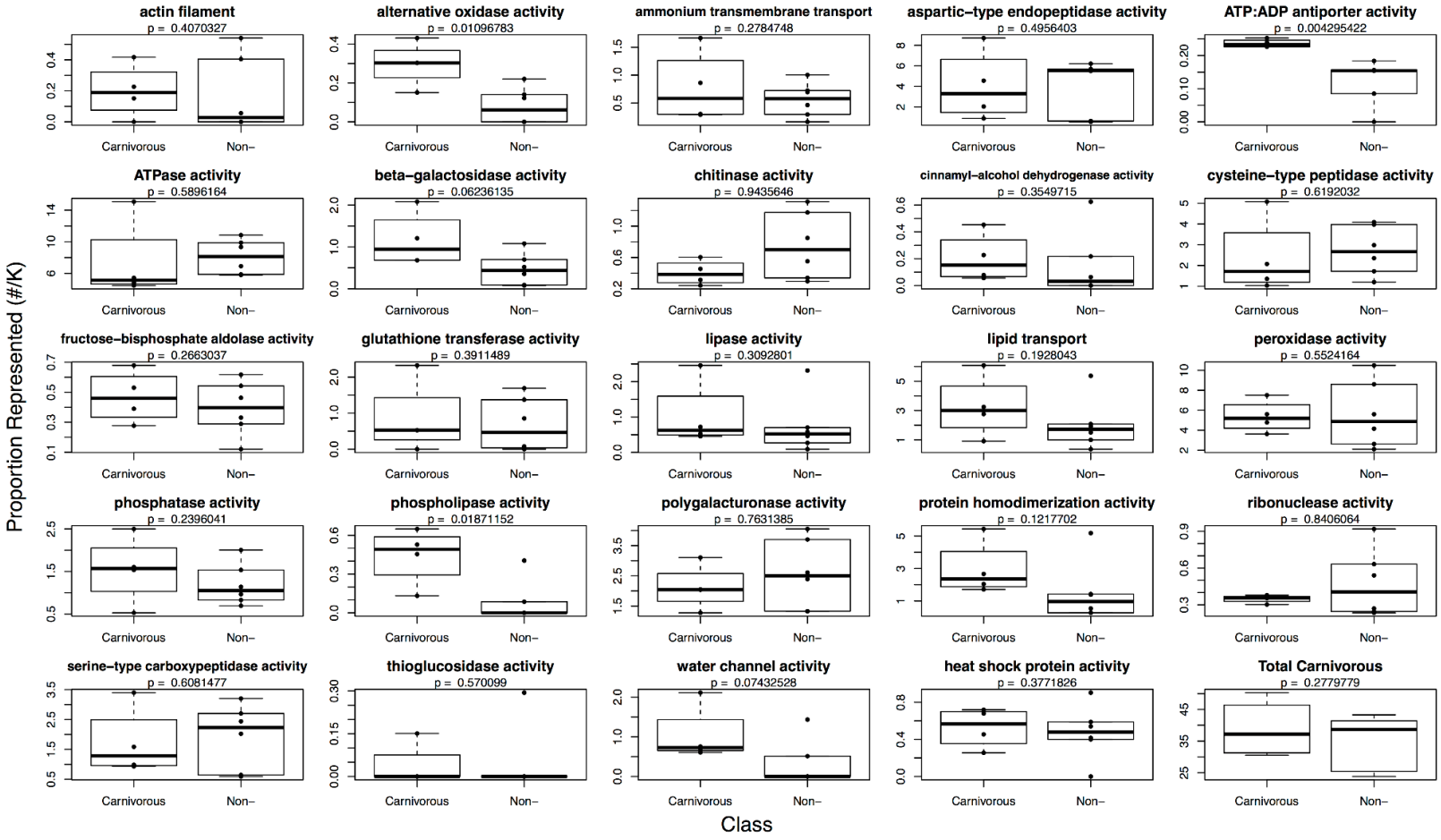

Figure 12: Graphical depiction of data presented in Table 3. *Each boxplot depicts one of 24 comparisons between the relative proportion of a carnivory-associated function in carnivorous vs. non-carnivorous plants, plus the sum of all these functions. Dots show the position and effects of individual samples within each distribution.*

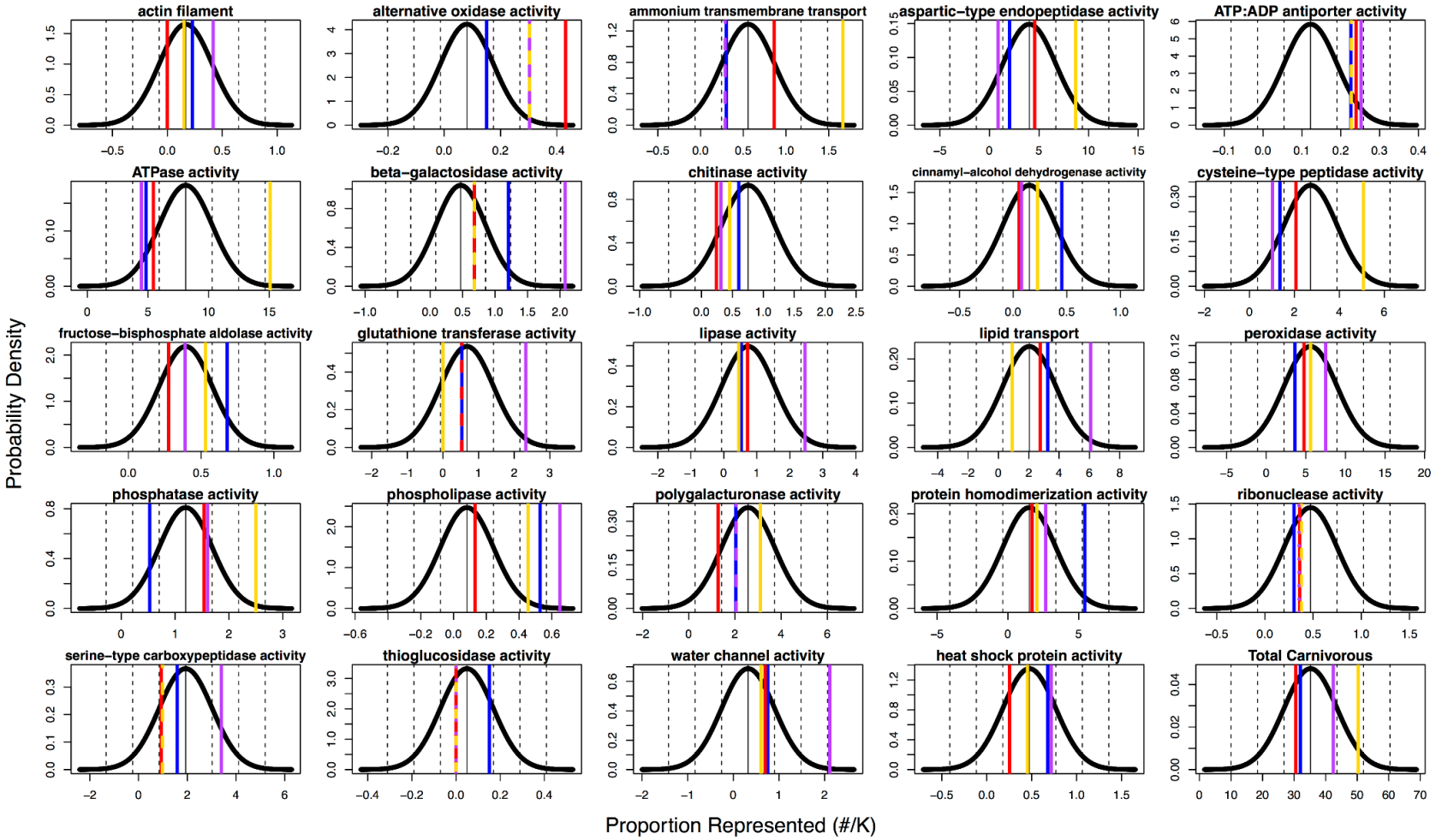

Figure 13: Graphical depiction of data presented in Table 4. *Each normal distribution represents the range of values found in non-carnivorous taxa for a function analyzed. Colored lines each indicate the value found for a carnivorous taxon (Blue: Genlisea aurea; Red: Drosera capensis; Yellow: Utricularia gibba; Purple: Cephalotus follicularis). Black vertical lines indicate mean (solid) and standard deviations from the mean (dashed).*

Appendix B: Chapter 3 Supplemental Materials

Chapter 3 Supplemental Tables

Metrics	SPAdes	SparseAssembler	DBG2OLC	Canu	Arrow	BioNano	Pilon
Coverage							
Total Sequence	1.54 Gbp	0.242 Gbp	0.767 Gbp	3.16 Gbp	3.16 Gbp	0.760 Gbp	3.16 Gbp
Genome Coverage	42.91%	6.75%	21.01%	87.76%	87.86%	2.11%	87.83%
Contiguity							
Contigs	1,520,620	941,462	87,884	122,036	122,036	306	122,036
Mean Length	1,015.9	257.9	8,608.1	25,889.1	25,910.9	248,239.6	25,910.1
N50	2,132.0	245.0	10,411.0	35,615.0	35,641.5	223,097.0	35,649.5
Length Quartiles							
Maximum	386,165	46,192	66,545	3,099,186	3,099,149	3,099,149	3,099,149
Q3	1,000	274	11,127	30,965	30,993	245,733	30,990
Median	430	235	7,840	17,603	17,621	208,119	17,619
Q1	290	214	4,970	11,140	11,150	189,439	11,150
Minimum	200*	200*	500*	1,000*	1,000	164,181	1,000
# Large							
>100kb	9	0	0	2,631	2,638	305	2,640
>250kb	1	0	0	73	73	72	73
>500kb	0	0	0	7	7	7	7
>1Mb	0	0	0	2	2	2	2

Table 17: Assembly metrics for* Sarracenia alata *genome for each pipeline stage. *Genome coverage is calculated based on an estimated haploid size of 3.60 Gb. For minimum size, values marked with an asterisk indicate the minimum value was produced by size filtering at this step. All assembly steps and their relation to one another can be viewed in Figure 6.*

Table 18: Genome annotation GO hits.

GO	Count	Definition
GO:0005524	2408	ATP binding
GO:0000166	2208	nucleotide binding
GO:0046872	2040	metal ion binding
GO:0016740	1962	transferase activity
GO:0016491	1814	oxidoreductase activity
GO:0016787	1725	hydrolase activity
GO:0003677	1406	DNA binding
GO:0008270	1398	zinc ion binding
GO:0003676	1368	nucleic acid binding
GO:0003824	1259	catalytic activity
GO:0016301	826	kinase activity
GO:0004672	742	protein kinase activity
GO:0004674	540	protein serine/threonine kinase activity
GO:0003700	524	transcription factor activity, sequence-specific DNA binding
GO:0003723	521	RNA binding
GO:0003735	457	structural constituent of ribosome
GO:0016874	434	ligase activity
GO:0016887	424	ATPase activity
GO:0016829	405	lyase activity
GO:0005215	399	transporter activity
GO:0016757	353	transferase activity, transferring glycosyl groups
GO:0008233	347	peptidase activity
GO:0008168	334	methyltransferase activity
GO:0005506	319	iron ion binding
GO:0020037	316	heme binding
GO:0046983	306	protein dimerization activity
GO:0043565	301	sequence-specific DNA binding
GO:0016853	284	isomerase activity
GO:0005525	278	GTP binding
GO:0000287	245	magnesium ion binding
GO:0016746	239	transferase activity, transferring acyl groups
GO:0004497	222	monooxygenase activity
GO:0016798	215	hydrolase activity, acting on glycosyl bonds
GO:0005509	213	calcium ion binding
GO:0004386	206	helicase activity
GO:0016705	201	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
GO:0009055	189	electron carrier activity
GO:0050660	172	flavin adenine dinucleotide binding
GO:0004553	171	hydrolase activity, hydrolyzing O-glycosyl compounds
GO:0051536	167	iron-sulfur cluster binding
GO:0003924	165	GTPase activity
GO:0016779	154	nucleotidyltransferase activity
GO:0030170	153	pyridoxal phosphate binding
GO:0019843	131	rRNA binding
GO:0016758	125	transferase activity, transferring hexosyl groups
GO:0032440	124	2-alkenal reductase [NAD(P)] activity
GO:0022857	124	transmembrane transporter activity
GO:0051287	124	NAD binding
GO:0004812	121	aminoacyl-tRNA ligase activity
GO:0016747	117	transferase activity, transferring acyl groups other than amino-acyl groups
GO:0016620	116	oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor

GO	Count	Definition
GO:0051082	116	unfolded protein binding
GO:0004842	114	ubiquitin-protein transferase activity
GO:0004252	108	serine-type endopeptidase activity
GO:0008483	105	transaminase activity
GO:0004871	104	signal transducer activity
GO:0042626	104	ATPase activity, coupled to transmembrane movement of substances
GO:0005507	103	copper ion binding
GO:0051539	101	4 iron, 4 sulfur cluster binding
GO:0051213	99	dioxygenase activity
GO:0003743	97	translation initiation factor activity
GO:0004601	96	peroxidase activity
GO:0000155	94	phosphorelay sensor kinase activity
GO:0003674	87	molecular_function
GO:0016788	87	hydrolase activity, acting on ester bonds
GO:0008137	86	NADH dehydrogenase (ubiquinone) activity
GO:0046982	86	protein heterodimerization activity
GO:0030246	85	carbohydrate binding
GO:0004190	82	aspartic-type endopeptidase activity
GO:0008017	82	microtubule binding
GO:0004527	81	exonuclease activity
GO:0010181	80	FMN binding
GO:0061630	79	ubiquitin protein ligase activity
GO:0005198	78	structural molecule activity
GO:0004721	78	phosphoprotein phosphatase activity
GO:0016597	78	amino acid binding
GO:0016614	76	oxidoreductase activity, acting on CH-OH group of donors
GO:0016616	76	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor
GO:0003755	76	peptidyl-prolyl cis-trans isomerase activity
GO:0008289	71	lipid binding
GO:0003899	71	DNA-directed 5'-3' RNA polymerase activity
GO:0015035	70	protein disulfide oxidoreductase activity
GO:0000977	70	RNA polymerase II regulatory region sequence-specific DNA binding
GO:0016820	70	hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances
GO:0008237	69	metallopeptidase activity
GO:0003746	66	translation elongation factor activity
GO:0050661	66	NADP binding
GO:0008236	66	serine-type peptidase activity
GO:0022891	66	substrate-specific transmembrane transporter activity
GO:0003678	66	DNA helicase activity
GO:0016772	65	transferase activity, transferring phosphorus-containing groups
GO:0015297	62	antiporter activity

GO	Count	Definition
GO:0005516	62	calmodulin binding
GO:0004222	62	metallo-endopeptidase activity
GO:0003779	61	actin binding
GO:0050662	61	coenzyme binding
GO:0005515	60	protein binding
GO:0048038	60	quinone binding
GO:0003777	60	microtubule motor activity
GO:0003690	58	double-stranded DNA binding
GO:0004004	58	ATP-dependent RNA helicase activity
GO:0008757	57	S-adenosylmethionine-dependent methyltransferase activity
GO:0000049	56	tRNA binding
GO:0004518	54	nuclease activity
GO:0008080	53	N-acetyltransferase activity
GO:0016627	51	oxidoreductase activity, acting on the CH-CH group of donors
GO:0016709	51	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, NAD(P)H as one donor, and incorporation of one atom of oxygen
GO:0008565	50	protein transporter activity
GO:0042803	50	protein homodimerization activity
GO:0080044	50	quercetin 7-O-glucosyltransferase activity
GO:0003682	50	chromatin binding
GO:0004519	50	endonuclease activity
GO:0080043	50	quercetin 3-O-glucosyltransferase activity

Table 19: Genome annotation GO hits. *Table shows GO code, number of hits to the* S. alata *genome, and the term's definition. List is limited to Molecular Function-class terms, with 50 hits or more.*

Table 20: Classes of identified elements found in the *Sarracenia alata* genome.

Class	Count	Length (bp)
Protein-coding Elements	**122,610**	**168,191,093**
Gene	**28,750**	**73,603,156**
Exon	**66,068**	**21,056,261**
mRNA	**27,792**	**73,531,676**
Pseudogenes	**36,979**	**53,043,265**
Functional RNAs	**1,570**	**370,453**
rRNA	**609**	**297,509**
tRNA	**961**	**72,944**
Repetitive Elements	**4,189,649**	**2,853,951,657**
DNA Transposons	**171,790**	**77,756,312**
CMC-Chapaev	90	12,561
CMC-EnSpm	21,144	8,909,628
CMC-Transib	1,028	324,463
MULE-MuDR	25,669	9,995,734
Maverick	3,634	2,723,827
PIF-Harbinger	5,441	3,326,128
TcMar-Tc2	610	916,859
TcMar-Tc4	3,305	2,425,762
TcMar (Uncategorized)	2,269	550,856
hAT-Ac	55,496	14,894,539
hAT-Charlie	1	82
hAT-Tag1	11,308	4,255,466
hAT-Tip100	9,394	3,868,436
hAT (Uncategorized)	32,284	17,227,016
LINE	**55,837**	**34,390,477**
LINE 1	1,123	180,701
I-Jockey	4,945	10,299,290
L1	38,617	20,717,653
L2	2,301	318,541
RTE-BovB	8,851	2,874,292
SINE	**9,447**	**1,389,567**
LTR	**1,166,601**	**1,276,836,130**
Cassandra	17,413	2,282,226
Caulimovirus	1,216	1,193,290
Copia	590,377	732,124,004
ERV1	11,637	12,930,050
ERVK	7,759	1,948,807

Class	Count	Length (bp)
Gypsy	538,196	526,355,990
Helitron	**6,679**	**3,475,420**
Retroposon	**18,124**	**3,883,619**
Satellite	**7,687**	**963,160**
Simple Repeat	**1,136,401**	**62,811,887**
Low Complexity	**222,975**	**11,788,884**
Artefacts	30	10,180
Bacterial Insertion	29	9,954
IS2	3	3,769
IS3	19	3,528
IS5	7	2,657
Tn1000	1	226

Table 21: Classes of identified elements found in the *Sarracenia alata* genome. Categories determined from MAKER-P annotation output GFF, summarized using the "SummarizeAnnotation" Perl script. Black headings indicate large general classifications, while grey indicates family and white indicates sub-type. Family levels show element and base pair counts higher than their constituent sub-types due to elements that could not be more specifically classified

Chapter 3 Supplemental Figures

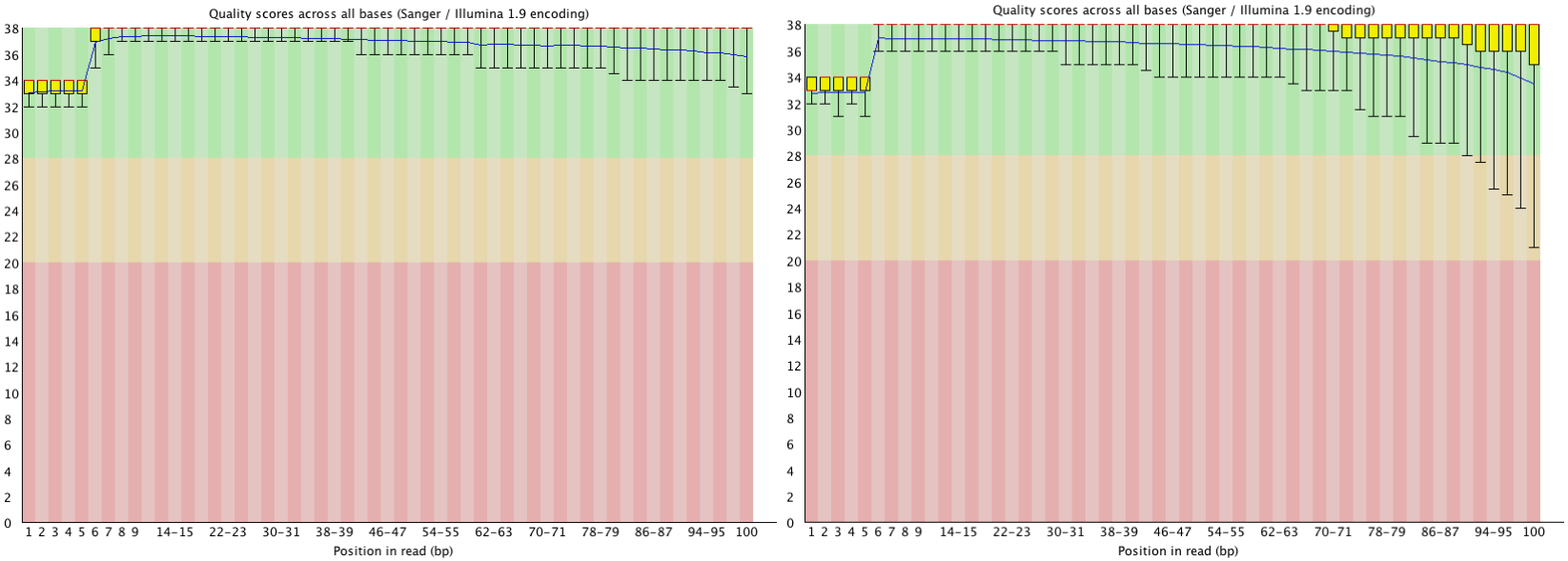

Figure 14: Illumina per-site sequence quality. *"Forward" paired-end reads left; "Reverse" paired-end reads right. Each box-and-whisker plot shows the Phred score distribution for a particular site (1-100) in a read*

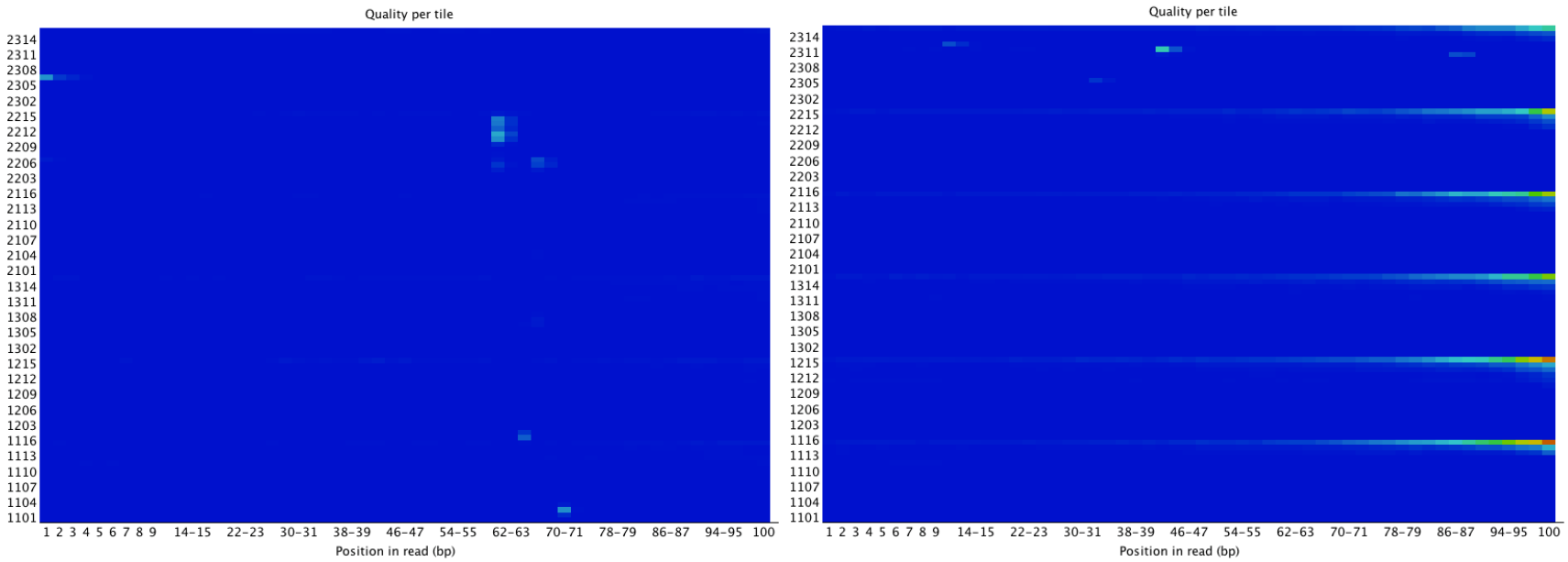

Figure 15: Illumina per-tile sequence quality. *"Forward" paired-end reads left; "Reverse" paired-end reads right. Reverse reads show some quality issues towards the end of sequences; however, these are typical of reverse reads and are insufficient to result in a failure of the test.*

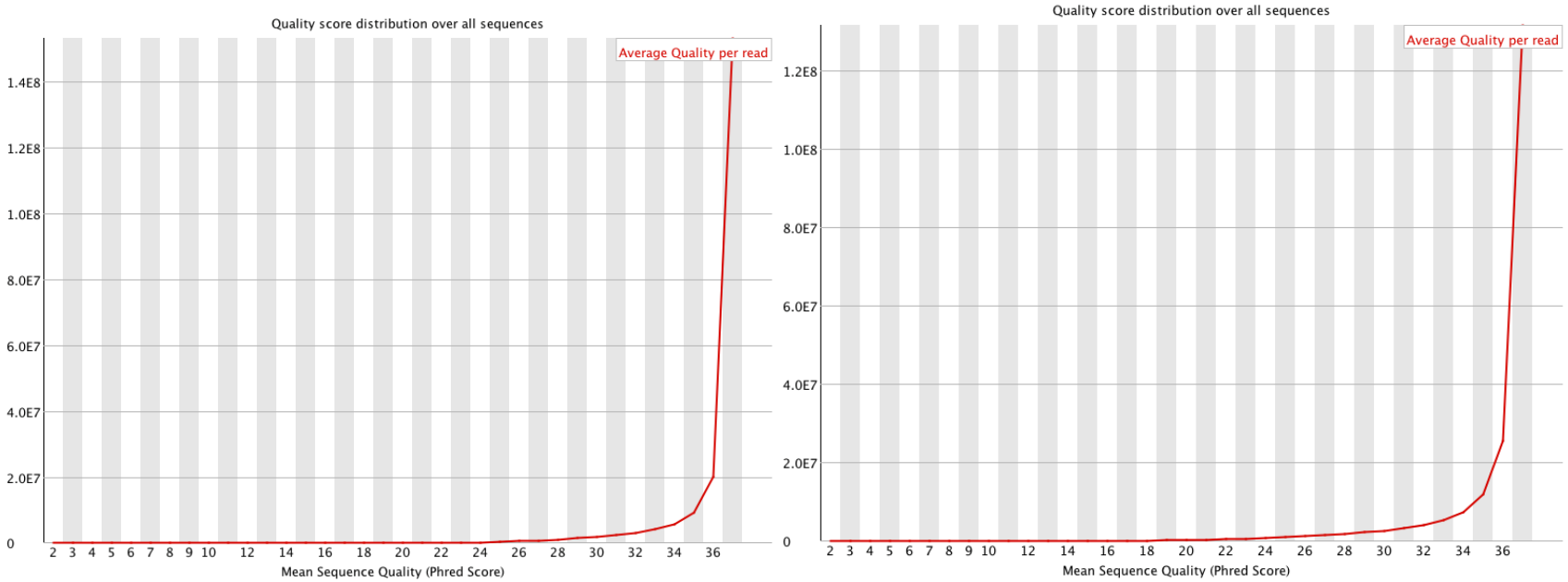

Figure 16: Illumina per-sequence quality score distribution. *"Forward" paired-end reads left; "Reverse" paired-end reads right. For both sets of reads, the vast majority of sequences possessed mean Phred scores of 37 or higher.*

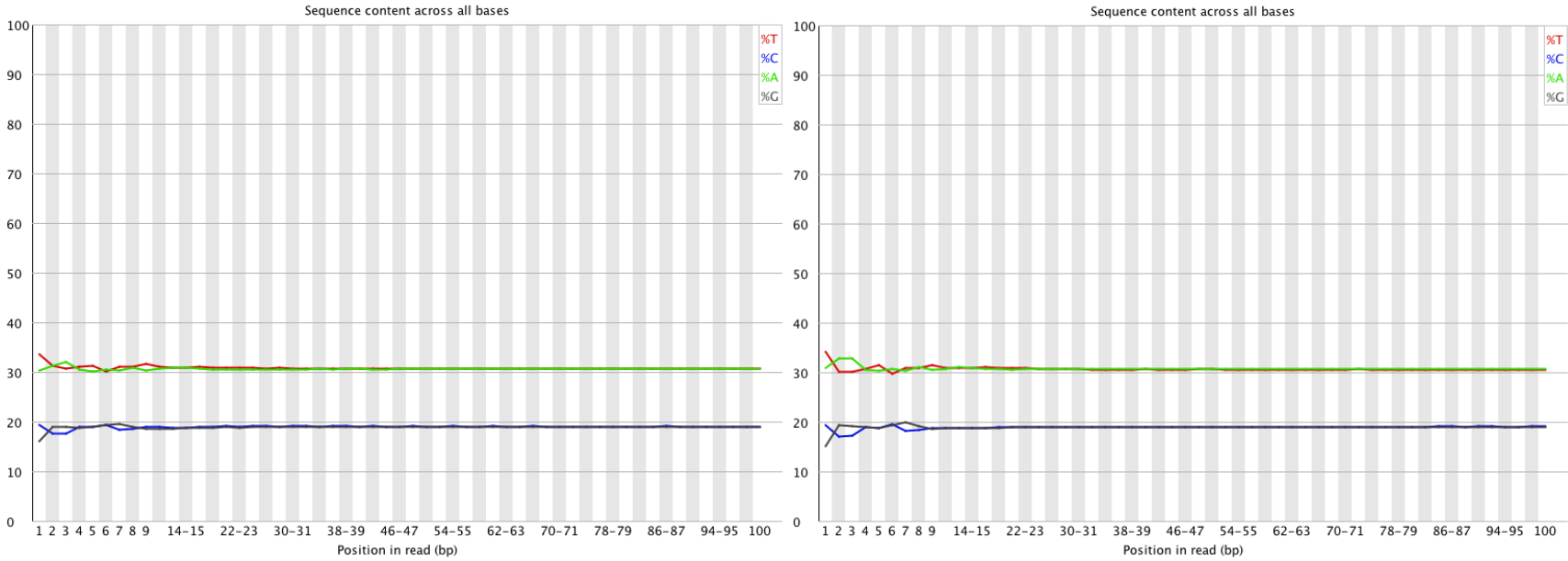

Figure 17: Illumina per-site base composition. *"Forward" paired-end reads left; "Reverse" paired-end reads right. Slight irregularity at the start of sequence is typical, with the effect somewhat exaggerated due to the scale of position 1-9 compared to the remaining sequence.*

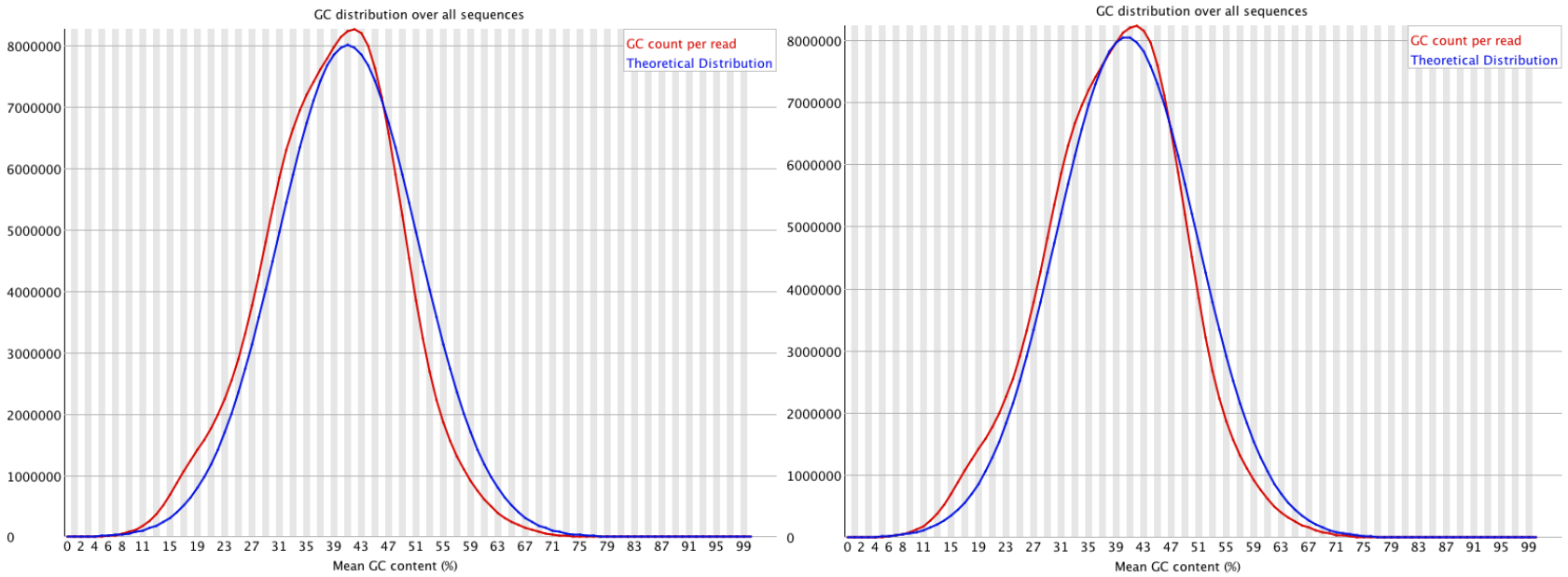

Figure 18: Illumina sequence GC content. *"Forward" paired-end reads left; "Reverse" paired-end reads right. The experimental distribution of sequence GC content (blue) closely approximates the theoretical distribution (red).*

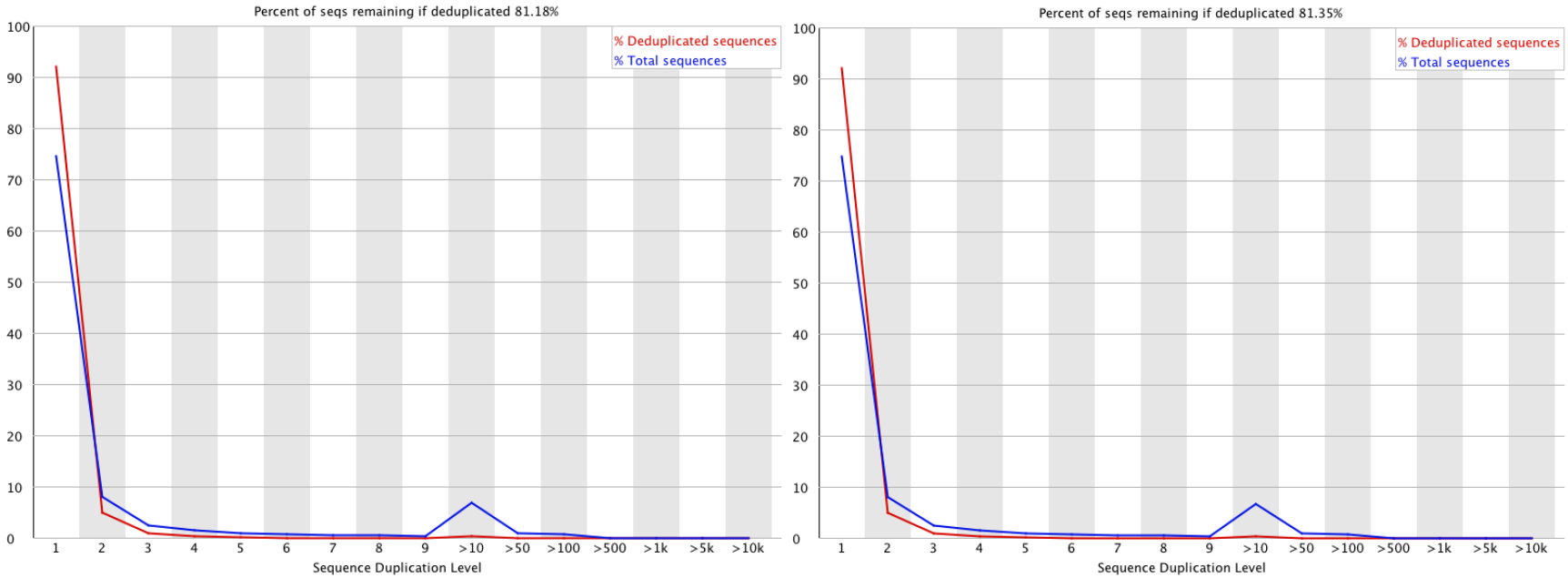

Figure 19: Illumina overrepresented sequences. *"Forward" paired-end reads left; "Reverse" paired-end reads right. The values detected for both forward and reverse sequences (blue) are a close fit to the theoretical distribution (red), only deviating at the ">10" level. This is likely due to repetitive elements common in plant genomes.*

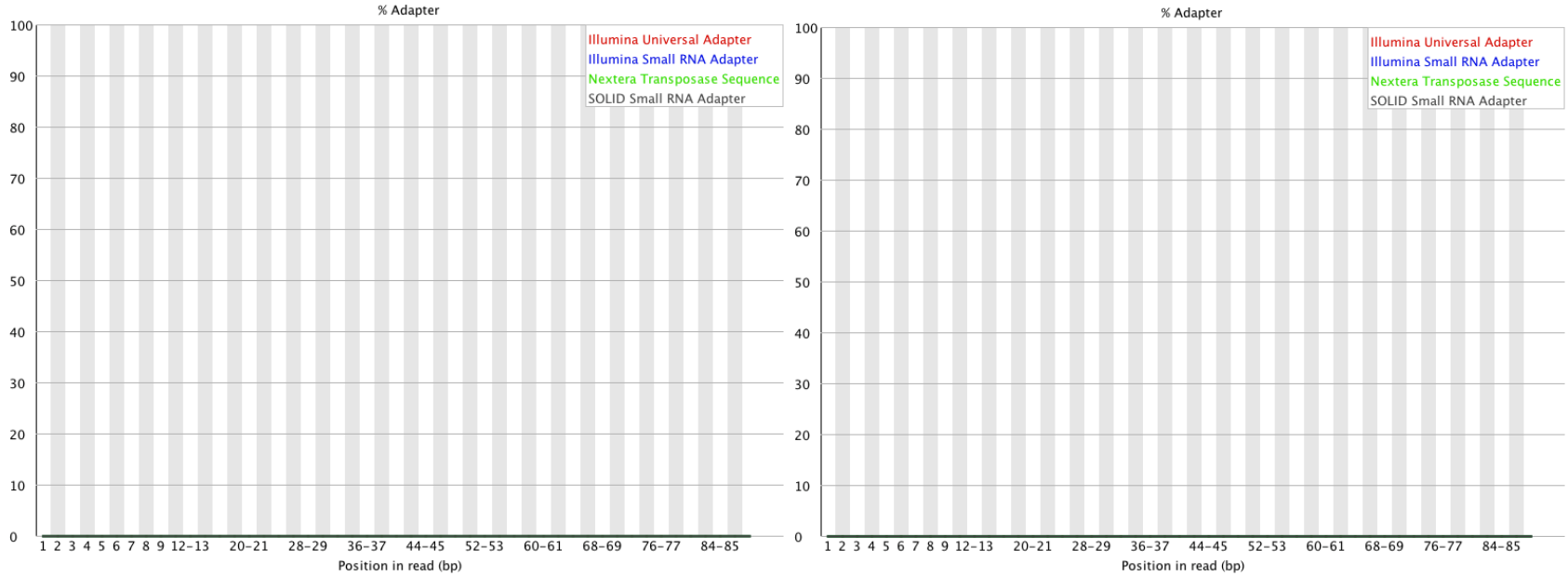

Figure 20: Percent adapter content. *"Forward" paired-end reads left; "Reverse" paired-end reads right. No adapter contamination was detected, nor were any overrepresented sequences or Kmers.*

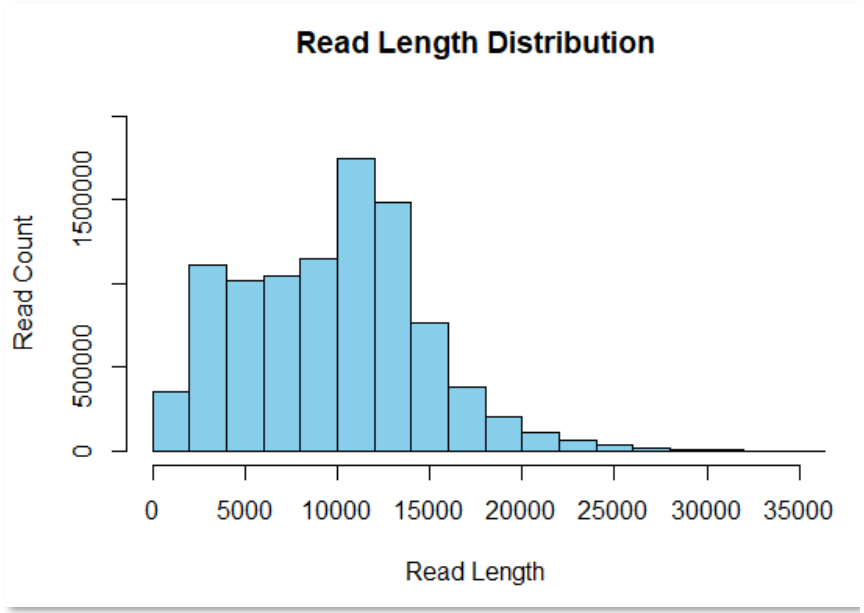

***Figure 21*: Histogram of molecule length distribution for PacBio SMRT sequence data.** Bins have a width of 2 kbp. (Mean: 6.3 kbp; N50: 21.3 kbp)

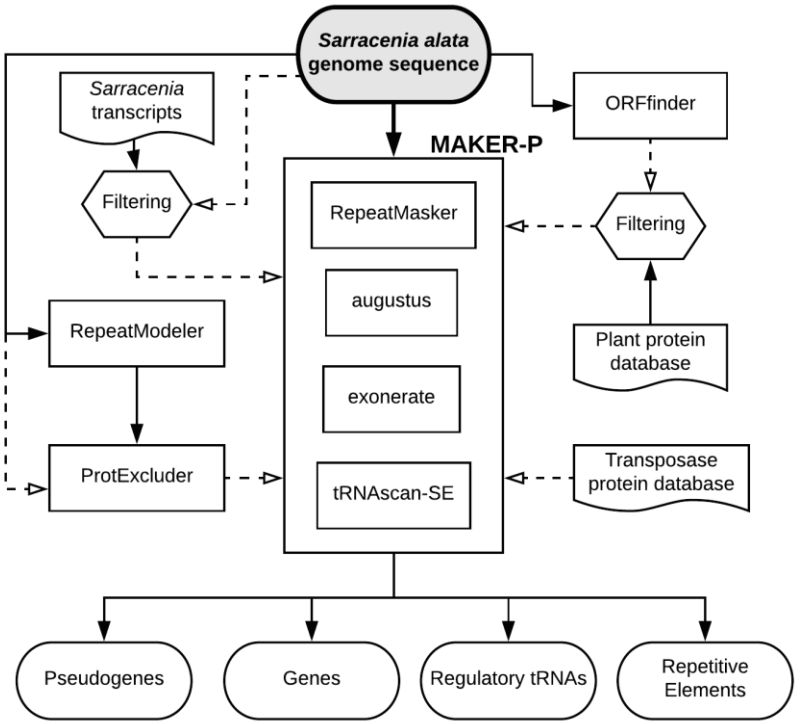

Figure 22: Automated annotation pipeline, using MAKER-P. *Dashed lines indicate sequences input as references only, which will not constitute part of the output data of the subsequent step.* Sarracenia alata *genome sequence is used to produce custom repeat libraries and filter protein databases and mRNA references for efficiency. These, along with a transposase sequence data are passed to MAKER-P as references, for use by its constituent programs. The product of this annotation is a list of regions identified as genes, pseudogenes, repetitive elements, and regulatory tRNAs, stored in GBFF format*

Appendix C: Chapter 4 Supplemental Materials

Chapter 4 Supplemental Tables

Function	Abbreviation	Microbiome	*Sarracenia*	Total
ammonium transmembrane transport	NHTrans	5170.6	4516.6	9687.1
sodium ion transmembrane transporter activity	NA+Trans	493.9	4512.6	5006.5
ATPase activity	ATPase	1029.1	0.0	1029.1
peroxidase activity	Perox	350.4	0.0	350.4
beta-glucanase activity	B-Gluc	340.1	0.0	340.1
ribonuclease activity	Ribonuc	313.2	0.0	313.2
cysteine-type peptidase activity	CystPep	302.2	0.0	302.2
chitinase activity	Chit	302.0	0.0	302.0
polygalacturonase activity	Polygal	297.6	0.0	297.6
xylanase activity	Xylan	293.9	0.0	293.9
thioglucosidase activity	Thiogluc	293.8	0.0	293.8
heat shock protein activity	HSP	253.5	0.0	253.5
protein homodimerization activity	ProtHomo	191.3	0.0	191.3
superoxide dismutase activity	SuperOx	185.2	0.0	185.2
phosphatase activity	Phosp	137.4	0.0	137.4
lipid transport	LipTrans	134.2	0.0	134.2
cinnamyl-alcohol dehydrogenase activity	CinAlc	103.4	0.0	103.4
serine-type carboxypeptidase activity	SerCarPep	98.5	0.0	98.5
lipase activity	Lipase	92.5	0.0	92.5
Other	Other	229.7	0.0	229.7
glutathione transferase activity	-	61.9	0.0	61.9
fructose-bisphosphate aldolase activity	-	44.4	0.0	44.4
polygalacturonase inhibitor activity	-	23.2	0.0	23.2
alpha-galactosidase activity	-	22.7	0.0	22.7
aspartic-type endopeptidase activity	-	19.6	0.0	19.6
alternative oxidase activity	-	17.7	0.0	17.7
formate dehydrogenase complex	-	11.1	0.0	11.1
ATP:ADP antiporter activity	-	10.3	0.0	10.3
cyclic-nucleotide phosphodiesterase activity	-	9.4	0.0	9.4
methylammonium channel activity	-	3.2	0.0	3.2
water channel activity	-	3.2	0.0	3.2
beta-galactosidase activity	-	2.4	0.0	2.4
phospholipase activity	-	0.5	0.0	0.5
actin filament	-	0.1	0.0	0.1
endonuclease complex	-	0.0	0.0	0.0
glucosidase complex	-	0.0	0.0	0.0
symplast	-	0.0	0.0	0.0

Table 22: Carnivory function expression-level data, as shown graphically in Figure 10. *Also includes key of abbreviated function to full-length Gene Ontology terms.*