The stress response of pokeweed and phylogeny of a defense gene family in plants

Kyra Dougherty

A thesis submitted to the Faculty of Graduate Studies

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Graduate Program in Biology

York University

Toronto, Ontario

August 2023

Abstract

Pokeweed (*Phytolacca americana L.*) is a non-model plant known for its resistance to a variety of biotic and abiotic stressors. Part of the reason for this is the presence of ribosome inactivating proteins (RIPs), which can hydrolyze adenine bases from nucleic acids. These proteins are upregulated in pokeweed by jasmonic acid, a plant hormone involved in stress response. The goal of this research was to gain a better understanding of how plants respond to stress and so two different, but complementary, approaches were taken. Firstly, an in-depth look at the diversity and evolution of RIPs in plants was undertaken by curating a dataset of RIPs from publicly available data and using computational approaches to characterize their domain architecture, identify conserved amino acids, and construct a gene tree. This research revealed that despite the damage that RIPs can potentially cause to the plant's own nucleic acids, RIPs are common among plants and their diversity indicates the potential for a multi-faceted impact on plant defense. Looking more closely at how pokeweed responds to stress, we applied jasmonic acid to leaves and analyzed changes in gene expression, through RNA sequencing (RNA-Seq). Identification of gene clusters involved in defense was aided by the generation of a pokeweed genome assembly. This research revealed that there is a variety of strategies that plants can implement to respond to stress, and that these strategies are applied differently by different species. Overall, this research contributes to a better understanding of the diversity and nuance present in the ways plants defend themselves.

Acknowledgements

Firstly, I would like to thank Dr. Kathi Hudak for being endlessly supportive and encouraging, even when my research was far outside anything either of us had ever done before. She continues to inspire me to be bold and brave in the pursuit of knowledge and science because she leads by example.

Next, I would like to thank my advisor Dr. Nik Kovinich for his support; his feedback provided another angle for the project that I hadn't thought of on my own. I'd also like to thank Dr. Robert Cribbie and Dr. Mark Bayfield for agreeing to be on my examination committee.

Although I haven't seen any of my past and present lab mates Jennifer, Tanya, Alex, and Fernand in person for over two years, our weekly lab meetings have always been a grounding force for me. I would also like to thank Kira, who was the first bioinformatician in the Hudak lab, for all her invaluable help when I have struggled over the years.

Lastly, I would like to thank my wife and my family for always believing with full confidence that I could do whatever I set my mind to. I never would have gotten here without everyone who supported me.

Table of Contents

List of Figures

List of Tables

List of Common Abbreviations

AOC	Allene oxide cyclase
AOS	Allene oxide synthase
bHLH	Basic-helix-loop-helix
COI1	Coronatine Insensitive 1
CYP94B1	Cytochrome P450 family protein
DBG	de Bruijn graph
DNA	Deoxyribonucleic Acid
Gb	Gigabase
GO	Gene ontology
Mb	Megabase
JA	Jasmonic acid
JAZ	Jasmonate-ZIM domain
JOX2	Jasmonic acid oxidase 2
Kb	Kilobase
LOX2	Lipoxygenase 2
OLC	Overlap-layout consensus
OPDA	12-oxo-phytodienoic acid
PAP	Pokeweed antiviral protein
RIP	Ribosome inactivating protein
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing

Chapter 1 - Introduction

1.1 Ribosome inactivating proteins (RIPs)

1.1.1 RIP function

Ribosome inactivating proteins (RIPs; EC 3.2.2.22) are N-glycosylases which hydrolyze purine bases from ribonucleic acid (RNA) (Karran and Hudak 2008; Fabbrini et al., 2017; Zhu et al., 2018). The key to their enzymatic activity is four highly conserved amino acids: valine 73 and serine 212 are involved in stabilizing the adenine substrate (Monzingo and Robertus, 1992; Gu and Xia, 2000), and two involved in base hydrolysis; arginine 179 protonates N-7 and glutamic acid 176 directs a water molecule to resolve the oxocarbenium ion that results from this destabilization (Frankel et al., 1990; Li et al., 1999). An illustration of the chemical mechanism for RIP depurination of RNA is shown in Figure 1.

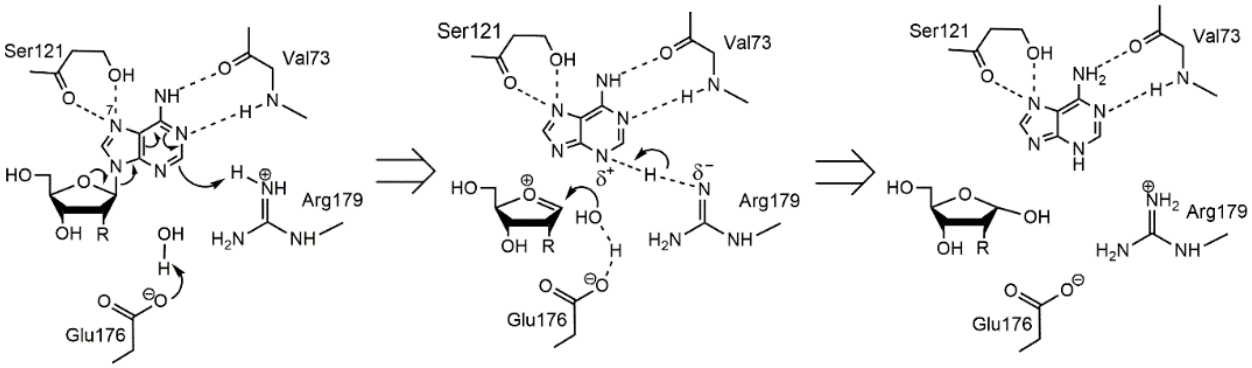


Figure 1 - Chemical mechanism for base excision of RNA by RIPs. Amino acid numbering from pokeweed antiviral protein is illustrated. Image modified from Prashar et al., 2023.

RIPs are best known to depurinate the highly conserved ribosomal RNA (rRNA) substructure called the sarcin/ricin loop (Endo and Tsurugi, 1988). Specifically, they hydrolyze the N-glycosidic bond of A_{4324} on 28S rRNA without cleaving the ribose phosphate backbone (Endo et al. 1987, Endo and Tsurugi 1987). The loss of functionality in the sarcin-ricin loop results in an inability to recruit protein factors needed for translation elongation (Montanaro et al. 1975) and reduces translation factor binding to ribosomes which therefore inhibits translation (Montanaro et al., 1975; Osborn and Hartley, 1990). This mechanism of action means RIPs are primarily involved in plant defense, but this also makes them a toxin when ingested (Montanaro et al., 1975; Osborn and Hartley, 1990; Grela et al., 2019). RIP activity on ribosomes can limit pathogen spread by causing cell death (Foa-Tomasi et al., 1982; Watanabe et al., 1997). In animals, damage to ribosomes, caused by UV radiation or depurination by a RIP, triggers a ribotoxic response and pyroptosis (Iordanov et al. 1997; Robinson et al. 2022). On the other hand, causing damage to a cell's own ribosomes can be beneficial in the case of infection by a pathogen because this damage could cause death of a host cell and therefore limit pathogen spread (Foa-Tomasi et al., 1982; Watanabe et al., 1997). Additionally, the potency of RIPs varies depending on environmental conditions, with some being far less enzymatically efficient (Bass et al., 1992; Hey et al., 1995) and some requiring ATP or other proteins to function optimally (Carnicelli et al., 1992) and others still cannot enter cells as easily (Stirpe et al. 1980). Therefore, by careful protein localization the plant could minimize damage to itself and maximize damage to attackers. For example, early identified RIPs were known to be sequestered co-translationally to the endoplasmic reticulum and then to the apoplast or to other membrane-bound subcellular compartments and the hypothesis was that this was to

protect the plant's ribosomes from depurination (Youle and Huang, 1976; Ready et al., 1986). Alternatively, some RIPs that are not sequestered in this way are synthesized as inactive precursors and their capacity to cause damage is limited (Walsh et al., 1991; Hey et al., 1995).

RIPs have antiviral activity against a number of viruses that target plants such as chilli venial mottle virus, cucumber mosaic virus (Zhu et al. 2013), turnip mosaic potyvirus (Yang et al. 2016), artichoke mottled crinkle virus (Bolognesi et al. 1997), tobacco mosaic virus (Dallal and Irvin 1978; Prasad et al. 1995; Verma et al. 1996; Vivanco and Tumer 2003; Yang et al. 2016), brome mosaic virus (Baranwal et al. 2002; Vivanco and Tumer et al. 2003), sunn-hemp rosette virus (Verma et al. 1996; Roy et al. 2006; Choudhary et al. 2008), pokeweed mosaic virus (Baranwal et al. 2002), papaya ringspot virus (Srivastava et al. 2009), and zucchini yellow mosaic virus (Sipahioğlu et al. 2017). This is because RIPs also depurinate messenger RNA, viral RNA, and deoxyribonucleic Acid (DNA) (Barbieri et al., 1997; Gandhi et al., 2008; Zhabokritsky et al., 2014). By depurinating viral RNA, RIPs hinder translation of viral proteins (Gandhi et al., 2008; Zhao et al., 2009; Zhabokritsky et al., 2014) and replication of viral RNA (Karran and Hudak 2008) resulting in reduced viral proliferation. Some RIPs have been observed to have a disproportionate impact on viral RNA compared to other cellular RNA. For example, a well-studied RIP called pokeweed antiviral protein (PAP, also known as PAP-I) was shown to inhibit 30% of cellular protein synthesis, but 90% of viral production (He et al., 2008). On the other hand, some viruses are more susceptible than others to depurination (Vivanco et al. 2003). The precise RNA structure required for binding and depurinating viral or messenger RNA is unclear but there is evidence that RIPs can bind to capped RNA (Hudak et al. 2000) but that this cap alone is not the only feature required for efficient depurination (Vivanco et al. 2003).

1.1.2 RIP classification

There were traditionally three types of RIPs described, classified based on their domain structures (reviewed in Stirpe 2004; Zhu et al. 2018). The type I RIP has a single-domain with N-glycosylase activity, which will henceforth be referred to as a RIP domain. The most well-known example of a type I RIP is PAP-I from pokeweed (Irvin 1975; Dallal and Irvin, 1978). It was first identified as an antiviral agent by Duggar and Armstrong (1925) and then later as a protein (Kassanis and Kleczkowski, 1948; Wyatt and Shepherd, 1969). Type II RIPs have the RIP domain plus an additional lectin-binding domain. These RIPs are more toxic than type I because the secondary domain can bind to membrane glycoproteins and facilitate entrance into cells, giving them a higher potential for toxicity (Stirpe and Barbieri, 1986; Sandvig and van Deurs, 1994; Steeves et al., 1999). Ricin from castor bean (*Ricinus communis*) is a well-known example of a type II RIP (Endo et al. 1987, Endo and Tsurugi 1987). It is the first RIP to be isolated (Stillmark 1888) and is famous for its use as a poison (Papaloucas et al. 2008). Ricin is also highly efficient at depurination, with the RIP domain capable of depurinating over 1000 ribosomes per minute (Endo and Tsurugi 1988). Type III RIPs, which are less toxic than other RIPs (Bass et al., 1992; Hey et al., 1995), are both unlike type I and II RIPs but also unlike each other. The first type III RIP identified was B-32 from corn (*Zea mays*), which has a RIP domain but also has a C-terminal domain that has amino acid similarity to the translation initiation factor eIF4E (Walsh et al., 1991). Shortly after this discovery, JIP60 was identified from barley (*Hordeum vulgare*), and it is unique because it is initially synthesized as an inactive proenzyme, but once a certain internal sequence is removed by proteolytic processing, the remaining N and C-terminal peptides combine into the active form of the RIP (Chaudhry et al., 1994; Rustgi et al., 2014). As more

sequencing data have become available other type III RIPs have been identified in silico, such as nuRIP from rice (Oryza sativa) which shares homology with JIP60 (Wytynck et al. 2021).

There is currently no consensus about how to classify RIPs. Some reject the notion of a type III classification given only two confirmed members (Stirpe 2004; Schrot et al. 2015), meanwhile a variety of other classification systems have been proposed. Lapadula and Ayub (2017) suggested that the differences between the known type III RIPs was sufficient to warrant this group splitting into two sub-categories (Figure 2). De Zaeytijd and Van Damme (2017) proposed that eight classifications would better reflect the sequence diversity observed, however this research failed to include the sequence data to support their claims. Schrot et al. (2015) proposed a classification system of at least seven groups based primarily on physicochemical properties and documented enzymatic activity. These classifications are summarized in Table 1.

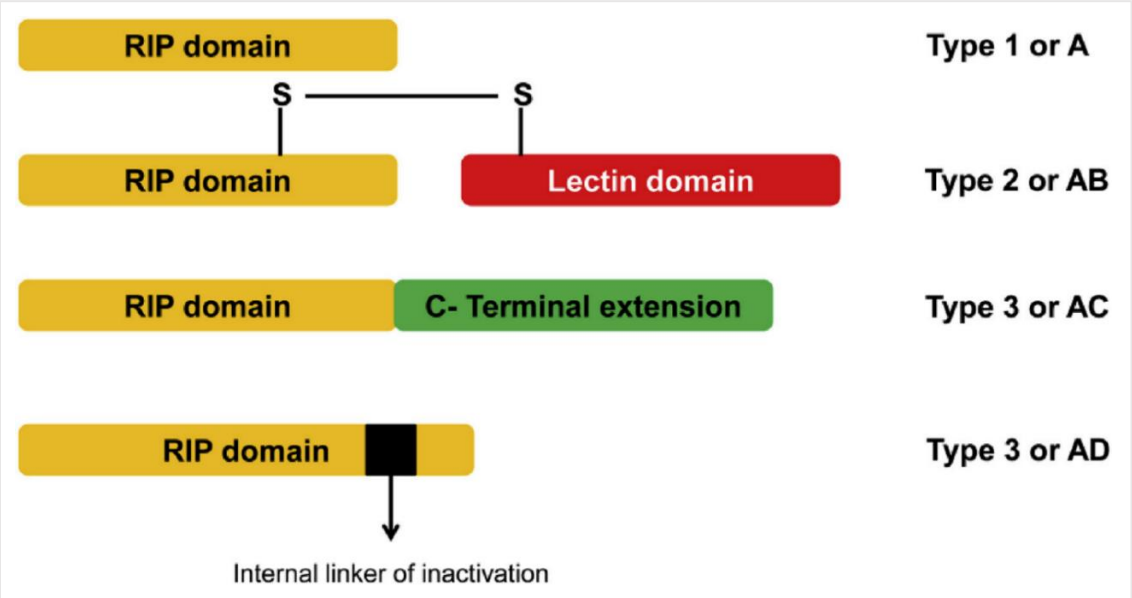


Figure 2 – Simplified outline of the two predominant RIP classification systems, and their approximate domain structures. Figure taken from Lapadula and Ayub (2017).

Table 1 - Summary of proposed RIP classification systems

Classical nomenclature	Lapadula and Ayub (2017)	De Zaeytijd and Van Damme (2017)	Schrot et al. (2015)
Type I / Type 1	Type A	Type A Type $A^{\Delta B}$ Type $A^{\Delta X}$	Small type 1 Type 1 Type 1 RIP candidate
Type II / Type 2	Type AB	Type AB	Small type 2 Type 2 Type 2 candidate
Type III / Type 3	Type AC Type AD	Type AX Type AC Type AD Type AP	
			RIP-like protein Peculiar RIP

1.1.3 RIP prevalence in plants

There was early speculation that RIPs may be widely distributed in plants which turned out to be the case. By the early 2000s, RIPs had been identified in about 70 plant species (Stirpe 2004), and by the late 2010s they had been identified in over 40 plant families (Schrot et al. 2015) and experimentally demonstrated in 12 plant orders (Lapadula and Ayub 2017); rice (*Oryza sativa*) alone has over 30 RIPs identified (Jiang et al. 2008). While not ubiquitous in plants because they are not present in well-studied species such as Arabidopsis (*Arabidopsis thaliana)*, RIPs are prevalent in both monocots and dicots indicating that they predate this important speciation event (Peumans et al., 2014; De Zaeytijd and Van Damme, 2017). Previous reports suggest that RIPs are present in Gnetophyta (Peumans and Van Damme, 2010; Peumans et al., 2014) and kelp (*Laminaria japonica A*) (Liu et al. 2002) although further research is needed to substantiate these results. While RIP genes have primarily been identified in plants, they have also been found in bacteria, fungi, and Metazoa, although the latter was

likely originated through horizontal gene transfer (Lam and Ng, 2001; Bergan et al., 2012; Lapadula et al., 2017; Citores et al., 2021).

1.1.4 *Phytolacca americana* and RIPs

Pokeweed (*Phytolacca americana L.*) is native to eastern North America and is a member of the taxonomic order Caryophyllales. While a primary area of research interest for this plant is in phytoremediation (Peng et al., 2008; Liu et al., 2010; Zhao et al., 2011), it is also an interesting candidate for studying plant defense because it is broadly resistant to many pathogens (Zoubenko et al., 1997; Lodge et al., 1993; Karran and Hudak, 2008; Mansouri et al., 2009). Part of this antipathogenic effect is due to the presence of RIPs, the best studied of these being PAP-I. This type I RIP has an N-terminal signal peptide that marks the protein for deposition into the apoplast (Ready et al., 1986) which means PAP-I can be stockpiled outside of the plasma membrane in preparation for an attack. In the case of viral infections, PAP-I has been shown to depurinate some plant and animal RNA viruses which prevents their replication (He et al., 2008; Karran and Hudak, 2008; Krivdova and Hudak, 2015; Mansouri et al., 2009) and limits translation of viral proteins in infected cells (Gessner and Irvin, 1980). Additionally, plants modified to express PAP-I are resistant to viruses and fungi (Lodge et al., 1993; Zoubenko et al., 1997).

PAP-I is not the only RIP present in pokeweed, however. Several others have been identified, isolated, classified, and named by various researchers such as PAP-II (Barbieri et al., 1982), PAP-C (Barbieri et al., 1989), PAP-R, PAP-H (Bolognesi et al., 1990; Park et al., 2002), PAP-S1, PAP-S2 (Honjo et al., 2002), PAP-III (Kurinov and Uckun, 2003), and PAP-α (Kataoka et al.,

1992; Neller et al. 2016). As of the first pokeweed genome assembly, the annotated RIPs present in pokeweed are PAP-I, PAP-II, PAP-α, PAP-S1, PAP-S2, and novel PAP (Neller et al. 2019). These PAP isoforms have variable depurination activity (Honjo et al., 2002; Kurinov and Uckun, 2003; Rajamohan et al., 1999) and do not all respond equally to defense-related hormones (Neller et al., 2016; 2018; 2019) indicating that different RIPs are required for different stress contexts. Notably, however, most respond strongly to jasmonic acid (JA), with PAP-I transcript levels increasing the most upon treatment of plants with this hormone (Neller et al. 2019).

1.2 Plant defense

1.2.1 Jasmonic acid signaling pathway

JA is a plant phytohormone that acts as a signaling molecule in response to biotic and abiotic stresses such as drought, salt, bacteria, viruses, and insects (Wasternack and Feussner 2017). At the same time, JA is required for male fertility in Arabidopsis (McConn and Browse, 1996; Stintzi and Browse, 2000; Browse, 2005). Therefore, the primary role of JA is to strike a balance between growth and defense (Huot et al. 2014; Chini et al., 2016). When JA is absent, Jasmonate-ZIM domain (JAZ) proteins bind to transcription factors involved in defense, thereby blocking the promoters of downstream genes which inhibits transcription (Chini et al. 2016).

When JA is present, JAR1 converts JA into its bioactive form jasmonoyl-L-isoleucine (Guranowski et al., 2007) which then binds to the F-box Coronatine Insensitive 1 (COI1) receptor (Fonseca et al. 2009). COI1 then physically associates with CUL1, Rbx1, and Skp1-like proteins which form a ubiquitin-ligase complex called SCF^{COI1} (Xu et al. 2002) and promotes the

ubiquitination of the JAZ proteins, marking them for degradation (Figure 3; Thines et al. 2007; Ali and Baek, 2020). JAZs have been shown to repress several stress-responsive transcription factors (Hu et al., 2013; Zhu et al., 2011; Boter et al., 2015), the most notable of these being MYC2 (Chini et al. 2016). MYC2 is a basic-helix-loop-helix (bHLH) transcription factor in Arabidopsis which regulates many genes involved in biotic and abiotic stress responses (Song et al., 2022) and is considered the 'master regulator' of the JA pathway (Kaza and Manners 2013). The MYC-based JA-mediated defense regulatory system is conserved in dicotyledonous plants (Boter et al. 2004) including Arabidopsis, tobacco, and tomato (Kazan and Manners 2013).

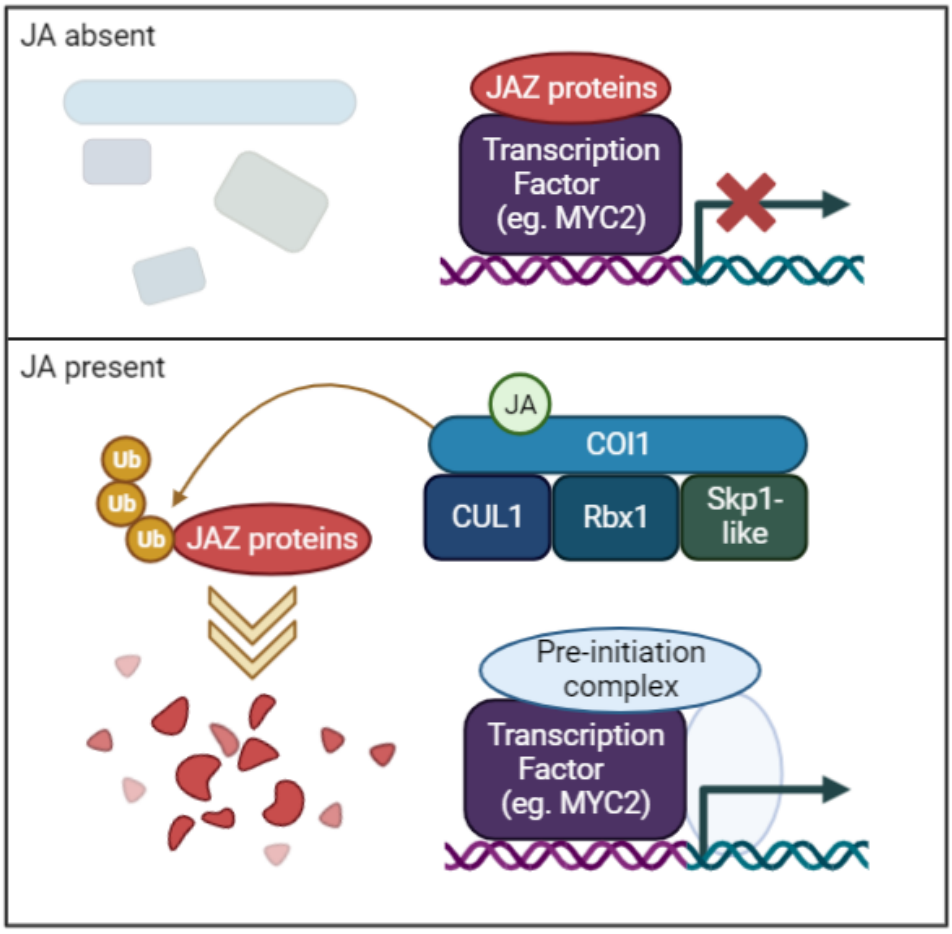


Figure 3 – Simplified illustration of transcription regulation by JA

1.2.2 Jasmonic acid biosynthesis

The JA biosynthesis pathway begins in the chloroplast with trienoic fatty acids which are then converted to 12-oxo-phytodienoic acid (OPDA) and dinor-OPDA (Schilmiller et al. 2006). This part of the process is called the oxylipin biosynthetic pathway and involves allene oxide synthase (AOS), lipoxygenase 2 (LOX2), and allene oxide cyclase (AOC) (He et al., 2002; Bannenberg et al., 2009; Pollmann et al., 2019). These precursors are then transported to peroxisomes where a 12-oxophytodienoate reductase (OPR3) reduces a certain double bond of the cyclopentenone moiety in 12-oxophytodienoic acid (Maynard et al., 2020). This intermediate is then converted to cyclopentanone compounds, then to CoA derivatives, and finally to JA (Figure 4; Schilmiller et al. 2006). If left unchecked, the biosynthesis of JA would lead to further biosynthesis of JA in a positive feedback loop. To combat this, jasmonic acid oxidase 2 (JOX2) catalyzes the oxidation of JA to 12OH-JA which attenuates this JA biosynthesis (Smirnova et al., 2017). Additionally, a cytochrome P450 family protein (CYP94B1) catalyzes the hydroxylation of the bioactive JA-Ile (Poudel et al., 2016) and MYC2 upregulates JAZ expression (Chini et al. 2007) which limits the expression of JA biosynthesis genes.

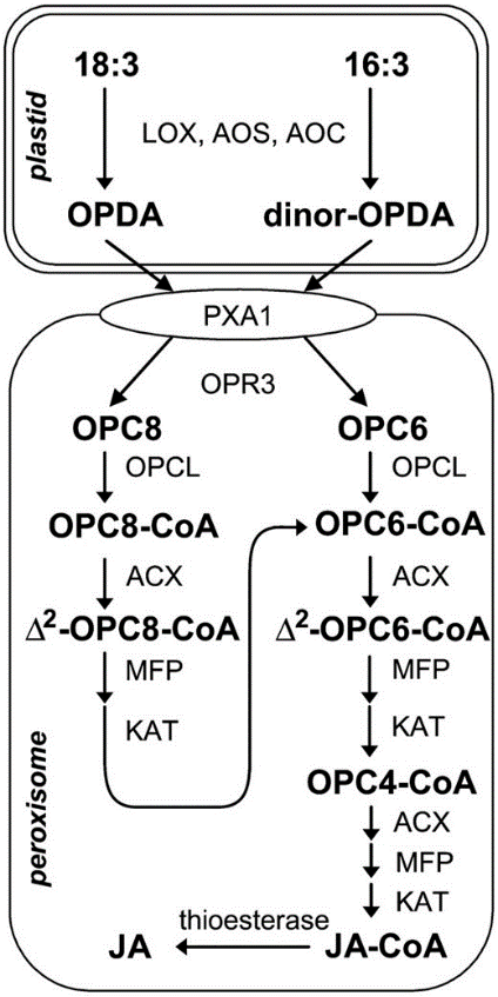


Figure 4 – Biosynthetic pathway converting trienoic fatty acids (18:3 and 16:3) to jasmonic acid (JA). Figure taken from Schilmiller et al. (2006).

1.3 Genome sequencing and assembly

Understanding complicated signaling pathways associated with some fundamental plant functions, such as stress response, is facilitated by good sequencing data. Over the last 20 years there has been an exponential increase in the number of sequenced plant genomes available; currently over 1000 genomes have been published, from non-vascular plants to flowering plants. The quality of these genomes is also much higher, with many at the chromosome level

(Figure 5; Sun et al. 2022). However, this lags behind what is available for prokaryotes and other eukaryotic clades because plants can have much more complicated genomes. To manage this, many of the early assemblies were done on plants with deliberately simplified genomes. For example, the first plant genome assembly was in the model plant Arabidopsis (*Arabidopsis thaliana*) a diploid with five chromosomes (The Arabidopsis Genome Initiative 2000). Another example is the first genome assembly in potato (*Solanum tuberosum*), normally a tetraploid and highly heterozygous species, which was done on a homozygous doubled-monoploid potato clone to simplify the process and ensure a higher quality result (Potato genome consortium 2011). As sequencing options have increased and the technology is more accessible, it is becoming more common for researchers to utilize multiple sequencing strategies to produce high quality assemblies, even for complex polyploid species with a high percentage of repeats; by 2020, there were genome assemblies for 62 polyploid plant species (Sun et al. 2022).

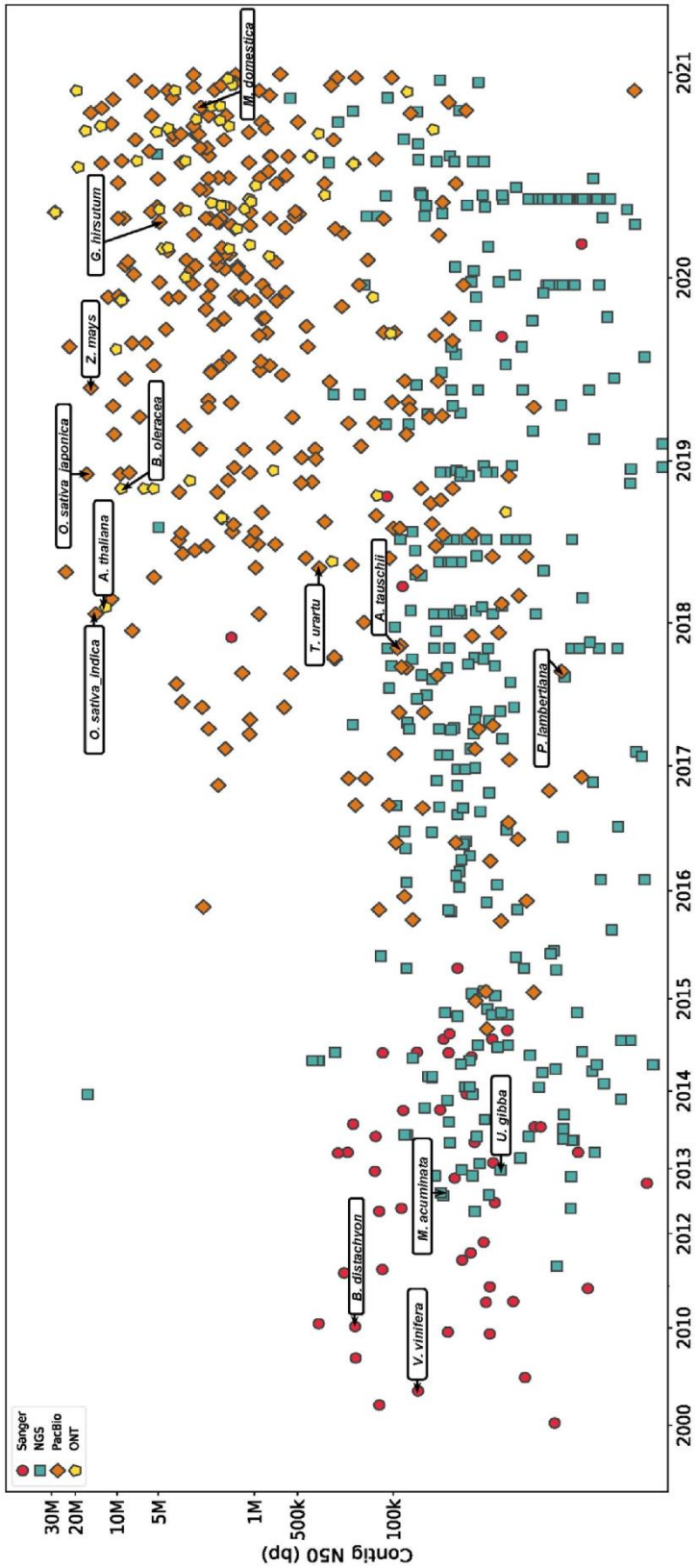


Figure 5 - Contiguity statistics of plant genomes published between 2000-2020. Each point is colour coded by the sequencing technology utilized: Sanger = Sanger sequencing; NGS = next generation sequencing/Illumina sequencing; PacBio = PacBio sequencing; ONT = Oxford nanopore sequencing. Contig N50 refers to the length of the contig at the halfway point when all contigs are ordered from smallest to largest. The x-axis refers to the year the plant genome assembly was published. Figure from Sun et al. (2022)

There are several whole genome sequencing technologies currently available, characterized by three major phases. First generation sequencing, also referred to as Sanger sequencing (Sanger, 1977) was used to generate the first genome assemblies but it is labor-intensive and expensive. Second generation sequencing, the most common of which being Illumina sequencing, is distinguished from first generation for being highly accurate while also being high throughput, although the read length is shorter than that of Sanger sequencing with a size between 50-300 bp per read (Slatko et al. 2018). Third generation sequencing is notable for having very long reads (Xiao and Zhou 2020). Oxford nanopore, one of the two main third-generation sequencing companies, has achieved the longest reads ever produced at nearly 1 Mb, but these reads tend to be error-prone (Jain et al. 2018). PacBio, the other main third-generation sequencing company, has resolved this issue with so-called 'PacBio HiFi' reads that are as accurate as Illumina reads (Hon et al. 2020).

To assemble a genome, sequencing reads are aligned into overlapping contiguous sections to create a sequence that became known as a 'contig' (Heather and Chain, 2016). The two most popular algorithms for genome assembly are the overlap-layout consensus (OLC) and the de Bruijn graph (DBG). Genome assembly programs using first-generation or error-prone long reads tended to favor the OLC method, whereas the DBG method tends to be used for high-accuracy and/or high-throughput reads such as Illumina and PacBio HiFi reads (Sun et al. 2022). Currently Hifiasm is the best *de novo* assembler for long reads. It uses a combination of both methods and produces accurate haplotype-resolved assemblies with reasonable computational resources (Cheng et al. 2021).

To make contig-level assemblies more contiguous, researchers can use algorithms or supplementary data to arrange the contigs in order in a process known as scaffolding. The two most popular methods for producing additional data are a version of chromosome conformation capture called Hi-C (Belton et al. 2012) and optical mapping techniques such as those produced by BioNano Genomics (Tang et al. 2015) and these methods have been capable of turning contig-level assemblies in plants into chromosome-level assemblies (Avni et al. 2017; Gui et al. 2018; Mitros et al. 2020). While there are several scaffolding programs available, they are prone to making errors (Hunt et al. 2014), therefore manual intervention is still often required.

1.4 Research objectives

The objective of this research overall is to provide insight into the ways pokeweed responds to and tolerates stress. This is explored more specifically by expanding our understanding of the structure, function, and evolution of RIPs, and by examining how pokeweed gene expression changes in response to a defense-related hormone.

Chapter 2 - Phylogeny and domain architecture of plant ribosome inactivating proteins

This chapter is presented as a peer-reviewed journal article.

Citation:

Dougherty K., Hudak K.A. (2022) Phylogeny and domain architecture of plant ribosome inactivating proteins. Phytochemistry. 202 doi: 10.1016/j.phytochem.2022.113337

Highlights

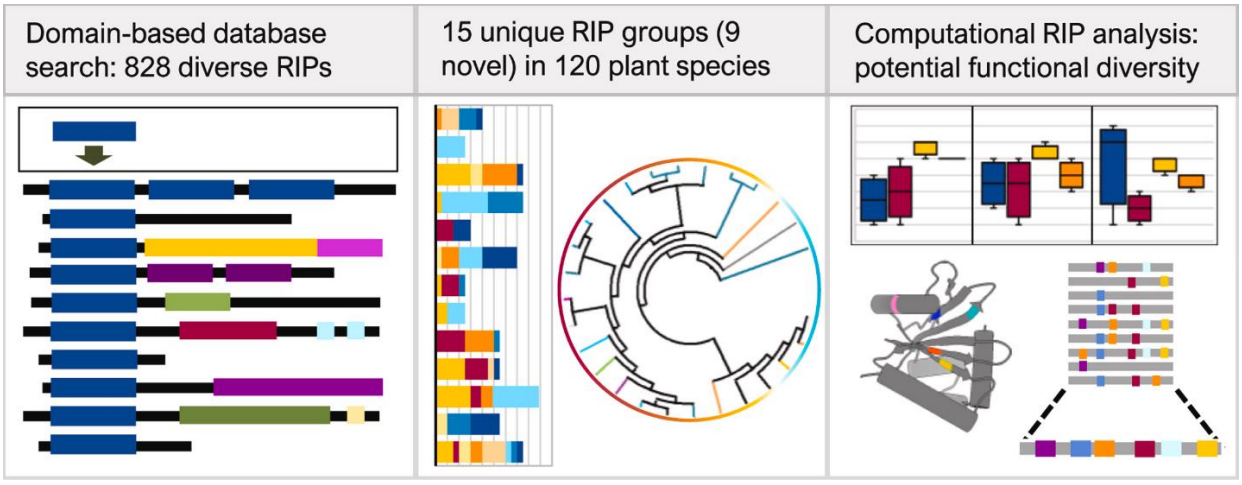
- Plant ribosome inactivating proteins (RIPs) are more diverse than previously known.
- We discovered 15 types of RIPs, based on protein domain configuration.
- Most RIPs lack signal peptides, suggesting their nucleocytoplasmic localization.
- Phylogenetic analysis shows early lectin domain-specific major clade separation.
- Unique physicochemical properties suggest varied functionality of the RIP family.

Abstract

Ribosome inactivating proteins (RIPs) are rRNA N-glycosylases (EC 3.2.2.22) best known for hydrolyzing an adenine base from the conserved sarcin/ricin loop of ribosomal RNA. Protein translation is inhibited by ribosome depurination; therefore, RIPs are generally considered toxic to cells. The expression of some RIPs is upregulated by biotic and abiotic stress, though the connection between RNA depurination and defense response is not well understood. Despite their prevalence in approximately one-third of flowering plant orders, our knowledge of RIPs stems primarily from biochemical analyses of individuals or genomics-scale analyses of small

datasets from a limited number of species. Here, we performed an unbiased search for proteins with RIP domains and identified several-fold more RIPs than previously known – more than 800 from 120 species, many with novel associated domains and physicochemical characteristics. Based on protein domain configuration, we established 15 distinct groups, suggesting diverse functionality. Surprisingly, most of these RIPs lacked a signal peptide, indicating they may be localized to the nucleocytoplasm of cells, raising questions regarding their toxicity against conspecific ribosomes. Our phylogenetic analysis significantly extends previous models for RIP evolution in plants, predicting an original single-domain RIP that later evolved to acquire a signal peptide and different protein domains. We show that RIPs are distributed throughout 21 plant orders with many species maintaining genes for more than one RIP group. Our analyses provide the foundation for further characterization of these new RIP types, to understand how these enzymes function in plants.

Graphical abstract



Introduction

Ribosome inactivating proteins are rRNA N-glycosylases (RIPs; EC 3.2.2.22) that hydrolyze purine bases from various RNAs (Fabbrini et al., 2017; Zhu et al., 2018). Theses enzymes are best known for their depurination of the highly conserved sarcin/ricin loop structure of the large ribosomal RNA, though other nucleic acid substrates such as DNA, poly(A), and viral RNA have been described (Endo and Tsurugi, 1988; Barbieri et al., 1997). RIPs are synthesized primarily by plants, but also by bacteria and fungi, with recent evidence indicating horizontal gene transfer to some insects, the only reported examples of RIP expression in the Metazoa (Lam and Ng, 2001; Bergan et al., 2012; Lapadula et al., 2017; Citores et al., 2021). Though not ubiquitous, our current study found that approximately one-third of all flowering plant orders contain RIPs, indicating that they are prevalent enzymes in plants.

Given that depurination of ribosomal RNA reduces translation factor binding to ribosomes and inhibits translation, RIPs are primarily viewed as toxins that mediate plant defense (Montanaro et al., 1975; Osborn and Hartley, 1990; Grela et al., 2019). Damage to ribosomal RNA could cause death of a host cell, limiting pathogen spread (Foa-Tomasi et al., 1982; Watanabe et al., 1997). Also, direct depurination of viral RNA by some RIPs hinders viral translation and replication, thereby reducing proliferation (Gandhi et al., 2008; Zhao et al., 2009; Zhabokritsky et al., 2014). The expression of some RIPs is also upregulated by biotic and abiotic stress, supporting their roles in defense. For example, transcript levels of SoRIP2, a RIP synthesized by spinach (*Spinacea oleracea*) increased in plants treated with salicylic acid, a hormone that regulates responses to pathogen infection (Kawade and Masuda, 2009). In addition, ectopic overexpression of a RIP in rice, OSRIP18, increased tolerance to drought and

salt stress (Jiang et al., 2012). However, much of the biochemical characterization of RIPs and mechanistic details of their enzyme activity have been conducted either *in vitro* or when expressed in heterologous systems; evidence of how RIPs function in the plants that produce them is still lacking.

Traditionally, RIPs are divided into three types based on their protein domain structure (Stirpe 2004). Type I is a single domain protein containing N-glycosylase activity. For example, pokeweed antiviral protein (PAP) is a classic type I RIP with an N-terminal signal peptide that co-translationally directs the protein into the lumen of the endoplasmic reticulum for subsequent exocytosis to the apoplast (Ready et al., 1986). Type II has two domains, the RIP domain linked to a lectin by a disulfide bond. Ricin, a classic type II RIP, also travels the endomembrane system for localization in protein storage vacuoles (Lord, 1985; Frigerio et al., 2001). The high toxicity of type II RIPs is attributed to lectin binding of membrane glycoproteins, thereby facilitating RIP entrance into cells (Sandvig and van Deurs, 1994; Steeves et al., 1999). Type III proteins are either single RIP domains formed from post-translational proteolytic processing, or RIP domains fused to other non-lectin domains. For example, Type III RIPs from maize (b-32) and barley (JIP60) were the early characterized members. B-32 is synthesized as an inactive proenzyme that is activated following the removal of an internal peptide and association of the remaining N- and C-terminal peptides into an active RIP (Walsh et al., 1991). JIP60, a methyl jasmonate-induced RIP, consists of an N-terminal domain with glycosylase activity linked to a C-terminal domain with amino acid similarity to eIF4E (Chaudhry et al., 1994; Rustgi et al., 2014). Curiously, type III RIPs are less toxic, and

unlike many type I and II RIPs, do not contain signal peptides to sequester them from ribosomes into subcellular or apoplastic spaces (Bass et al., 1992; Hey et al., 1995).

The question of how RIP types evolved and are related to each other has been examined by different groups (Di Maro et al., 2014; Peumans et al., 2014; De Zaeytijd and Van Damme, 2017; Lapadula and Ayub, 2017). One evolutionary model posits the existence of a type I RIP with a domain of unknown origin that arose before the separation of dicots from monocots and subsequently gave rise to all the RIPs of the Poaceae (Peumans et al., 2014; De Zaeytijd and Van Damme, 2017). Lapadula and Ayub (2017) do not recognize this RIP group as unique and based on sequence similarities, group them with Type III RIPs. In addition, a notable distinction among these models lies in the first appearance of signal peptides and the emergence of other domains within different taxonomic groups (De Zaeytijd and Van Damme, 2017; Lapadula and Ayub, 2017). Currently, there is lack of agreement regarding whether type II RIPs form a monophyletic group. Though these models have extended our knowledge of RIP evolution, unifying them is hampered by the fact that each chose biased and different criteria for selecting specific subgroups of RIPs. Since these studies, the number of sequenced plant genomes from non-model species has substantially increased (Sun et al., 2021), allowing for more thorough analysis of RIP phylogeny. In this work, we performed an unbiased search for plant RIPs based on genes with annotated RIP domains and discovered 15 different groups of RIPs across 21 plant orders, indicating that RNA glycosylases are significantly more diverse than previously suspected. During their evolution, RIPs have gained and lost protein domains, resulting in enzymes with different properties and conserved amino acids. We clearly define our data set

and make it accessible to support further characterization and a more comprehensive understanding of RIP function in plants.

Results

To identify RIPs in plants, all sequences with a RIP domain as identified by the Conserved Protein Domain Architecture Retrieval Tool (Geer et al., 2002; Yang et al., 2020) were downloaded from NCBI Entrez (Sayers et al., 2022) and filtered to remove low quality sequences, partial sequences, mature peptides, and duplicates; 828 proteins comprised the final dataset. Chemical properties of each protein were calculated from their respective sequences using the R package Peptides (version 2.4.4, Osorio et al., 2015). A comprehensive table with the NCBI accession number, species name, sequence, domain information, and all calculated properties is provided (Supplementary Data 1). The accession numbers and sequences of the proteins that were removed as duplicates are provided in a separate table (Supplementary Data 2).

RIP groups

Though RIPs are historically categorized as either type I, II or III (Stirpe, 2004), subsequent discovery of RIPs with different C-terminal domains necessitated the expansion of the traditional three types with regards to nomenclature. In this scheme, type 1 is termed A, type II is AB, type III RIPs are denoted with other letters in addition to the A, to indicate the presence of other C-terminal domains (Peumans et al., 2014). Though this nomenclature is useful when describing relatedness of a limited number of RIP types, for the sake of clarity, we chose to denote RIP groups by their literal protein domains and the presence or absence of signal peptides. To our surprise we discovered 15 categories of RIPs (Table 1). The

predominant form was the single RIP domain protein without a signal peptide (noSP-RIP), which comprised 56% of all identified RIPs. The single RIP domain proteins with a signal peptide (SP-RIP) were the second most prevalent at approximately 24% of RIPs. RIPs containing a lectin domain, written as RICIN domain, were divided into four groups: either with a single or repeated RICIN domain, and with or without a signal peptide. Taken together, this RICIN group comprised approximately 18% of all RIPs. The remaining nine categories of RIPs either contained repeating RIP domains or a single RIP domain accompanied by the addition of other domains. Though these nine categories comprised only 2.4% of all RIPs, we identified several domains involved in different functions. For example, six RIPs had BTB_POZ domains which are protein-protein interaction motifs involved in many cellular functions including transcriptional regulation, cytoskeletal dynamics and targeting proteins for ubiquitination (Stogios et al., 2005). Two RIPs were identified with WD40 repeat (WDR) domains which function as protein interaction scaffolds in multiprotein complexes (Stirnimann et al., 2010). They are a large family implicated in transcription regulation, cell cycle control and apoptosis (Jain and Pandey, 2018). The RNA recognition motif (RRM) involved in binding single-stranded RNA (Maris et al., 2005) was identified in five RIPs. RRM proteins have a variety of functions including regulation of RNA stability, translation and alternative splicing (Sachs et al., 1987; Query et al., 1989). Finally, unexpected domains were identified, such as Ras, which is characteristic of small GTPases involved in signaling pathways that regulate cell growth and differentiation (Hancock, 2003; Patel and Côté, 2013). We also identified glycoside hydrolase-3 (GH3) domains that regulate the levels of phytohormones. Specifically, these domains are characteristic of enzymes that conjugate amino acids to jasmonic acid and auxin, thereby controlling the

activity of these hormones (Staswick et al., 2005; Wakuta et al., 2011). None of these remaining nine RIP categories contained a signal peptide, suggesting that they may be localized to the nucleocytoplasm of cells during some portion of their expression. All protein domains with their features and superfamily names (conserved models) are listed in Supplementary Data 3. In addition to domain annotations extracted from NCBI, we searched for domains other than RIP and RICIN using InterPro and identified their associated gene ontology (GO) terms for biological process and molecular function. Classification of these unexpected domains from InterPro validated our NCBI-annotated protein domains, and their associated GO terms were consistent with the biochemical characteristics of proteins with these domains (Supplementary Data 4).

Table 1. Count of all RIPs in our dataset based on signal peptide and protein domains. Rows are divided based on presence/absence of a signal peptide predicted by SignalP or annotated in NCBI's Protein database, and domains listed in NCBI's Conserved Domain and Protein databases. Candidate proteins were identified using the Conserved Domain Architecture Retrieval Tool.

RIP	Count	Percentage of total proteins	
noSP-RIP	465	56.09%	79.61%
SP-RIP	195	23.52%	
noSP-RIP-RICIN	9	1.09%	2.05%
SP-RIP-RICIN	8	0.97%	
noSP-RIP-RICIN-RICIN	30	3.62%	15.92%
SP-RIP-RICIN-RICIN	102	12.30%	
noSP-RIP-RIP	1	0.12%	2.41%
noSP-RIP-RIP-RIP	3	0.36%	
noSP-RIP-BTB_POZ	6	0.72%	
noSP-RRM-RIP	1	0.12%	
noSP-RRM-RRM-RRM-RIP	2	0.24%	
noSP-RRM-RIP-RIP-RIP	2	0.24%	
noSP-RIP-Ras	2	0.24%	
noSP-GH3-GH3-RIP	1	0.12%	
noSP-RIP-WD40	2	0.24%	

To further investigate our suggestion that RIPs without signal peptides may be nucleocytoplasmic, we predicted cellular localization using ApoplastP, a program that predicts plant apoplast-localized proteins. Unlike SignalP, ApoplastP does not base its predictions on the presence of a signal peptide but rather the amino acid character of apoplastic proteins (Sperschneider et al., 2017). There was overall consistency between our initial SignalP predictions and the ApoplastP results, and as expected not all proteins with a signal peptide were predicted to be apoplastic (Supplementary Data 5). Specifically, the nine types of multi-domain RIPs, excluding those with a RICIN domain, were predicted to be non-apoplastic, in agreement with our signal peptide predictions. Curiously, ApoplastP could not efficiently predict cellular localization of proteins with a RICIN domain and we attribute this to the processing of characterized RIPs with a lectin – they are localized to storage vacuoles (Lord, 1985). Therefore, based on amino acid character and lack of signal peptide, we speculate that many of the RIPs in our database may not be sequestered via conventional protein secretion.

RIP physicochemical characteristics

Given the variety of domains found among the different plant RIPs, we were interested in comparing their physicochemical characteristics. The group containing a single RIP domain was the shortest in length (approximately 30 KDa; Fig. 1A). Curiously, those without a signal peptide (noSP-RIPs) were generally longer than the SP-RIPs and had the greatest range of lengths among all groups, suggesting that the extra amino acids may reflect a cytosolic localization with subsequent processing, given the lack of signal peptides for this group. Alternatively, these proteins may contain domains that have yet to be characterized or annotated. As expected, the RIP-RICIN domain proteins were on average twice the molecular

weight (60 KDa) of the RIP alone proteins. These RIP-RICIN and RIP-RICIN-RICIN proteins fall into the classic type II RIPs, with the lectin domain generally doubling protein size. The addition of other domains to the RIP domain increased the molecular weight of these protein groups, in particular the noSP-RRM-RIP-RIP-RIP and noSP-RIP-WD40 groups that were 4.5 and 6.5-fold larger, respectively, than the SP-RIP group.

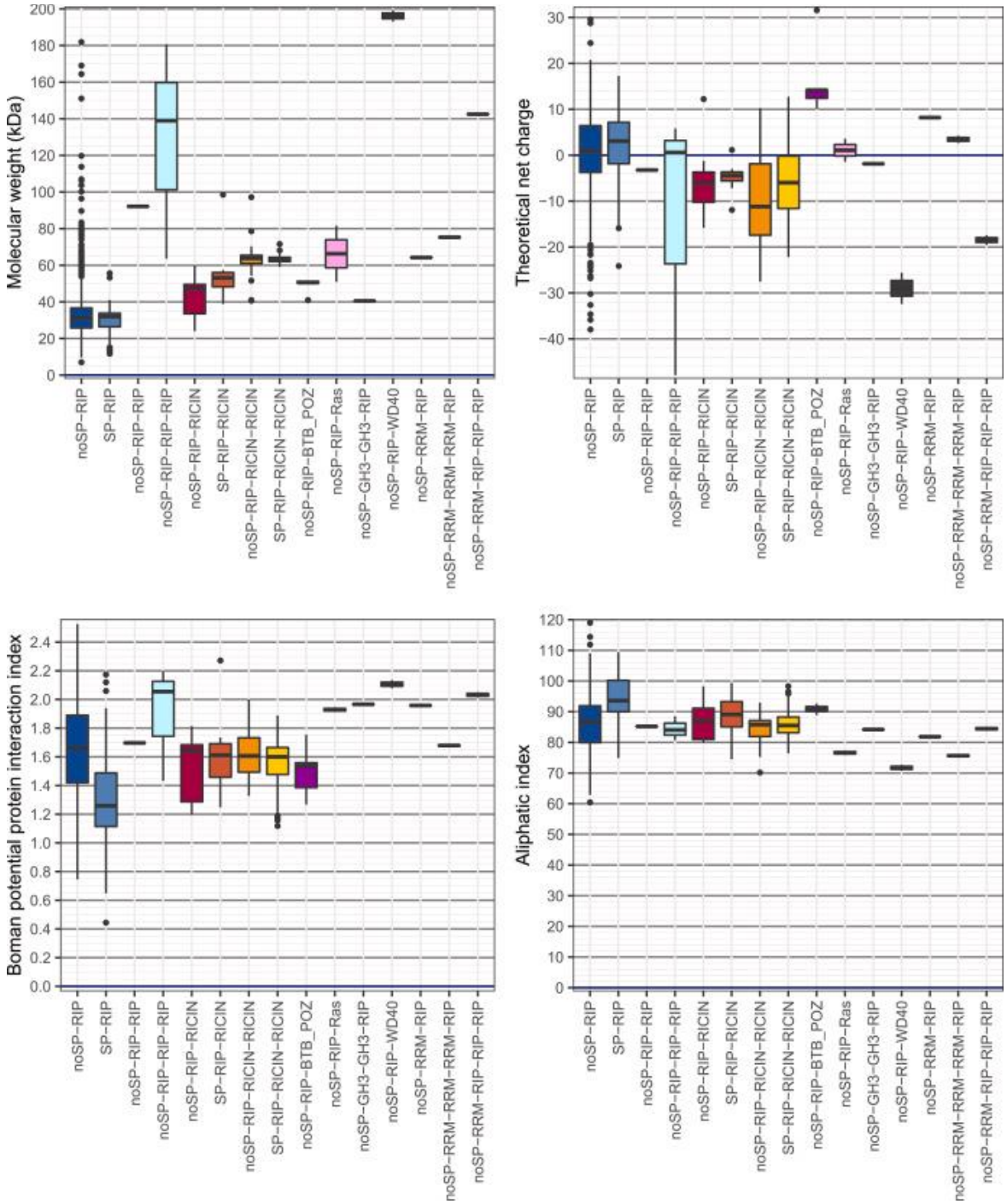


Fig. 1. – Physicochemical characteristics of different RIP groups. Values were calculated from the amino acid coding sequence of each protein in our dataset with the R package 'peptides' and presented on boxplot format. Our dataset comprised a curated selection from all the proteins available within NCBI's Conserved Domain Database containing a RIP domain. The groups were sorted based on presence/absence of a signal peptide and domains listed in both NCBI's Conserved Domain and Protein databases. (A) molecular weight prediction; (B) theoretical net charge prediction; (C) Boman potential protein interaction index prediction; (D) aliphatic index prediction.

The theoretical net charge, or overall charge of a protein at neutral pH (Bjellqvist et al., 1993), was calculated for each group of proteins and varied based on the domain type and number (Fig. 1B). For example, the four groups containing a RICIN domain were all negatively charged at pH 7.0, whereas proteins composed of only a RIP domain varied in charge based on the presence or absence of a signal peptide and number of repeating RIP domains. Moreover, addition of other domains to the RIP domain correlated with an increase in either the basic or acidic nature of these proteins, supporting the possibility of different functional interactions or cellular localizations. The greatest range in theoretical net charge was observed in the noSP-RIP group, correlating with their varied length and potential for different functional domains.

The Boman potential protein interaction index provides an overall estimate of the potential of a protein to bind to membranes or other proteins (Boman, 2003). The lowest values were associated with members of the SP-RIP group, which are sequestered from the nucleocytoplasm. The lack of a signal peptide for the noSP-RIP group correlated with the greatest range for potential interactions suggesting that this group consisted of proteins with different functions (Fig. 1C). While no proteins met the threshold for 'high binding potential' as a whole, the addition of domains to the single RIP domain tended to increase this index, indicating that these domains, such as the noSP-RRM-RIP-RIP-RIP and noSP-RIP-WD40, may be responsible for increased interaction with other cellular proteins or membranes. Accordingly, these two groups were also the most charged among the 15 different RIP groups, which may facilitate their interactions.

The aliphatic index, which is the measure of the relative volume of a protein occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine), is an indicator of globular

protein thermostability (Ikai 1980). Averages among the 15 RIP groups ranged from 68 to 95, with the noSP-RIP-WD40 group having the lowest value and the SP-RIP group with the highest thermostability (Fig. 1D). Once again, the noSP-RIP group showed the greatest range, suggesting they are the most diverse group of RIPs. Given that most globular proteins range in aliphatic index from 80 to 100 (Ikai 1980; Amblee and Jeffery, 2015), we considered the RIP groups to be generally thermostable.

RIP distribution among plant orders and species

We investigated the distribution of these 15 groups of RIPs among the 21 plant orders represented in our database of identified RIPs. The greatest number of RIPs and diversity of RIP groups was found in the Poales – 400 RIPs from 13 different groups (Fig. 2). The large number of sequenced species within this order may have contributed to the high number; however, the Poales contained 48% of all identified RIPs and only 22% of the species with RIPs, indicating that species within this order were enriched in RIP content relative to species within other orders. Most orders had more than one group of RIPs and only five orders contained one RIP group. Specifically, Celastrales, Trochodendrales, Myrtales each had one noSP-RIP identified, Oxalidales had two noSP-RIPs and Santalales had only SP-RIP-RICIN-RICIN members identified. As already noted, the dominant RIP group was the single domain RIP and only three orders, the Santalales, Dipsacales, and Apiales, did not have this group identified among their RIPs. Proteins containing a Ras or GH3 domain were found in the Asterales, whereas all other non-lectin domains were identified only in the Poales. In addition, noSP-RIPs were found primarily in the Poales but also distributed throughout the other orders, with only four (Asparagales, Santalales, Dipsacales, and Apiales) lacking the noSP-RIPs.

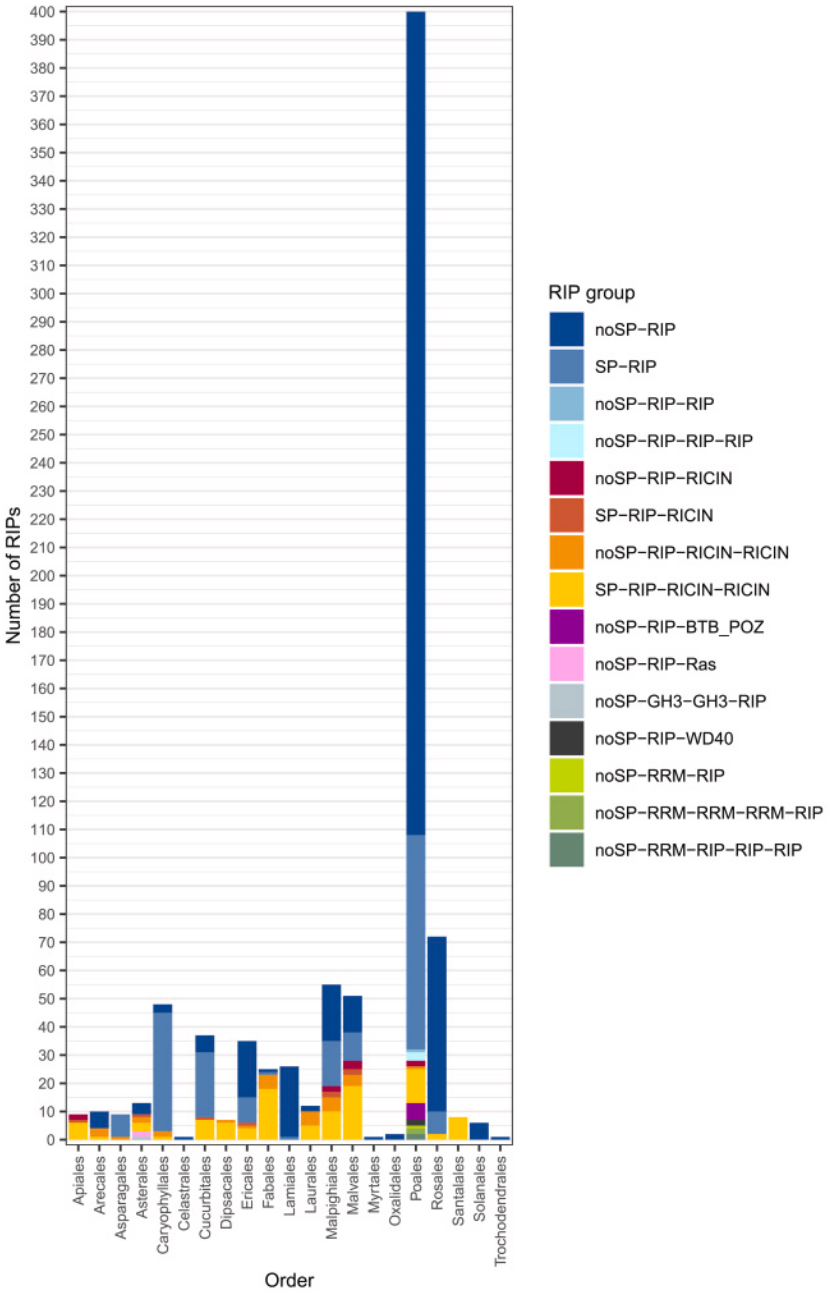


Fig. 2. – The number and type of RIPs within each plant order. Our dataset comprised a curated selection from all the proteins available within NCBI's Conserved Domain Database containing a RIP domain. The groups were sorted based on presence/absence of a signal peptide and domains listed in both NCBI's Conserved Domain and Protein databases. Colours represent different RIP groups. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Of the 120 plant species we identified with RIPs, 54 species (45%) had more than one type of RIP with one to several RIPs within each type, while 31 species (26%) had only a single RIP of a single type. A stacked bar graph of each species and the number and type of RIP identified in each illustrates that many species contained more than one RIP and many had more than one type (Fig. 3). The predominant form in most species was the noSP-RIP.

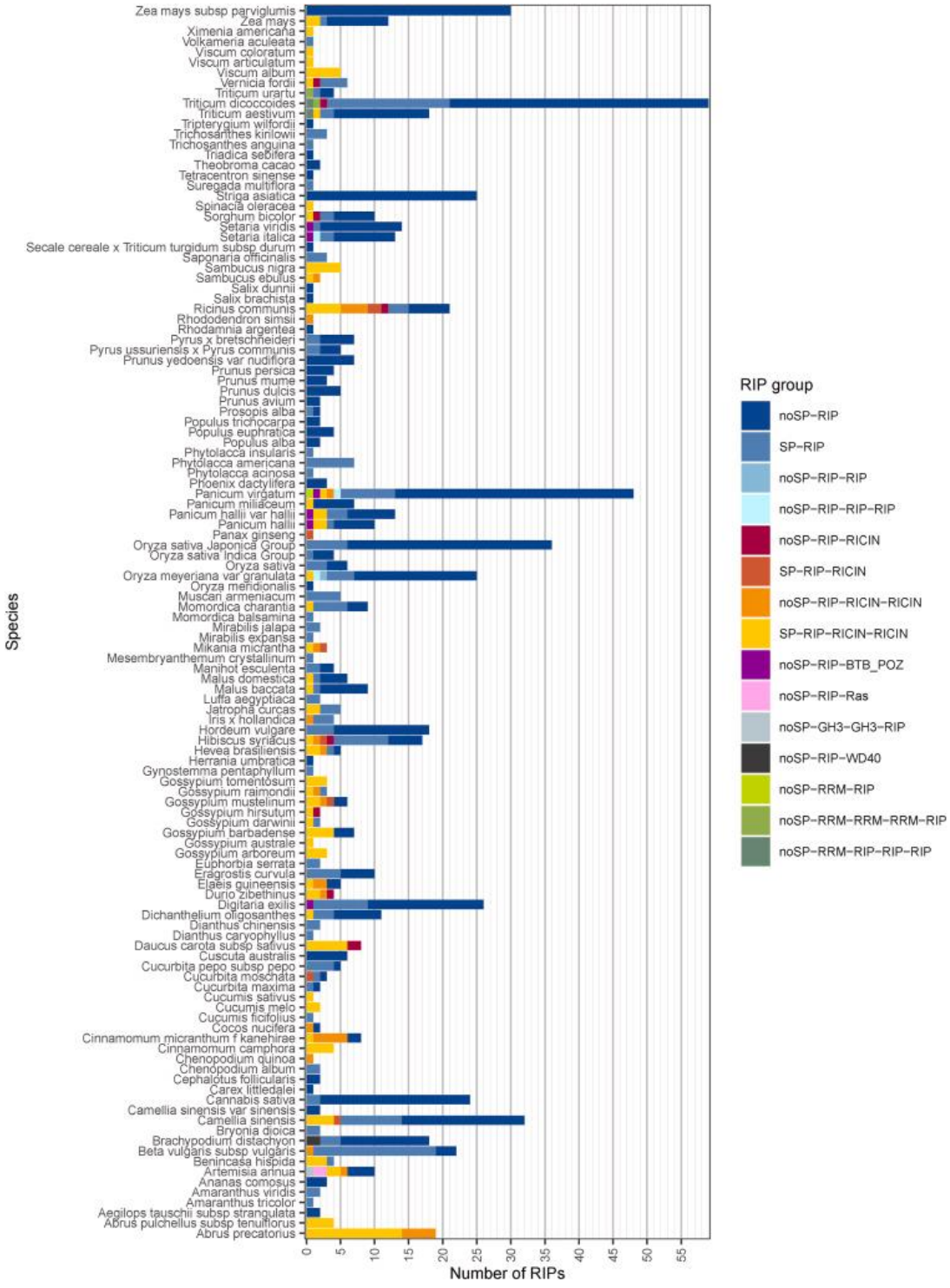


Fig. 3. - The number of RIPs within each plant species. Our dataset comprised a curated selection from all the proteins available within NCBI's Conserved Domain Database containing a RIP domain. The groups were sorted based on presence/absence of a signal peptide and domains listed in both NCBI's Conserved Domain and Protein databases. Colours represent different RIP groups. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

RIP phylogenetic tree

Given the number of RIP groups and the presence of more than one group in many plants, we wanted to visualize the evolutionary relationships among these proteins in the 120 species of RIP-containing plants. The colour coding of the phylogenetic tree illustrated in Fig. 4 is based on the grouping of structurally similar RIPs and indicates the plant orders as a coloured band around the circumference of the tree. The chosen outgroup enzymes were DNA-3-methyladenine glycosylase (https://www.uniprot.org/uniprot/A0A0D6R531) and alpha-amylase (https://www.uniprot.org/uniprot/A0A0D6QTL5), selected from hoop pine (*Araucaria cunninghamii*) because this species does not contain a RIP and is not within the division Magnoliophyta. From the RIP-based phylogenetic tree, we predict that the original parent group of all identified RIPs is the noSP-RIP; it was most closely aligned to the outgroup, suggesting that it evolved before the other RIPs. There is also an early division which separates the tree into two major clades: one comprising most of the lectin-containing proteins, with the other comprising primarily proteins in the noSP-RIP group and most of the other non-lectin-containing multi-domain proteins. This division is correlated with the division between monocots and dicots, as the proteins within dicots were primarily present in the former clade while monocots comprised the majority of the latter clade. Because proteins with a signal peptide tended to be closely related to proteins containing a lectin domain, it is likely that the emergence of these two structural features is closely aligned. The four lectin-containing RIP groups were closely related, apart from those within Poales, suggesting that the common ancestor of these proteins contained a lectin domain. In some lineages, the lectin domains were either duplicated or lost, to either revert the protein to a single-domain RIP, or

produce the double and single lectin domain proteins. As the other groups of multi-domain proteins are closely related to noSP-RIPs and not closely related to each other, it is likely that these proteins emerged independently through domain fusion events with noSP-RIPs. Some of these RIPs, specifically, noSP-RIP-Ras and noSP-GH3-GH3-RIP were present in the Asterales, indicating an independent evolution from those RIPs with non-lectin domains found within the Poales. A more detailed version of this phylogenetic tree, which includes bootstrap supports and individual protein labels, is available in Newick format in Supplementary Data 6.

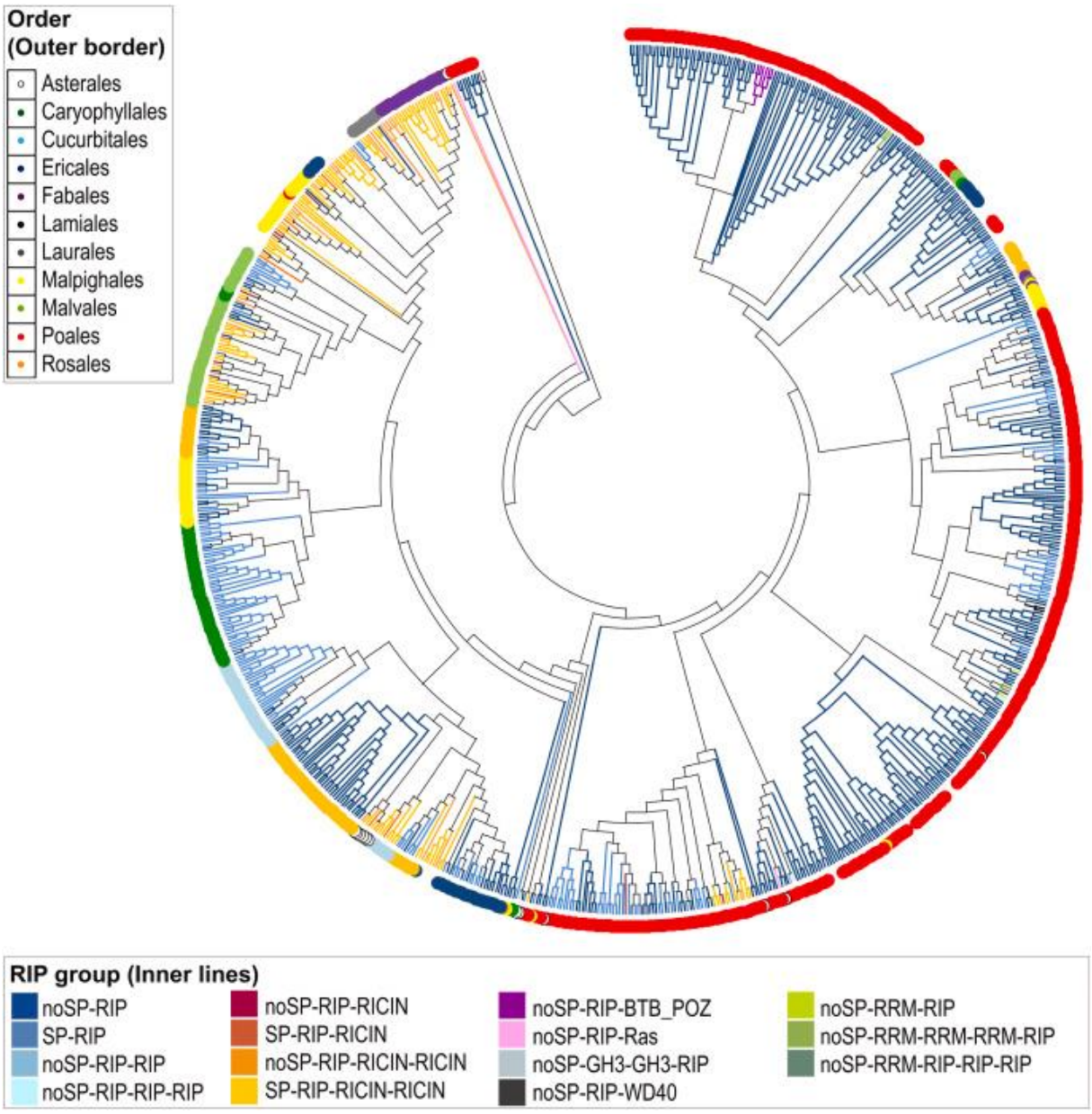


Fig. 4. – Circular gene tree of RIPs. Our dataset comprised a curated selection from all the proteins available within NCBI's Conserved Domain Database containing a RIP domain. The phylogenetic tree was constructed using the maximum-likelihood method with the amino acid substitution model WAG + R9, ultrafast bootstrapping approximation (UFBoot) with 1000 iterations and the SH-like approximate likelihood ratio test with 1000 iterations. The legend titled 'Order (outer ring)' outlines the coloured dots around the perimeter of the tree and represents the phylogenetic order from which each protein sequence originated. Proteins without a coloured dot belong to orders with less than ten proteins. The legend titled 'RIP domains (inner lines)' defines colours in the branches of the tree, each representing a RIP group, based on presence/absence of a signal peptide and domains listed in both NCBI's Conserved Domain and Protein databases. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Amino acid conservation across RIP groups

To consider what constitutes a RIP domain and identify conserved amino acids, we performed a multiple sequence alignment of all NCBI-annotated RIP domains within our dataset and obtained the consensus sequence as predicted by Jalview (v2.11.1.7; Waterhouse et al., 2009). Any amino acid in this alignment with 70% identity or greater was considered 'conserved' and was mapped onto the crystal structure of pokeweed antiviral protein 1QCI (Kurinov et al., 1999a, 1999b), a well characterized SP-RIP. The crystal structure is of mature pokeweed antiviral protein lacking the signal peptide; therefore, its first amino acid is valine (V). Based on protein domain databases (InterPro and Expasy Prosite) a single RIP domain comprises the majority of the amino acid sequence of pokeweed antiviral protein. Comparison of the consensus sequence with 1QCI indicated two positions with conserved amino acids that were not identical but semi-conserved in the sequence of 1QCI, (I vs L at positions 117 and 152), so the amino acids present at those positions in 1QCI were highlighted in their place (Fig. 5). Although the conserved amino acids were distributed throughout the sequences (Fig. 5A), when mapped to the crystal structure of 1QCI most clustered, predictably, within the enzyme active-site pocket (Fig. 5B). Previous reports of conserved amino acids within a RIP domain cite five essential, often non-variable amino acids (Van Damme et al., 2001; Fabbrini et al., 2017), whereas our analysis shows that E176 and R179 were the most conserved amino acids in the protein (95% and 92% respectively), while the remaining were less conserved (35–69%; data not shown). The glutamic acid and arginine are directly involved in glycosylase activity (Frankel et al., 1990; Li et al., 1999), which is consistent with their clustering at the active site of the enzyme structure (Fig. 5B). We observed the two tyrosines (Y72, Y123) involved in base-

stacking with the substrate adenine (Monzingo and Robertus, 1992; Gu and Xia, 2000) and W208 with S212 postulated to stabilize the adenine within the active site (Chambery et al., 2007; Shi et al., 2016). G141, I152 (L152 in the consensus sequence) and M173 are buried within the protein suggesting their role in maintaining structural integrity or facilitating minor mobility during catalysis. I117 (L117 in the consensus sequence) is the only surface-exposed amino acid not within the active site, and its function is not clear. We speculate that I177/L177 may contribute by maintaining structure required for protein-RNA interactions. In addition to this amino acid alignment, we also created a sequence logo illustrating the conserved amino acids among all RIP domains in our database (Supplementary Data 7). Extensions at the N and C termini were removed because they represented a large gap in the multiple sequence alignment. This representation allowed for a more detailed comparison among specific amino acids at a given location. There was consistency with regards to the conservation of amino acids described above. Though other conserved amino acids were evident, they were present at frequencies below 70%. Furthermore, many variable regions tended to have an amino acid of similar character suggesting the overall conserved nature of the RIP domain.

A

```
RIP domain consensus   VNFIIYDCAKATSDSYRDFIESLRNRLTGPKHSSHGIPVLPPQSVPQRFFLVELTNGRTS 60
1QCI                   VNTIIYNVGSTTISKYATFLNDLRNEAKDPSLKCYGIPMLPNTNTNPKYVLVELQGSNKK 60
                       ** ***: ..:* ..*  *::.***. ..*. ..:***:**  ..  ::.**** .....

RIP domain consensus   TITLAIRVDNLYLVGFRNAFNDGNERSYWFKDAPATERPDAITNLFPEAPQRTTETLGFG 120
1QCI                   TITLMLRRNNLYVMGYSDPFETNKCRYHIFNDISGTERQDVETTLCPNANSRVSKNINFD 120
                       **** :* :***::*: : *: .: * : *:*  .*** *. *.* *:* .*.::.:.*.

RIP domain consensus   GSYRDLERAAGVGDRENIPLGRQELDDAINALADPSAAAPDAQAKALLVLIQMVSEAARF 180
1QCI                   SRYPTLESKAGVKSRSQVQLGIQILDSNIGKISGVMSFTEKTEAEFLLVAIQMVSEAARF 180
                       . *  **  *** .*.:: ** * **. *. ::.  : : .::*: *** **********

RIP domain consensus   KYIEQTVADGFYGTFNPDPAMLSLENNWGDLSKAILLVKNGVLPKPLELVDASGAKWIVL 240
1QCI                   KYIENQVKTNFNRAFNPNPKVLNLQETWGKISTAIHDAKNGVLPKPLELVDASGAKWIVL 240
                       ****: *  .*  :***:* :*.*::.**.:*.**  .********************

RIP domain consensus   RVDEIKPDVALLNYVGGSCQD- 261
1QCI                   RVDEIKPDVALLNYVGGSCQTT 262
                       ********************
```

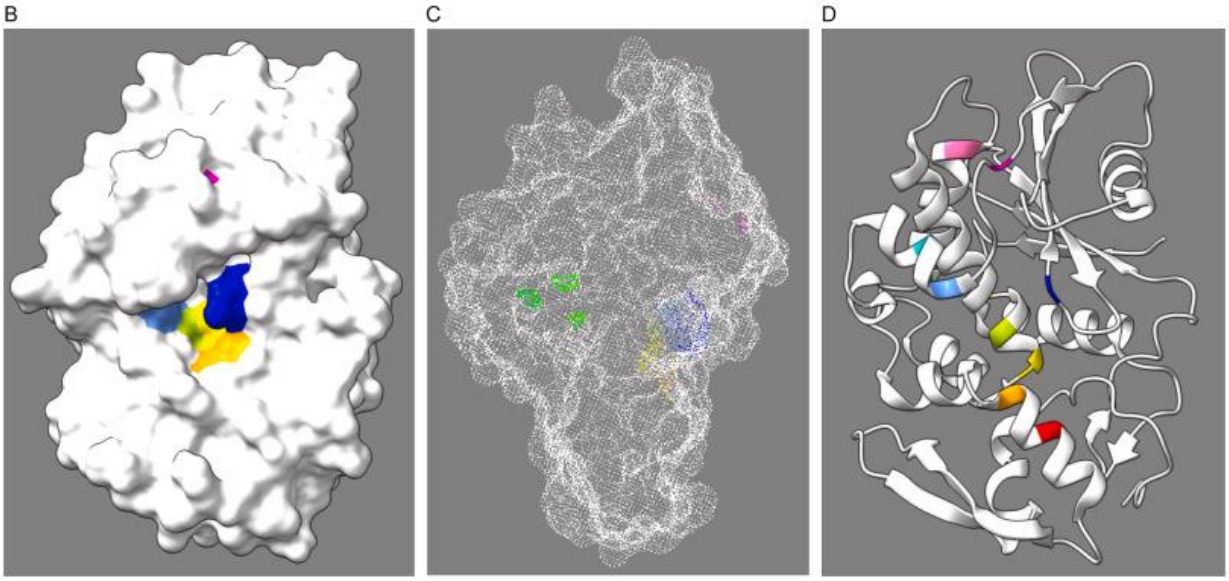


Fig. 5. - The most highly conserved RIP amino acids. Our dataset comprised a curated selection from all the proteins available within NCBI's Conserved Domain Database containing a RIP domain. Colours indicate amino acids conserved in at least 70% of sequences. For the two pink-highlighted proteins, the black bolded amino acids were present in 70% of sequences at that position but were not present in the amino acid sequence of the crystal structure. (A) Sequence alignment. RIP domain consensus: the consensus sequence generated from the multiple sequence alignment in Jalview excluding gaps; 1QCI: the amino acid sequence of pokeweed antiviral protein (protein databank: 1QCI). The third line denotes the similarity in the two sequences as determined by Clustal Omega. (B) The crystal structure of 1QCI visualized in UCSF ChimeraX in surface representation; (C) mesh representation; and (D) cartoon representation. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

To determine whether the addition of non-RIP domains or lack of signal peptides altered the conservation of RIP domain amino acids, the RIP domain sequences of the different protein groups were separated into individual fasta files based on their domain configuration and signal peptide presence or absence. The “other” group was comprised of all RIPs with the addition of non-lectin domains, which were pooled together because these proteins had so few members. Each of these groups underwent multiple sequence alignment. Fig. 6 lists the conserved amino acids within RIP domains, as previously defined, and indicates their presence or absence within each group; amino acid positions with less than three conserved amino acids between the groups were not included. Three amino acids were present in all protein groups (G141, E176, R179); G141 is hypothesized to maintain protein structure, and E176 and R179 are required for base hydrolysis (Frankel et al., 1990; Li et al., 1999). The remaining conserved amino acids identified in Fig. 5A were not conserved at least 70% in all RIP groups. For example, W and S did not meet the 70% threshold in the “other” group. Comparison of the extent of amino acid conservation across RIP groups indicates that noSP-RIP had the least number of conserved amino acids, followed by SP-RIP, suggesting that these groups are the most divergent RIPs. The “other” group also had few conserved amino acids, as was expected for a group comprised of different domains. The four groups of RIPs with RICIN (lectin) domains were more similar to each other than to the other RIP groups, based on the conservation of their amino acids, suggesting that their evolution is more tightly constrained than the noSP-RIP and SP-RIP groups. The amino acid conservation also suggests that the RIPs with RICIN domains likely evolved from a single ancestor, which is consistent with their distribution in the phylogenetic tree.

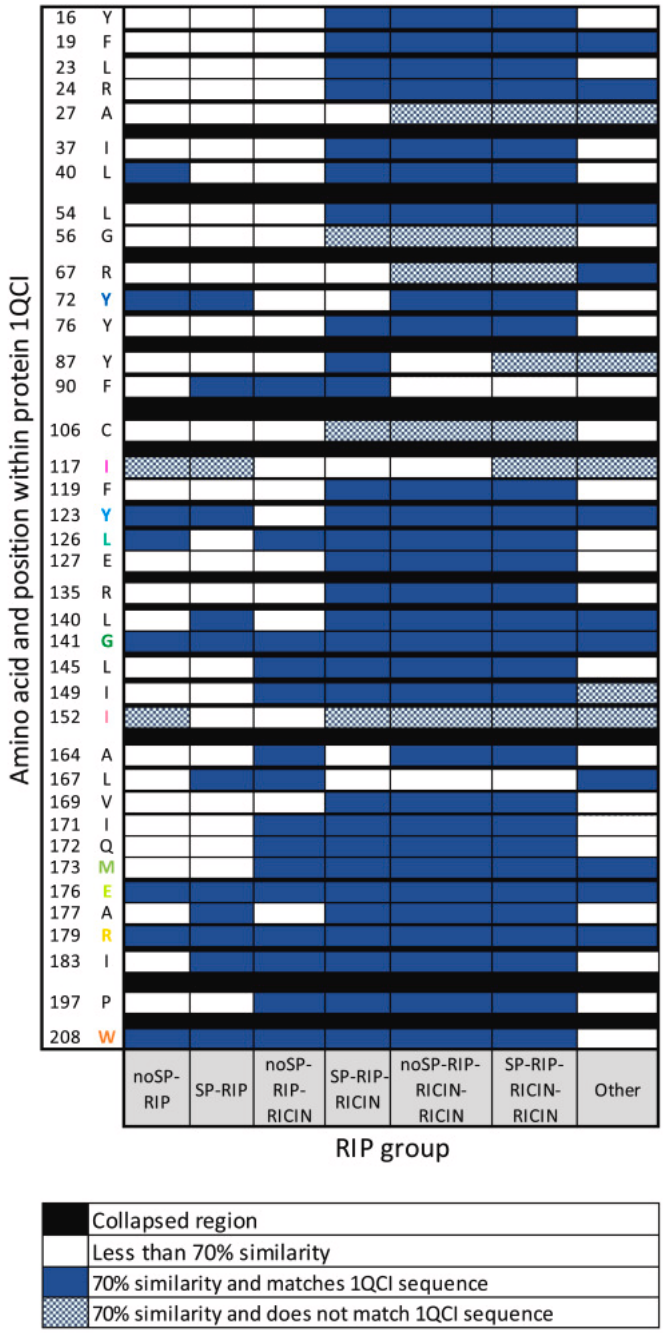


Fig. 6. Tile plot of the most conserved amino acids within RIP domains of each protein group. The Y axis depicts the amino acid and corresponding position within the pokeweed antiviral protein (protein databank: 1QCI), the X axis depicts the proteins groups. For the two pink-highlighted amino acids on the Y axis, different amino acids were present in 70% of sequences at that position but were not present in the amino acid sequence of the crystal structure. The solid blue cells represent amino acids conserved at least 70% within each RIP group and with shared identity to the reference sequence 1QCI. Patterned blue cells represent amino acids conserved at least 70% within RIP groups but without identity to 1QCI. Any amino acid positions with 70% consensus in less than three protein groups were collapsed and shaded black.

Discussion

By conducting an unbiased search of all sequenced plant genomes, we identified several new types of RIPs based on their protein domains. Surprisingly, most are predicted to be nucleocytoplasmic, indicating they may not be toxic to the plants that synthesize them. This, together with their range of calculated physicochemical characteristics, suggests they have different functions. Our phylogenetic tree indicates that of the different types, the noSP-RIP was the initial RIP in plants. Furthermore, the inclusion of signal peptides evolved in tandem with the addition of the lectin domain, which clarifies some disparity among previous evolutionary models. The distribution of RIPs throughout 21 plant orders, with many species expressing more than one type of RIP, indicates a more diverse group of proteins than previously known.

Signal peptides and domain diversity

Upon their synthesis, many RIPs are sequestered to membrane-bound subcellular compartments or the apoplast, presumably to protect ribosomes from depurination (Youle and Huang, 1976; Ready et al., 1986). Contrary to this expectation, our sorting of RIPs based on presence of a signal peptide indicated that most RIPs in plants, including the nine new RIP groups with non-lectin domains, do not have signal peptides. These RIPs may therefore be nucleocytoplasmic with potential access to conspecific ribosomes. Characterization of some RIPs without signal peptides has shown their synthesis as inactive precursors, which may be a mechanism to avoid disabling ribosomes (Walsh et al., 1991; Hey et al., 1995). In addition, there is also variety with regards to their enzymatic activity. The level of ribosome depurination *in vitro* by fully processed maize RIP1, which lacks a signal peptide, was orders of magnitude less

than depurination by single domain RIPs with signal peptides, indicating that some nucleocytoplasmic RIPs may be less toxic than sequestered RIPs (Bass et al., 1992; Hey et al., 1995). Over 60% of the RIP genes we identified did not encode a signal peptide and because our screening included recently sequenced plant genomes, most of these RIPs have not been characterized. The lack of signal peptide required for co-translational insertion into the endoplasmic reticulum lumen does not preclude the possibility that some RIPs without these signals are localized to the extracellular space by unconventional secretion (Krause et al., 2013). Indeed, increasing evidence shows cytosolic proteins that lack signal peptides can gain access to the apoplast and that their extracellular numbers increase during stress (reviewed in Wang et al., 2018; Ruano and Scheuring, 2020). We speculate that some leaderless RIPs may be localized to the apoplast via these unconventional routes; however, further analysis is required to identify their processing and localization. Our finding that the majority of plant RIPs lack a signal peptide significantly extends our understanding of these enzymes beyond sequestered toxins to proteins whose localization may vary with environmental conditions.

We have uncovered considerable diversity among RIPs based on domain number and type. We expected to see single domain RIPs and RIP domains bound to RICIN (lectin) domains, but we also observed gene models with duplicated RIP or other domains. The domain architecture of ricin, a RIP from castor bean (*Ricinus communis*), falls within the SP-RIP-RICIN-RICIN group with a duplicated lectin domain. Examples of some tetrameric RIPs have been described; however, these are the result of disulfide bonds that link proteins post-translationally in a RIP-RICIN-RICIN-RIP conformation rather than genetically encoded duplicated domains (Chandran et al., 2010; Iglesias et al., 2010). In addition to these two main

groups RIP and RIP-RICIN, we documented nine novel types of RIPs, each comprised of domains in addition to the RIP domain. The greatest diversity was found in the order Poales and their lack of signal peptides is a common feature among RIPs produced in cereals (Jiang et al., 2008; Wytynck et al., 2017). Many of these genes were differentially expressed in different plant tissues and under stress conditions, suggesting a range of functions for the RIPs within this single species (Jiang et al., 2008). The variety of domains we identified, ranging from RNA recognition motifs to peptides involved in protein-protein interaction, indicates that in addition to their RNA glycosylase activity, these RIPs have other functions. The additional domains may also modulate or control their glycosylase activity, or indeed they may be released from the RIP domain during processing and maturation of the protein *in vivo*. Such a scenario has been shown for the methyl-jasmonate induced JIP60 from barley, a noSP-RIP (Dunaeva et al., 1999; Rustgi et al., 2014; Przydacz et al., 2020). A recent database search for proteins bearing a RIP domain using JIP60 as the query resulted in less than half of the protein lengths comprised of the RIP domain, suggesting the presence of additional domains (Przydacz et al., 2020). We anticipate that some of the proteins we assigned as noSP-RIP may have additional domains identified in future, as the protein and domain databases become more complete. Given the diversity of RIP groups we present with regards to their physicochemical properties and potential for association with RNAs and proteins, we suggest that they function beyond rRNA depurination and that their activities may be affected by environmental factors, not unlike that of JIP60.

Numbers and groups of RIPs

Our search for RIPs showed that most plants have more than one RIP. While most RIPs to date have been identified through genome and transcriptome data (see Supplementary Data 1 for full list of reference materials for each sequence), in some cases more than one RIP within a single species has also been characterized biochemically. For example, the type III RIPs, b-32 and JIP60, synthesized in *Zea mays* and *Hordeum vulgare*, respectively, were characterized (Walsh et al., 1991; Chaudhry et al., 1994) following the identification of other single domain RIPs within both species (Coleman and Roberts, 1982; Leah et al., 1991). Both single-domain and lectin-containing RIPs are also produced in *Iris hollandica* (Hao et al., 2001). Moreover, 18 different RIPs have been identified in *Ricinis communis*, including four single-domain and six lectin-containing RIP genes that were expressed in unique patterns during seed development (Loss-Morais et al., 2013). Because RIPs are not ubiquitous and therefore not essential for plant survival, it is plausible that this number allows for a greater variability in RIP evolution and thus explains the variety of different forms and functions RIPs have taken in different species. Furthermore, as genome-wide duplication events have given rise to multiple copies of these genes within certain species, particularly in the Poales (Jiang et al., 2008; Di Maro et al., 2014; McKain et al., 2016; Lapadula and Ayub, 2017), maintenance of more than one group of RIP could benefit the plant with a new function or stress adaptation.

RIP phylogeny

A gene tree is distinct from a species tree because it describes the phylogeny of a particular DNA region or gene of interest, where the operational taxonomic units for the phylogeny are the alleles of the gene family of interest rather than populations or species

(Avise, 1989). We chose to represent RIP phylogeny from the perspective of their different domains and the presence or absence of a signal peptide rather than using plant taxonomy as the basis for our phylogenetic tree, as we focus on the evolution of these proteins rather than the plant species containing them. Therefore, our data presentation allows us to visualize each clade as a modification to the sequence, and by extension, structure and function of these diversely represented proteins.

Since the most recent RIP phylogenetic trees (De Zaeytijd and Van Damme, 2017; Lapadula and Ayub, 2017) many more genomes have been sequenced to the chromosome level from a wide variety of plant species (Sun et al., 2021), allowing for a more thorough investigation of plants with RIPs, especially for non-model organisms. With this increased diversity we were able to identify more species in a wider number of orders than identified previously, resulting in a significantly larger dataset of RIPs compared to previous research (Di Maro et al., 2014; Lapadula et al., 2017). Our unbiased method of identifying RIPs also allowed us to detect more unusual and novel sequences. Previous strategies selected existing RIPs either as candidates for BLAST searches (Peumans and Van Damme, 2010; Peumans et al., 2014; De Zaeytijd and Van Damme, 2017), or to construct a matrix to identify homologous proteins (Lapadula et al., 2013, 2017; Lapadula and Ayub, 2017). Both these techniques select for proteins that are sufficiently similar to the candidate list. In contrast, our method searched for candidate RIPs by the presence of an annotated RIP domain within NCBI's conserved domain database, which enabled us to identify novel RIPs.

Authors of RIP phylogeny have proposed various models over time, with consensus in some areas but not in others. It is widely believed that the first RIPs evolved as single-domain

proteins (Lapadula and Ayub, 2017; Peumans et al., 2014), which is supported by our phylogenetic tree showing the ancestral protein is a RIP with no other domains, based on its position relative to the outgroup. There is also consensus that chimeric proteins comprised of a RIP and lectin domain emerged from a fusion event of these domains before the separation of monocots and dicots. However, some propose that this occurred only once (Lapadula et al., 2013; Di Maro et al., 2014; Lapadula and Ayub, 2017) or more than once (Peumans and Van Damme 2010; Peumans et al., 2014; De Zaeytijd and Van Damme, 2017). Our analysis supports the introduction of the lectin domain prior to the separation of monocots and dicots due to its presence in both lineages. Furthermore, our tree concurs with studies by the Ayub group (Lapadula et al., 2013; Di Maro et al., 2014; Lapadula and Ayub, 2017) showing that there was subsequent domain loss in some lineages; this is indicated in our tree as clades of lectin-containing RIPs that are closely related to single-domain RIPs. There is also consensus that a sub-section of RIPs with a single lectin domain emerged from the loss of one copy of this domain, though to what extent has varied depending on the size of the dataset. For example, Peumans et al. (2014) noted that there are instances of proteins with a RIP domain and a non-duplicated lectin domain which they proposed was due to a deletion of one of the duplicated domains. This is supported by our tree as these proteins are closely related to RIPs with duplicated lectin domains rather than being clustered into one isolated clade. Our tree also supports the hypothesis outlined by De Zaeytijd and Van Damme (2017) that the introduction of the signal peptide happened in tandem with the addition of other domains. Therefore, our results clarify several aspects of RIP evolution.

While the scale and diversity of this dataset introduces more uncertainty into our final tree compared to previous trees composed of proteins with greater sequence similarity, our results support and build on this research. Our large dataset also allowed for the discovery of RIPs with additional non-lectin domains outside the family Poaceae, specifically within the Asteraceae that were unlike those found in Poaceae. We are also the first to report the existence of multi-domain proteins other than RIPs with a duplicated lectin, and the first to identify RIPs with a single lectin domain in addition to a RIP domain in the context of a phylogenetic tree. Moreover, we found no evidence of plant RIPs outside of flowering plants which calls into question the hypothesis that RIPs evolved before the separation of Gnetophyta and Magnoliophyta (Peumans and Van Damme, 2010; Peumans et al., 2014). With significantly greater species number and RIP types, we have gained new understanding of how these proteins arose and diverged in different plant taxa.

RIP consensus and conserved amino acids

Differences in the conservation of amino acids within the RIP domains of the 15 protein groups we identified suggest that they may vary regarding activity or substrate interaction. Even though all characterized RIPs depurinate rRNA, they show varying degrees of binding and enzyme activity toward ribosomes from different taxa (Massiah and Hartley, 1995; De Zaeytijd et al., 2019). For example, ricin, a SP-RIP-RICIN-RICIN, accesses ribosomes by binding to stalk proteins characteristic of eukaryotes (May et al., 2012), and has little activity against prokaryotic ribosomes (Endo et al., 1988). By comparison, pokeweed antiviral protein, a SP-RIP, is active against both eukaryotic and prokaryotic ribosomes and binds ribosomal protein L3 (Hudak et al., 1999), which is conserved across taxa. In addition to interaction with ribosomes,

RIPs also differ with regards to their RNA substrates apart from rRNA. Several single domain RIPs have antiviral activity and cleave adenine bases from plant and animal viral RNA (Barbieri et al., 1997; Gandhi et al., 2008; Zhabokritsky et al., 2014). Structural analyses (Kurinov et al., 1999c; Gu and Xia, 2000) and depurination assays (Hudak et al., 2001) of some SP-RIPs have shown guanine as a substrate in addition to adenine. The lack of conservation of W208 and S212 in the non-lectin domain (Other) group was unexpected, given that these amino acids stabilize the ligand inside the active site pocket (Shi et al., 2016). The absence of their conservation suggests that these proteins may substitute functionally equivalent amino acids, or alternatively, they may interact differently with their substrates. The diversity of sequence among the different RIP groups supports the potential for a range of depurination levels from different substrates, expanding the variety of activities of these proteins. Further structural analyses of RIPs bound to ribosomes and rRNA is needed to understand how the conserved amino acids of the different groups contribute to their substrate specificity.

RIPs are characterized largely as defense proteins, though the connection between their activities as glycosylases and their responses to biotic and abiotic stress is not well understood. Our extensive search for RIPs among complete plant genome sequences has uncovered 15 different groups with unique physicochemical properties, suggesting diverse functionality within this family of proteins. Our phylogenetic analysis significantly extends our knowledge of RIP distribution and evolution, showing how the forms arose and diversified across 21 plant orders. This comprehensive study supports further characterization of the different RIPs and their activities, to contribute better understanding of their functions in plants.

Experimental

Data gathering and filtration

All amino acid sequences, within the clade *Viridiplantae,* containing the domain pfam00161 (RIP) (Lu et al., 2019) as identified by the Conserved Protein Domain Architecture Retrieval Tool (Geer et al., 2002; Yang et al., 2020) were downloaded from NCBI's Protein database with Entrez (Sayers et al., 2022) along with their corresponding Genpept files. In addition to this, the amino acid sequences of pokeweed antiviral protein isoforms from the pokeweed genome (Neller et al., 2019) and two outgroups from *Araucaria cunninghamii*, UniProt ID A0A0D6QTL5 and A0A0D6R531, (The UniProt Consortium, 2021) were included.

To filter partial proteins, any sequence name containing the words 'partial', 'chain', 'fragment', or 'truncated' was removed, along with any sequence not starting with a methionine. Similarly, proteins labeled as low quality were removed, and mature peptides were filtered by removing protein names containing 'protein product'. The Genpept files of proteins without a signal peptide were checked manually to ensure they were not mature proteins (missing signal peptides); sequences that were not genomic in origin, labeled as clones from cDNA, or labeled as mature proteins were removed. Duplicate sequences were removed by calculating the pairwise percentage identity of all sequences with the R package Biostrings (v2.62.0; Pagès et al., 2021) and removing one member of any pair with greater than 99% similarity within a single species.

The domain organization was manually tabulated from the annotations in the Genpept files; when annotations were not listed, the structures presented in the Conserved Protein Domain Architecture Retrieval Tool were used instead. The presence of a signal peptide was

predicted from all the amino acid sequences using SignalP 6.0 (Teufel et al., 2022); when results conflicted between the annotations in NCBI and predictions from SignalP, the NCBI annotation was selected. For simplicity, the domains “cl40832: Ricin_B_lectin Superfamily”, “cl40779: RICIN Superfamily”, “cl23784: RICIN Superfamily”, “pfam00652: Ricin_B_lectin”, “pfam14200: RicinB_lectin_2”, and “smart00458: RICIN” were all referred to here as RICIN. Proteins were grouped by the organization of these domains and the presence or absence of a signal peptide; all groupings are summarized in Table 1 while the domain description and identification code within the Conserved Domain Database is available in Supplementary Data 3.

To further validate and characterize the RIPs containing domains other than ‘RIP’ or ‘RICIN’, the amino acid sequences of these proteins were used as input for the InterPro web browser (Jones et al., 2014; Blum et al., 2020) and the domain configuration with the greatest consensus among top hits was manually recorded, along with the InterPro-predicted gene ontology (GO) terms for each sequence (Supplementary Data 4); the GO terms for ‘cellular component’ (CC) were not included because none were predicted by InterPro for these sequences. To determine the cellular localization of proteins based on a metric other than the presence of a signal peptide, ApoplastP (Sperschneider et al., 2017) was used to computationally predict whether the protein is apoplastic based on the amino acid sequence, and the output was imported to R for plotting (Supplementary Data 5) and tabulation with other data (Supplementary Data 1).

Data analysis in R

The fasta file downloaded from NCBI was imported into R with the seqinr package (Charif and Lobry, 2007). The species names were extracted from the fasta sequence names,

and the full NCBI taxonomy was tabulated with the R package Taxize (Chamberlain and Szocs, 2013). For each amino acid sequence, the molecular weight, theoretical net charge, Boman potential protein interaction index, and aliphatic index were calculated using the R package Peptides (Osorio et al., 2015). These data were organized and plotted in R using the tidyverse suite of packages (Wickham et al., 2019).

Phylogenetic tree construction

Multiple sequence alignment was performed with MAFFT (version 7.505; Katoh et al., 2002) using the '--auto' option which allows the program to compute and select the best alignment strategy. The amino acid substitution models were calculated using IQ-TREE (v1.6.12; Nguyen et al., 2015) with ModelFinder (Kalyaanamoorthy et al., 2017) which computes log-likelihood, Akaike information criterion (AIC), corrected Akaike information criterion (AICc), and the Bayesian information criterion (BIC). The calculated best model was WAG + R9 for all three criteria, and therefore this model was selected. Partition models (Chernomor et al., 2016) were also calculated with the single-domain proteins separate from the multi-domain proteins, but ultimately the partition model was not used as it did not improve the quality of the output (data not shown). A maximum-likelihood tree was constructed with IQ-TREE using the model WAG + R9, ultrafast bootstrap approximation (UFBoot) 1000 times (Hoang et al., 2018) and the SH-aLRT test, which is a SH-like approximate likelihood ratio test (Guindon et al., 2010) with 1000 iterations. The resulting tree was visualized with MEGA11 (Tamura et al., 2021), rooted with both outgroups, and colour coded based on RIP group (inner lines) and order (outer ring).

Map to crystal structure and sequence logo

The amino acid sequences of the RIP domains in the six largest groups were separated into fasta files. Because the remaining protein groups had few members, their RIP domain sequences were pooled into a group labeled ‘other’. The amino acid sequence for 1QCI (Kurinov et al., 1999a, 1999b) was added as the first entry for each fasta file. Each of these fasta files underwent multiple sequence alignment with MAAFT using default settings in Jalview (v2.11.1.7, Waterhouse et al., 2009); the process was also repeated with all sequences pooled into one file. All sequences with more than 70% consensus at an amino acid were selected. The crystal structure of 1QCI was visualized with UCSF ChimeraX (Pettersen et al., 2021).
The sequence logo was constructed in R using ggseqlogo (Wagih, 2017) from the same multiple sequence alignment used for comparison of amino acid conservation to the crystal structure of 1QCI.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

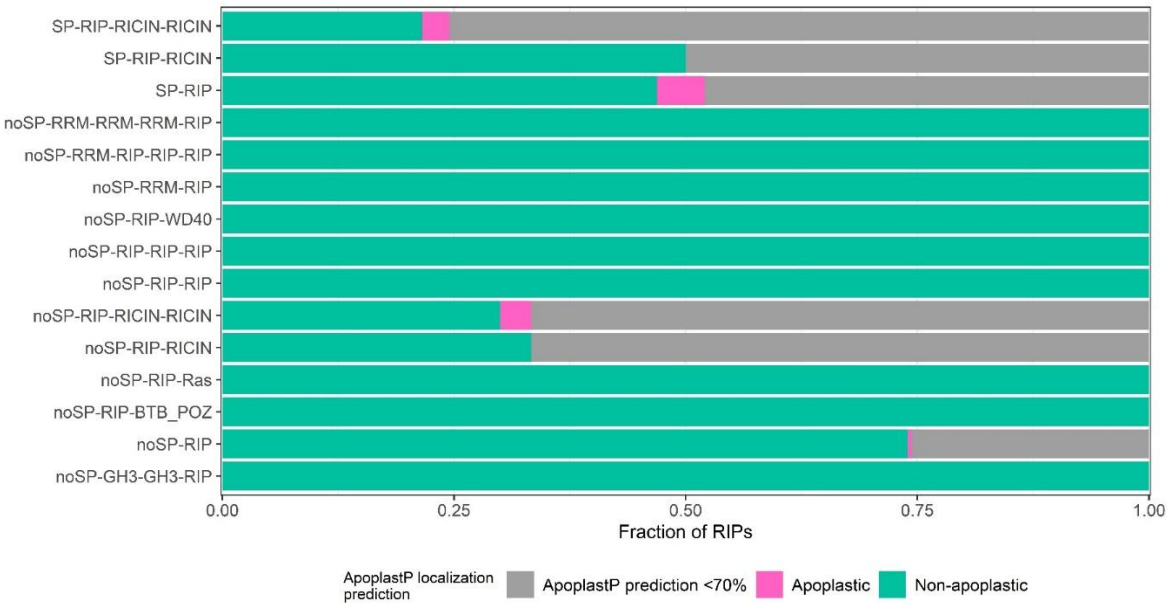
Acknowledgements

FUNDING: This work was supported by a Discovery Grant to K.A.H. from the Natural Sciences and Engineering Research Council of Canada, and a Canada Graduate Scholarship – Master's (CGS M) to K.D.

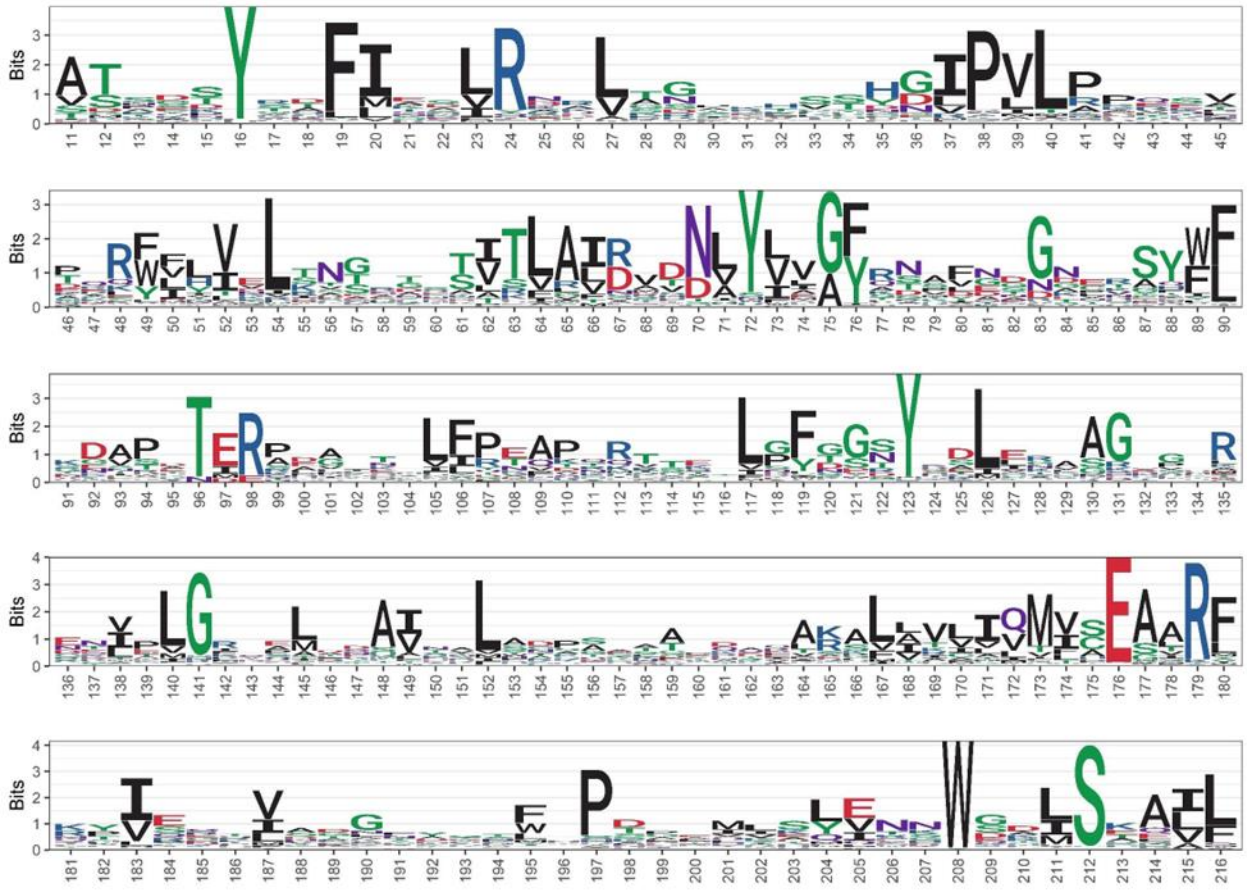
Molecular graphics and analyses were performed with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San

Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases.

Supplementary figures



Supplementary Data 5 – Bar graph of the fraction of RIPs predicted to be apoplastic versus non-apoplastic per RIP group. Our dataset comprised a curated selection from all the proteins available within NCBI's Conserved Domain Database containing a RIP domain. Predictions were made from amino acid sequence data using the ApoplastP web server; any predictions with a probability score lower than 70% were re-labelled as a separate category titled 'ApoplastP prediction <70%'.



Supplementary Data 7 – Sequence logo of the multiple sequence alignment against the amino acid sequence of pokeweed antiviral protein (protein databank: 1QCI). Multiple sequence alignment was generated using MAFFT in Jalview and then gaps introduced into the reference sequence 1QCI were collapsed; all positions with consensus greater than 70% were retained. The first ten amino acids and the last 46 amino acids were also removed as they represented large gaps in the multiple sequence alignment. Our dataset comprised a curated selection from all the proteins available within NCBI's Conserved Domain Database containing a RIP domain.

References

Amblee, V., Jeffery, C.J. 2015. Physical features of intracellular proteins that moonlight on the cell surface. PLoS One 10(6):e0130575. doi: 10.1371/journal.pone.0130575.

Avise, J.C. 1989. Gene trees and organismal histories: a phylogenetic approach to population biology. Evolution 43, 1192-1208. doi: 10.1111/j.1558-5646.1989.tb02568.x.

Barbieri, L., Valbonesi, P., Bonora, E., Gorini, P., Bolognesi, A., Stirpe, F. 1997. Polynucleotide: adenosine glycosidase activity of ribosome-inactivating protein s: effect on DNA, RNA and poly(A). Nucleic Acids Res. 25, 518-522. doi: 10.1093/nar/25.3.518.

Bass, H.W., Webster, C., O'Brian, G.R., Roberts, J.K., Boston, R.S. 1992. A maize ribosome-inactivating protein is controlled by the transcriptional activator Opaque-2. Plant Cell 4, 225–234. doi: 10.1105/tpc.4.2.225.

Bergan, J., Dyve Lingelem, A.B., Simm, R., Skotland, T., Sandvig, K. 2012. Shiga toxins. Toxicon 60, 1085–107. doi: 10.1016/j.toxicon.2012.07.016.

Bjellqvist, B., Hughes, G.J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J.C., Frutiger, S., Hochstrasser D. 1993. The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. Electrophoresis 14, 1023-1031. doi: 10.1002/elps.11501401163.

Blum, M., Chang, H., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G.A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D.H., Letunic, I., Marchler-Bauer, A., Mi, H., Natale, D.A., Necci, M., Orengo, C.A., Pandurangan, A.P., Rivoire, C., Sigrist, C.J.A., Sillitoe, I., Thanki, N., Thomas, P.D., Tosatto, S.C.E., Wu, C.H., Bateman, A. and Finn, R.D. 2020. The InterPro protein families and domains database: 20 years on. Nucleic Acids Research. doi: 10.1093/nar/gkaa977

Boman, H.G. 2003. Antibacterial peptides: basic facts and emerging concepts. J. Intern. Med. 254, 197-215. doi: 10.1046/j.1365-2796.2003.01228.x.

Chamberlain, S., Szocs, E. 2013. taxize - taxonomic search and retrieval in R. F1000Res. 2, 191. https://f1000research.com/articles/2-191/v2

Chambery, A., Pisante, M., Di Maro, A., Di Zazzo, E., Ruvo, M., Costantini, S., Colonna, G., Parente, A. 2007. Invariant Ser211 is involved in the catalysis of PD-L4, type I RIP from Phytolacca dioica leaves. Proteins 67, 209–218. doi: 10.1002/prot.21271.

Chandran, T., Sharma, A., Vijayan, M. 2010. Crystallization and preliminary X-ray studies of a galactose-specific lectin from the seeds of bitter gourd (*Momordica charantia*). Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. 66, 1037-1040. doi: 10.1107/S174430911002659X.

Charif, D., Lobry, J.R. 2007. Seqin{R} 1.0-2: a contributed package to the {R} project for statistical computing devoted to biological sequences retrieval and analysis, in: Bastolla, U., Porto, M., Roman, H.E., Vendruscolo, M. (Eds.), Structural Approaches to Sequence Evolution.

Biological and Medical Physics, Biomedical Engineering. Springer, Berlin, Heidelberg, pp. 207-232. https://doi.org/10.1007/978-3-540-35306-5_10

Chaudhry, B., Müller-Uri, F., Cameron-Mills, V., Gough, S., Simpson, D., Skriver, K., Mundy, J. 1994. The barley 60 kDa jasmonate-induced protein (JIP60) is a novel ribosome-inactivating protein. Plant J. 6, 815-24. doi: 10.1046/j.1365-313x.1994.6060815.x.

Chernomor, O., von Haeseler, A., Minh, B.Q. 2016. Terrace aware data structure for phylogenomic inference from supermatrices. Syst. Biol. 65, 997-1008. https://doi.org/10.1093/sysbio/syw037

Citores, L., Iglesias, R., Ferreras, J.M. 2021. Antiviral activity of ribosome-inactivating proteins. Toxins 13, 80. https://doi.org/ 10.3390/toxins13020080.

Coleman, W.H., Roberts, W.K. 1982. Inhibitors of animal cell-free protein synthesis from grains. Biochim. Biophys. Acta 696, 239-224. doi: 10.1016/0167-4781(82)90053-7.

De Zaeytijd, J., Van Damme, E.J.M. 2017. Extensive evolution of cereal ribosome-inactivating proteins translates into unique structural features, activation mechanisms, and physiological roles. Toxins 9, 123. https://doi.org/10.3390/toxins9040123

De Zaeytijd, J., Rouge, P., Smagghe, G., Van Damme, E.J.M. 2019. Structure and activity of a cytosolic ribosome-inactivating protein from rice. Toxins 11, 325. doi: 10.3390/toxins11060325.

Di Maro, A., Citores, L., Russo, R., Iglesias, R., Ferreras, J.M. 2014. Sequence comparison and phylogenetic analysis by the maximum likelihood method of ribosome-inactivating proteins from angiosperms. Plant Mol. Biol. 85, 575–588. https://doi.org/10.1007/s11103-014-0204-y.

Dunaeva, M., Goebel, C., Wasternack, C., Parthier, B., Goerschen, E. 1999. The jasmonate-induced 60 kDa protein of barley exhibits N-glycosidase activity in vivo. FEBS Lett. 452, 263-266. doi: 10.1016/s0014-5793(99)00645-6.

Endo, Y., Tsurugi, K. 1988. The RNA N-glycosidase activity of ricin A-chain. The characteristics of the enzymatic activity of ricin A-chain with ribosomes and with rRNA. J. Biol. Chem. 263, 8735–8739. PMID: 3288622.

Endo, Y., Tsurugi, K., Lambert, J.M. 1988. The site of action of six different ribosome-inactivating proteins from plants on eukaryotic ribosomes: the RNA N-glycosidase activity of the proteins. Biochem. Biophys. Res. Commun. 150, 1032–1036. doi: 10.1016/0006-291x(88)90733-4.

Fabbrini, M.S., Katayama, M., Nakase, I., Vago, R. 2017. Plant Ribosome-inactivating proteins: Progresses, challenges and biotechnological applications (and a few digressions). Toxins 9, 314. doi: 10.3390/toxins9100314.

Foa-Tomasi, L., Campadelli-Fiume, G., Barbieri, L., Stirpe, F. 1982. Effect of ribosome-inactivating proteins on virus infected cells. Inhibition of virus multiplication and of protein synthesis. Arch. Virol. 71, 323–332. doi: 10.1007/BF01315062.

Frankel, A., Welsh, P., Richardson, J., Robertus, J.D. 1990. Role of arginine 180 and glutamic acid 177 of ricin toxin A chain in enzymatic inactivation of ribosomes. Mol. Cell Biol. 10, 6257-6263. doi: 10.1128/mcb.10.12.6257-6263.1990.

Frigerio, L., Jolliffe, N.A., Di Cola, A., Felipe, D.H., Paris, N., Neuhaus, J.M., Lord, J.M., Ceriotti, A., Roberts, L.M. 2001. The internal propeptide of the ricin precursor carries a sequence-specific determinant for vacuolar sorting. Plant Physiol. 126, 167–175. doi: 10.1104/pp.126.1.167.

Gandhi, R., Manzoor, M., Hudak, K.A. 2008. Depurination of Brome mosaic virus RNA3 in vivo results in translation-dependent accelerated degradation of the viral RNA. J. Biol. Chem. 283, 32218-28. doi: 10.1074/jbc.M803785200

Geer, L.Y., Domrachev, M., Lipman, D.J., Bryant, S.H. 2002. CDART: Protein homology by domain architecture. Genome Res. 12, 1619–1623. https://doi.org/10.1101/gr.278202

Grela, P., Szajwaj, M., Horbowicz-Drożdżal, P., Tchórzewski, M. 2019. How ricin damages the ribosome. Toxins 11, 241. doi: 10.3390/toxins11050241.

Gu, Y.J. Xia , Z.X. 2000. Crystal structures of the complexes of trichosanthin with four substrate analogs and catalytic mechanism of RNA N-glycosidase. Proteins 39, 37-46. PMID: 10737925

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321. https://doi.org/10.1093/sysbio/syq010

Ras proteins: different signals from different locations. Hancock JF. *Nat. Rev. Mol. Cell Biol.* 4, 373-84, (2003). PMID: 12728271

Hao, Q., Van Damme, E.J., Hause, B., Barre, A., Chen, Y., Rougé, P., Peumans, W.J. 2001. Iris bulbs express type 1 and type 2 ribosome-inactivating proteins with unusual properties. Plant Physiol. 125, 866–876. https://doi.org/10.1104/pp.125.2.866

Hey, T.D., Hartley, M., Walsh, T.A. 1995. Maize ribosome-inactivating protein (b-32). Homologs in related species, effects on maize ribosomes, and modulation of activity by pro-peptide deletions. Plant Physiol. 107, 1323–1332. doi: 10.1104/pp.107.4.1323.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., Vinh, L.S. 2018. UFBoot2: Improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35, 518–522. https://doi.org/10.1093/molbev/msx281

Hudak, K.A., Dinman, J.D., Tumer, N.E. 1999. Pokeweed antiviral protein accesses ribosomes by binding to L3. J. Biol. Chem. 274, 3859–3864. doi: 10.1074/jbc.274.6.3859.

Hudak, K.A., Hammell, A.B., Yasenchak, J., Tumer, N.E., Dinman, J.D. 2001. A C-terminal deletion mutant of pokeweed antiviral protein inhibits programmed +1 ribosomal frameshifting and Ty1 retrotransposition without depurinating the sarcin/ricin loop of rRNA. Virology 279, 292-301. doi: 10.1006/viro.2000.0647.

Iglesias, R., Citores, L., Ferreras, J.M., Pérez, Y., Jiménez, P., Gayoso, M.J., Olsnes, S., Tamburino, R., Di Maro, A., Parente, A., Girbés, T. 2010. Sialic acid-binding dwarf elder four-chain lectin

displays nucleic acid N-glycosidase activity. Biochimie 92, 71-80. doi: 10.1016/j.biochi.2009.09.011

Ikai, A. 1980. Thermostability and aliphatic index of globular proteins. J. Biochem. 88, 1895-1898. PMID: 7462208

WD40 Repeat Proteins: Signalling Scaffold with Diverse Functions. Jain BP, Pandey S. *Protein J.* 37, 391-406, (2018). PMID: 30069656

Jiang, S.-Y., Bhalla, R., Ramamoorthy, R., Luan, H.-F., Nori Venkatesh, P., Cai, M., Ramachandran, S. 2012. Over-expression of OSRIP18 increases drought and salt tolerance in transgenic rice plants. Transgenic Res. 21, 785–795. doi: 10.1007/s11248-011-9568-9.

Jiang, S.Y., Ramamoorthy, R., Bhalla, R., Luan, H.-F., Nori Venkatesh, P., Cai, M., Ramachandran, S. 2008. Genome-wide survey of the RIP domain family in Oryza sativa and their expression profiles under various abiotic and biotic stresses. Plant Mol. Biol. 67, 603–614. https://doi.org/10.1007/s11103-008-9342-4

Jones, P., Binns, D., Chang, H., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S., Lopez, R., Hunter, S. 2014. InterProScan 5: genome-scale protein function classification, Bioinformatics, Volume 30, Issue 9 Pages 1236–1240, https://doi.org/10.1093/bioinformatics/btu031

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., Jermiin, L.S. 2017. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587-589. https://doi.org/10.1038/nmeth.4285

Katoh, K., Misawa, K., Kuma, K., Miyata, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066. https://doi.org/10.1093/nar/gkf436

Kawade, K., Masuda, K. 2009. Transcriptional control of two ribosome inactivating protein genes expressed in spinach (Spinacia oleracea) embryos. Plant Physiol. Biochem. 47, 327-334. doi: 10.1016/j.plaphy.2008.12.020.

Krause, C., Richter, S., Knöll, C., Jürgens, G. 2013. Plant secretome-from cellular process to biological activity. Biochim. Biophys. Acta 1834, 2429–2441. doi: 10.1016/j.bbapap.2013.03.024.

Kurinov, I.V., Myers, D.E., Irvin, J.D., Uckun, F.M. 1999a. Low temperature structure of pokeweed antiviral protein complexed with adenine from *Phytolacca americana.* PDB DOI: 10.2210/pdb1qci/pdb

Kurinov, I.V., Myers, D.E., Irvin, J.D., Uckun, F.M. 1999b. X-ray crystallographic analysis of the structural basis for the interactions of pokeweed antiviral protein with its active site inhibitor and ribosomal RNA substrate analogs. Protein Sci. 8, 1765-1772. DOI: 10.1110/ps.8.9.1765

Kurinov, I.V., Rajamohan, E., Venkatachalam, T.K., Uckun, F.M. 1999c. X-ray crystallographic analysis of the structural basis for the interaction of pokeweed antiviral protein with guanine residues of ribosomal RNA. Protein Sci. 8, 2399-2405. doi: 10.1110/ps.8.11.2399.

Lam, S.K., Ng, T.B. 2001. First simultaneous isolation of a ribosome inactivating protein and an antifungal protein from a mushroom (Lyophyllum shimeji) together with evidence for synergism of their antifungal effects. Arch. Biochem. Biophys. 393, 271-280. doi: 10.1006/abbi.2001.2506.

Lapadula, W.J., Ayub, M.J. 2017. Ribosome inactivating proteins from an evolutionary perspective. Toxicon 136, 6-14. https://doi.org/10.1016/j.toxicon.2017.06.012.

Lapadula, W.J., Marcet, P.L., Mascotti, M.L., Sanchez-Puerta, M.V., Ayub, M.J. 2017. Metazoan ribosome inactivating protein encoding genes acquired by horizontal gene transfer. Sci. Rep. 7, 1863. https://doi.org/10.1038/s41598-017-01859-1

Lapadula, W.J., Sánchez Puerta, M.V., Ayub, M.J. 2013. Revising the taxonomic distribution, origin and evolution of ribosome inactivating protein genes. PLoS One, 8(9):e72825. https://doi.org/10.1371/journal.pone.0072825

Leah, R., Tommerup, H., Svendsen, I., Mundy, J. 1991. Biochemical and molecular characterization of three barley seed proteins with antifungal properties. J. Biol. Chem. 266, 1564-1573. PMID: 1899089.

Li, H.G., Xu, S.Z., Wu, S., Yan, L., Li, J.H., Wong, R.N., Shi, Q.L., Dong, Y.C. 1999. Role of Arg163 in the *N*-glycosidase activity of neo-trichosanthin. Protein Eng. 12, 999-1004. doi: 10.1093/protein/12.11.999.

Lord, J.M. 1985. Precursors of ricin and Ricinus communis agglutinin. Glycosylation and processing during synthesis and intracellular transport. Eur. J. Biochem. 146, 411–416. doi: 10.1111/j.1432-1033.1985.tb08667.x.

Loss-Morais, G., Turchetto-Zolet, A.C., Etges, M., Cagliari, A., Körbes, A.P., Maraschin, F., Margis-Pinheiro, M., Margis, R. 2013. Analysis of castor bean ribosome-inactivating proteins and their gene expression during seed development. Genet. Mol. Biol. 36, 74–86. https://doi.org/10.1590/S1415-47572013005000005

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Marchler, G.H., Song, J.S., Thanki, N., Yamashita, R.A., Yang, M., Zhang, D., Zheng, C., Lanczycki, C.J., Marchler-Bauer, A. 2019. CDD/SPARCLE: The conserved domain database in 2020. Nucleic Acids Res. 48(D1):D265-D268. https://doi.org/10.1093/nar/gkz991

Maris, C., Dominguez, C., Allain, F.H.-T. 2005. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J. 272, 2118-2131. doi: 10.1111/j.1742-4658.2005.04653.x.

Massiah, A.J., Hartley, M.R. 1995. Wheat ribosome-inactivating proteins: seed and leaf forms with different specificities and cofactor requirements. Planta 197, 633-640. doi: 10.1007/BF00191571.

May, K.L., Li, X.P., Martinez-Azorin, F., Ballesta, J.P., Grela, P., Tchorzewski, M., Tumer, N.E. 2012. The P1/P2 proteins of the human ribosomal stalk are required for ribosome binding and depurination by ricin in human cells. FEBS J. 279, 3925–3936. doi: 10.1111/j.1742-4658.2012.08752.x.

McKain, M.R., Tang, H., McNeal, J.R., Ayyampalayam, S., Davis, J.I., dePamphilis, C.W., Givnish, T.J., Pires, J.C., Stevenson, D.W., Leebens-Mack, J.H. 2016. A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. Genome Biol. Evol. 8, 1150–1164. https://doi.org/10.1093/gbe/evw060

Montanaro, L., Sperti, S., Mattioli, A., Testoni, G., Stirpe F. 1975. Inhibition by ricin of protein synthesis in vitro. Inhibition of the binding of elongation factor 2 and of adenosine diphosphate-ribosylated elongation factor 2 to ribosomes. Biochem. J. 146, 127–131. doi: 10.1042/bj1460127.

Monzingo, A.F., Robertus, J.D. 1992. X-ray analysis of substrate analogues in the ricin-A chain active site. J. Mol. Biol. 227: 1136-11 45. doi: 10.1016/0022-2836(92)90526-p.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Mol. Biol. Evol. 32, 268-274. https://doi.org/10.1093/molbev/msu300

Neller, K.C.M., Diaz, C.A., Platts, A.E., Hudak, K.A. 2019. De novo assembly of the pokeweed genome provides insight into pokeweed antiviral protein (PAP) gene expression. Front. Plant Sci. 10, 1002. https://doi.org/10.3389/fpls.2019.01002

Osborn, R.W., Hartley, M.R. 1990. Dual effects of the ricin A chain on protein synthesis in rabbit reticulocyte lysate. Inhibition of initiation and translocation. Eur. J. Biochem. 193, 401–407. doi: 10.1111/j.1432-1033.1990.tb19353.x.

Osorio, D., Rondon-Villarreal, P., Torres, R. 2015. Peptides: A package for data mining of antimicrobial peptides. R J. 7, 4-14. doi: 10.32614/RJ-2015-001.

Pagès, H., Aboyoun, P., Gentleman, R., DebRoy, S. 2021. Biostrings: Efficient manipulation of biological strings. R package version 2.62.0. https://bioconductor.org/packages/Biostrings.

Patel, M., Côté, J.-F. 2013. Ras GTPases' interaction with effector domains: Breaking the families' barrier. Commun. Integr. Biol. 6(4):e24298. doi: 10.4161/cib.24298.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., Ferrin, T.E. 2021. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci. 30, 70-82. doi: 10.1002/pro.3943.

Peumans, W.J., Shang, C., Van Damme, E.J.M. 2014. Updated model of the molecular evolution of RIP genes, in: Stirpe, F., Lappi, D.A. (Eds.), Ribosome-inactivating Proteins, ricin and related proteins. John Wiley & Sons, Inc., Oxford, pp. 134-150.

Peumans, W. J., & Van Damme, E. 2010. Evolution of plant ribosome-inactivating proteins, in: Lord, J.M., Hartley, M.R. (Eds.), Toxic plant proteins. Berlin, Germany: Springer, pp. 1-26. https://doi.org/10.1007/978-3-642-12176-0_1

Przydacz, M., Jones, R., Pennington, H.G., Belmans, G., Bruderer, M., Greenhill, R., Salter, T., Wellham, P.A.D., Cota, E., Spanu, P.D. 2020. Mode of action of the catalytic site in the N-terminal ribosome-inactivating domain of JIP60. Plant Physiol. 183, 385-398. doi: 10.1104/pp.19.01029

A common RNA recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNP protein. Query CC, Bentley RC, Keene JD. *Cell* 57, 89-101, (1989). View articlePMID: 2467746

Ready, M.P., Brown, D.T., Robertus, J.D. 1986. Extracellular localization of pokeweed antiviral protein. Proc. Natl. Acad. Sci. USA. 83, 5053–5056. doi: 10.1073/pnas.83.14.5053.

Reinbothe, S., Reinbothe, C., Lehmann, J., Becker, W., Apel, K., Parthier, B. 1994. Jip60, a methyl jasmonate-induced ribosome-inactivating protein involved in plant stress reactions. Proc. Natl. Acad. Sci. USA. 91, 7012–7016. doi: 10.1073/pnas.91.15.7012.

Rustgi, S., Pollmann, S., Buhr, F., Springer, A., Reinbothe, C., von Wettstein, D., Reinbothe, S. 2014. Jip60-mediated, jasmonate- and senescence-induced molecular switch in translation toward stress and defense protein synthesis. Proc. Natl. Acad. Sci. USA. 111, 14181–14186. doi: 10.1073/pnas.1415690111.

A single domain of yeast poly(A)-binding protein is necessary and sufficient for RNA binding and cell viability. Sachs AB, Davis RW, Kornberg RD. *Mol. Cell. Biol.* 7, 3268-76, (1987). View articlePMID: 3313012

Sandvig, K., van Deurs, B. 1994. Endocytosis and intracellular sorting of ricin and shiga toxin. FEBS Lett. 346, 99–102. doi: 10.1016/0014-5793(94)00281-9.

Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., Wang, J., Williams, R., Trawick, B.W., Pruitt, K.D., Sherry, S.T. 2022. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 50(D1):D20-D26. doi: 10.1093/nar/gkab1112.

Shi., W.-W., Nga-Sze Mak, A., Wong, K.-B., Shaw, P.-C. 2016. Structures and ribosomal interaction of ribosome-inactivating proteins. Molecules 21, 1588. doi: 10.3390/molecules21111588.

Sperschneider J et al. (2017) ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning. New Phytologist. doi:10.1111/nph.14946.

Staswick, P.E., Serban, B., Rowe, M., Tiryaki, I., Maldonado, M.T., Maldonado, M.C., Suza, W. 2005. Characterization of an Arabidopsis enzyme family that conjugates amino acids to indole-3-acetic acid. Plant Cell 17, 616-627. doi: 10.1105/tpc.104.026690.

Steeves, R.M., Denton, M.E., Barnard, F.C., Henry, A., Lambert, J.M. 1999. Identification of three oligosaccharide binding sites in ricin. Biochemistry 38, 11677–11685. doi: 10.1021/bi990493o

Stirnimann, C.U., Petsalaki, E., Russell, R.B., Muller, C.W. 2010. WD40 proteins propel cellular networks. Trends Biochem. Sci. 35, 565– 574. DOI: 10.1016/j.tibs.2010.04.003

Stirpe, F. 2004. Ribosome-inactivating proteins. Toxicon 44, 371-383. doi: 10.1016/j.toxicon.2004.05.004.

Stogios, P.J., Downs, G.S., Jauhal, J.J.S., Nandra, S.K., Privé, G.G, 2005. Sequence and structural analysis of BTB domain proteins. Genome Biol. 6, R82. doi: 10.1186/gb-2005-6-10-r82.

Sun, Y., Shang, L., Zhu, Q., Fan, L., Guo, L. 2021. Twenty years of plant genome sequencing: achievements and challenges. Trends Plant Sci. 27, 391-401. doi: 10.1016/j.tplants.2021.10.006.

Tamura, K., Stecher, G., Kumar, S. 2021. MEGA11: Molecular evolutionary genetics analysis version 11. Mol. Biol. Evol. 38, 3022–3027. https://doi.org/10.1093/molbev/msab120

Teufel, F., Almagro Armenteros, J.J., Johansen, A.R., Gislason, M.H., Pihl, S.I., Tsirigos, K.D., Winther, O., Brunak, S., von Heijne, G., Neilsen, H. 2022. SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat. Biotechnol. https://doi.org/10.1038/s41587-021-01156-3

The UniProt Consortium. 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 49, D480–D489, https://doi.org/10.1093/nar/gkaa1100

Van Damme, E.J.M., Hao, Q., Chen Y., Barre, A., Vandenbussche, F., Desmyter, S., Rougé, P., Peumans, W.J. 2001. Ribosome-inactivating proteins: a family of plant proteins that do more than inactivate ribosomes. Crit. Rev. Plant Sci. 20, 395-465. DOI: 10.1080/07352689.2001.10131826

OsJAR1 and OsJAR2 are jasmonyl-L-isoleucine synthases involved in wound- and pathogen-induced jasmonic acid signalling. Wakuta S, Suzuki E, Saburi W, Matsuura H, Nabeta K, Imai R, Matsui H. *Biochem. Biophys. Res. Commun.* 409, 634-9, (2011). View articlePMID: 21619871

Walsh, T.A., Morgan, A.E., Hey, T.D. 1991. Characterization and molecular cloning of a proenzyme form of a ribosome-inactivating protein from maize. Novel mechanism of proenzyme activation by proteolytic removal of a 2.8-kilodalton internal peptide segment. J. Biol. Chem. 266, 23422-23427. https://doi.org/10.1016/S0021-9258(18)54513-0.

Watanabe, K., Kawasaki, T., Sako, N., Funatsu, G. 1997. Actions of pokeweed antiviral protein on virus-infected protoplasts. Biosci. Biotechnol. Biochem. 61, 994–997. doi: 10.1271/bbb.61.994.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., Barton, G.J. 2009. Jalview version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189-1191. doi:10.1093/bioinformatics/btp033

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L, Miller, E., Bache, S.M.,

Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H. 2019. Welcome to the tidyverse. J. Open Source Softw. 4, 1686. https://doi.org/10.21105/joss.01686.

Wytynck, P., Rougé, P., Van Damme, E.J.M. 2017. Genome-wide screening of Oryza sativa ssp. japonica and indica reveals a complex family of proteins with ribosome-inactivating protein domains. Phytochemistry 143, 87-97. doi: 10.1016/j.phytochem.2017.07.009.

Yang, M., Derbyshire, M.K., Yamashita, R.A., Marchler-Bauer, A. 2020. NCBI's conserved domain database and tools for protein domain analysis. Curr. Protoc. Bioinformatics 69(1):e90. doi: 10.1002/cpbi.90.

Youle, R.J., Huang, A.H. 1976. Protein bodies from the endosperm of castor bean: subfractionation, protein components, lectins, and changes during germination. Plant Physiol. 58, 703-709. doi: 10.1104/pp.58.6.703.

Zhabokritsky, A., Mansouri, S., Hudak, K.A. 2014. Pokeweed antiviral protein alters splicing of HIV-1 RNAs, resulting in reduced virus production. RNA 20, 1238–1247. doi: 10.1261/rna.043141.113.

Zhao, W.-L. Feng, D. Wu, J. Sui, S.-F. 2009. Trichosanthin inhibits integration of human immunodeficiency type 1 through depurinating the long-terminal repeats. Mol. Biol. Rep. 37, 2093–2098. doi: 10.1007/s11033-009-9668-2.

Zhu, F., Zhou, Y.K., Ji, Z.L., Chen, X.R. 2018. The plant ribosome-inactivating proteins play important roles in defense against pathogens and insect pest attacks. Front. Plant Sci. 9, 146. doi: 10.3389/fpls.2018.00146

Chapter 3 - Response of pokeweed to jasmonic acid reveals early defense strategies

This chapter is presented as a submitted manuscript by Kyra Dougherty and Katalin A. Hudak
Submitted on Jun 28, 2023, Document ID: TPJ-00841-2023

SUMMARY

Jasmonic acid (JA) is a plant phytohormone involved in regulating responses to biotic and abiotic stress. Although the JA pathway is well characterized in *Arabidopsis thaliana*, less is known about many non-model plants. *Phytolacca americana* (pokeweed) is native to eastern North Americana and is resilient to a variety of environmental conditions. In this work, we have assembled and annotated the pokeweed genome and made it publicly available to serve as a research resource. The genome assembly had an NG50 of ~13.2 Mb and the gene annotations had a minimum 93.9% complete BUSCO score. To investigate the early response of pokeweed to stress, we sprayed leaves with JA and collected samples for transcriptome analysis over a six-hour period. Approximately 5,100 genes were differentially expressed during the time course with almost equal number of up- and down-regulated genes. Cluster and gene ontology analyses indicated the down-regulation of genes associated with photosynthesis and up-regulation of genes involved in secondary metabolite synthesis and hormone signaling. We constructed a gene regulatory network of pokeweed response to JA and integrated transcriptomic data from orthologues of Arabidopsis genes. Similarities between the two plants existed; however, unlike Arabidopsis, pokeweed does not use leaf senescence as a means of reallocating resources during stress. In addition, many secondary metabolite synthesis genes

were constitutively expressed, suggesting that pokeweed directs its resources for survival over the long term compared to Arabidopsis. The assembled pokeweed genome, and the investigation of pokeweed response to JA, provide insights into alternative approaches plants use during defence.

SIGNIFICANCE STATEMENT

Pokeweed (*Phytolacca americana*) is a non-model plant known for its resilience to different environmental conditions and our publicly available genome assembly will serve as a resource for research in plant defense. Analysis of pokeweed gene expression during early response to jasmonic acid illustrates contrasts with expression patterns in Arabidopsis; notably, pokeweed mounts a defense response without sacrificing leaf tissue and constitutively synthesizes a broad range of secondary metabolites which may contribute to its long-term survival.

INTRODUCTION

Jasmonic acid (JA) is a phytohormone that is involved in regulating plant development and adaptation to various environmental conditions by mediating the balance between growth and defense (Chini et al., 2016). JA is involved in response to biotic and abiotic stresses such as drought, salt, pathogens, and insect herbivory (Wasternack and Feussner, 2017) and several excellent reviews detailing JA signaling in Arabidopsis (*Arabidopsis thaliana* L. Heynh.) have been published (Kazan and Manners, 2013; Liu and Timko, 2021; Song et al., 2022). In the absence of JA, activated transcription of genes in the JA pathway is repressed by binding of JASMONATE-ZIM DOMAIN (JAZ) proteins to transcription factors bound to their target promoters. When JA is present, it is converted to its bioactive form (+)-7-iso-JA-isoleucine (JA-Ile) which binds to the CORONATINE INSENSITIVE 1 (COI1) receptor (Chini et al. 2016) prompting formation of the SKP1-CULLIN1-F-box-type (SCF) E3 ubiquitin ligase complex (SCF^{COI1}) complex. This complex ubiquitinates the JAZ proteins marking them for degradation (Ali and Baek, 2020). The ubiquitin-mediated degradation of the JAZ proteins (Thines et al., 2007) de-represses previously bound transcription factors allowing for downstream gene expression. JAZs have been shown to repress several stress-responsive transcription factors (Hu et al., 2013; Zhu et al., 2011; Boter et al., 2015). Most importantly, JAZs repress the MYCs, of which MYC2 is considered the master regulator for the JA pathway (Chini et al., 2016). MYC2 regulates many genes involved in growth and development, biotic and abiotic stress response (Song et al., 2022) and does so by forming a variety of regulatory complexes in Arabidopsis; there are at least 100 proteins that can interact with MYC2 (Chen et al., 2012).

Though the JA response pathway is well characterized in Arabidopsis, less is known about many non-model plants. For example, pokeweed (*Phytolacca americana* L.) is native to eastern North America and is a member of the taxonomic order Caryophyllales. The plant is recognized as a heavy metal hyperaccumulator (Peng et al., 2008; Liu et al., 2010; Zhao et al., 2011) and is broadly resistant to many pathogens (Zoubenko et al., 1997; Lodge et al., 1993; Karran and Hudak, 2008;). Resistance is attributed to the presence of pokeweed antiviral protein, an RNA N-glycosylase that depurinates the sarcin/ricin loop of 28S rRNA, which either limits translation needed for pathogen replication or signals a defense response (Zhabokritsky et al., 2011; Citores et al., 2021). Though these enzymes are present in approximately one-third of flowering plant orders, Arabidopsis does not synthesize an RNA N-glycosylase (Dougherty and Hudak, 2022). Moreover, Arabidopsis is not known for its tolerance to environmental change. For example, older Arabidopsis plants are more susceptible to pathogens, younger plants poorly tolerate temperatures over 25 °C, and plants of all ages can be killed by high or low light intensity (Rivero et al., 2014). Furthermore, Arabidopsis is more sensitive to salt than some other species (van Delden et al., 2020), and induces leaf senescence upon JA application (He et al., 2002). We therefore hypothesize that while pokeweed will respond similarly to stress as Arabidopsis in some ways, differences may contribute to its increased hardiness.

Our previous study showed that the transcriptome of pokeweed is responsive to JA 24 hours after treatment (Neller et al., 2016; 2018; 2019). However, the gene expression patterns in pokeweed within the first six hours following JA exposure, or the early JA response, remains unknown. Furthermore, there are currently no publicly available genome assemblies in the taxonomic family Phytolaccaceae, which hinders further genomics research in pokeweed and

other closely related species. The goal of this study was to produce a quality publicly available pokeweed genome assembly and annotations, and use this resource to determine how the early JA response changes gene expression, with particular focus on genes that precede, and therefore may be involved with, the upregulation of defense. We anticipated that the MYC-based JA-mediated defense regulatory system is conserved in pokeweed; however, changes at the transcript level in the early gene expression patterns will differ from Arabidopsis and will contribute to the long-term survival of pokeweed.

RESULTS

Genome assembly and annotation quality

The pokeweed genome was assembled de novo from approximately 30 Gb of PacBio Hifi long reads using hifiasm and represents the first publicly available genome assembly for this non-model plant. The quality of this assembly was assessed with several metrics. Genomescope2 was used to predict the read error rate, the percentage of genome homozygosity, the haploid genome length, and the percentage of repeat content. BUSCO scores were generated using the orthologue databases ‘eukaryota_odb10’ and ‘eudicots_odb10’ to determine what percentage of the highly conserved coding sequences in the two respective taxonomic groups were present in the genome assembly.

The initial profile generated from the long reads indicated that the genome of our pokeweed sample was 99.8% homozygous which, given that pokeweed is a tetraploid plant, lends confidence to the quality of the final assembly. The BUSCO scores of the genome assembly were high, with 100% complete BUSCOs using the ‘eukaryota_odb10’ database, and 97.4% complete BUSCOs using the ‘eudicots_odb10’ database (Table 1). While the total

assembled genome length changed very little from the previous published assembly with short-read data only (Neller et al., 2019), the contiguity of the assembly improved greatly. The current assembly has a total of 1,058 contigs, with ~98% of all identified genes (33,410 of 34,107) present on the largest 100 contigs. The N50 of this assembly is ~14,000,000 bp, and its longest contig is ~51,000,000 bp. (Table 2).

Table 1 – Quality scores for the primary haplotype-resolved genome assembly using the programs Genomescope2 and BUSCO with two different databases from OrthoDB (eukaryota_odb10 and eudicots_odb10).

Program	**Measure**	**Result**
Genomescope2 initial statistics	Minimum predicted percent homozygous	99.8%
	Minimum predicted repeat content	56.4%
	Minimum predicted haploid genome length	1,181,754,405 bp
BUSCO on eukaryota_odb10 (255 proteins)	Complete single-copy	76.1%
	Complete duplicated	23.9%
	Fragmented	0.0%
	Missing	0.0%
BUSCO on eudicots_odb10 (2326 proteins)	Complete single-copy	90.5%
	Complete duplicated	6.9%
	Fragmented	0.8%
	Missing	1.8%

Table 2 - QUAST output comparing the pokeweed genome assembly based on short reads alone to the assembly based on long reads alone. The short read assembly is from Neller et al., (2019), and all parameters were identical.

Assembly	**Primary contig assembly**	**Fold improvement (x) from assembly by Neller et al. 2019**
Total number of bases in contigs of length >= 0 bp	1,124,849,122 bp	-0.05
Total number of bases in contigs of length >= 1000 bp	1,124,849,122 bp	0.17
Total number of contigs	1,058	800.29
Largest contig	51,110,158 bp	138.37
N50	13,995,428 bp	351.01
NG50	13,284,801 bp	423.41
L50	26	276.31
LG50	29	334.90
N75	8,669,510 bp	506.91
NG75	7,920,507 bp	1984.09
L75	51	325.35
LG75	56	562.61

Several attempts were made to annotate the genome assembly using BRAKER, MAKER, and a combination of the two with different evidence sources. The transcript evidence was the 0-6 hour RNA-seq JA time-course samples described in this study, sequences previously published by our lab (Neller et al., 2019), and sequences downloaded from NCBI (PRJEB21674, One Thousand Plant Transcriptomes Initiative, 2019; PRJNA649785, Zhao et al., 2021; PRJNA623405, Jing et al., 2022; PRJNA669370; PRJNA384358). For protein data, the 'plants' database from OrthoDB (Zdobnov et al., 2021) and all proteins, both canonical and isoforms, available for the taxonomic order Caryophyllales in the UniProt database were used (Boutet et al., 2007, retrieved November 2022). The BUSCO results from the annotation produced by MAKER were lower quality than the BRAKER annotation attempts with an average of 22%

fragmented or missing BUSCOs between both databases searched. This was unexpected as the only input sources were the BRAKER annotations and the Stringtie transcriptome assembly, even though both had higher BUSCO scores. The BRAKER-only version had the highest BUSCO score (93.9% complete BUSCOs with the 'eudicots_odb10' database and 96.5% with the 'eukaryota_odb10' database) and therefore, was selected as the final annotation set (Supporting Figure S1). The annotated genome assembly is available on NCBI under the accession [released upon manuscript acceptance]. Additionally, the Gene Ontology (GO) annotation file that accompanies the structural annotations and was used for GO analysis of the RNA-Seq data is available in Supporting Data S1).

Time-course RNA-Seq, cluster analysis, and GO analysis in pokeweed

Our previous study showed significant changes in the pokeweed transcriptome following 24 hours of jasmonic acid treatment (Neller et al., 2019). To identify the early response of pokeweed to JA, an RNA-Seq time-course experiment was performed for six hours comparing plants sprayed with JA to the control plants sprayed with ethanol. All differential expression analysis results including log 2 fold change (logFC), counts per million (CPM), and false discovery rate (FDR) for each contrast calculated are available in Supporting Data S2. To visualize how many genes were downregulated and upregulated throughout the time course, an upset plot was generated in R. The left bar graph shows how many genes fall within a particular category and the top graph indicates how many of those genes intersect in more than one category. 5133 genes were differentially expressed in at least one time point throughout the time course, and of these a similar number were upregulated compared to downregulated at a particular time point, with an increasing number of differentially expressed genes over

time. When considering the intersection between time points throughout the time course, the largest group of genes (~850) was not differentially expressed at one and two hours but was downregulated at four and six hours. The next largest group (~600) was similarly not differentially expressed at one and two hours then was upregulated at four and six hours. A minority of genes fluctuated between downregulated and upregulated; the largest of these groups (~30 genes) was downregulated at one hour, not differentially expressed at two hours, then upregulated at four and six hours (Figure 1).

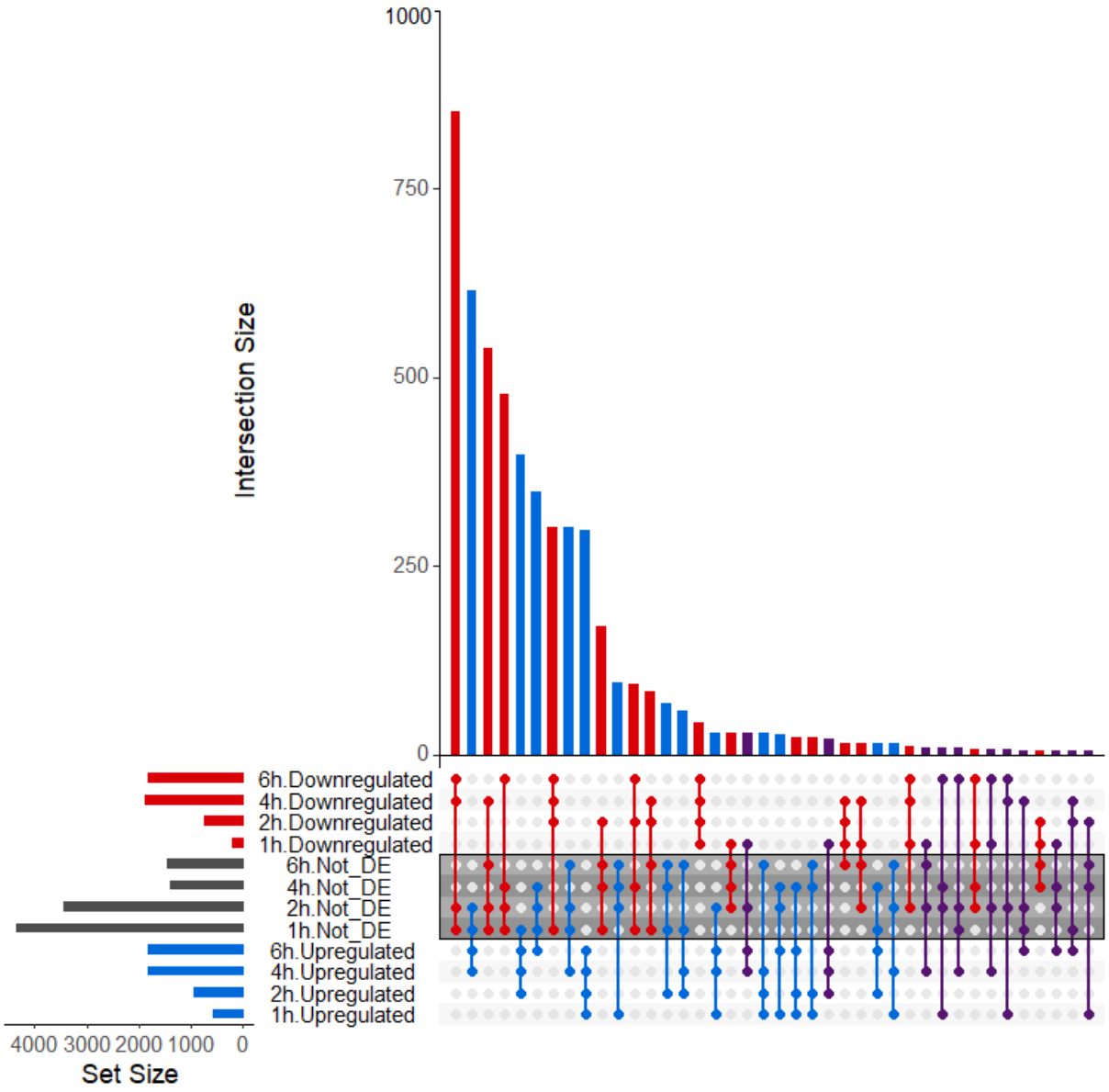


Figure 1 - Upset plot of the number of genes in various combinations of differential expression groups over time (top 40 largest). All time points are specified in hours (h). Genes not differentially expressed (FDR 1%) in all four contrasts were not included. “Downregulated” refers to genes that have a FC < -1.5 for the time point specified. “Upregulated” refers to genes that have a FC > 1.5 for the time point specified. “Not_DE” refers to genes that are not differentially expressed (FDR 1%) for the time point specified.

To further investigate this dataset, the genes identified as differentially expressed in at least one time point were further separated by cluster analysis into four clusters whereby the genes in each cluster generally shared expression patterns. The functional characteristics of these clusters were explored with gene ontology (GO) analysis to determine which terms were differentially enriched in each cluster (Figure 2A). Overall, there were approximately 2600 genes in cluster 1, 1850 genes in cluster 2, 300 genes in cluster 3, and 350 genes in cluster 4. Genes within cluster 1 tended to be downregulated, genes of cluster 2 and 3 were upregulated, and the genes of cluster 4 fluctuated between upregulated and downregulated. Considered together, there was approximately a 1:1 ratio of downregulated compared to upregulated genes among the four clusters (Figure 2B, C). The genes in cluster 1 tended to function in photosynthesis (e.g. chlorophyll biosynthetic process, chloroplast RNA processing, protein targeting chloroplast) and cell division (e.g. regulation of mitotic spindle organization, mitotic chromosome condensation, mitotic spindle assembly) suggestive of growth. Likewise, cluster 4, whose genes fluctuated between downregulated and upregulated though rarely with a logFC larger than +/- 1, tended to involve DNA replication and mitosis (e.g. mitotic DNA replication, nucleosome assembly, regulation of sister chromatid cohesion). By comparison, cluster 2, which had genes that were increasingly upregulated over the six hours and in a pattern inverse to that of cluster 1, contained terms related to biotic and abiotic stress response (e.g. hypoxia, cadmium ion, salt stress, wounding), jasmonic acid (e.g. jasmonic acid metabolic process, oxylipin biosynthetic process, jasmonic acid mediated signaling pathway), and terpene biosynthesis (e.g. farnesyl diphosphate metabolic process, terpene biosynthetic process). The genes of cluster 3, which showed an increasing logFC until four hours after treatment and then

decreased to a mean close to zero, had GO terms enriched in a variety of specific stress responses (e.g. chitin catabolic process, killing cells of another organism, indole glucosinolate biosynthesis). Though the majority of upregulated genes increased in fold change over time, a smaller subset, those of cluster 3, initially increased at a higher rate than cluster 2 following JA application but declined at the four-hour mark. Therefore, there was a coordinated response to jasmonic acid such that photosynthesis genes were generally downregulated whereas genes involved in stress response were upregulated with groups of genes following different expression patterns. All differentially enriched GO terms, along with their GO IDs, respective groups, and statistical metrics are available in Supporting Data S3.

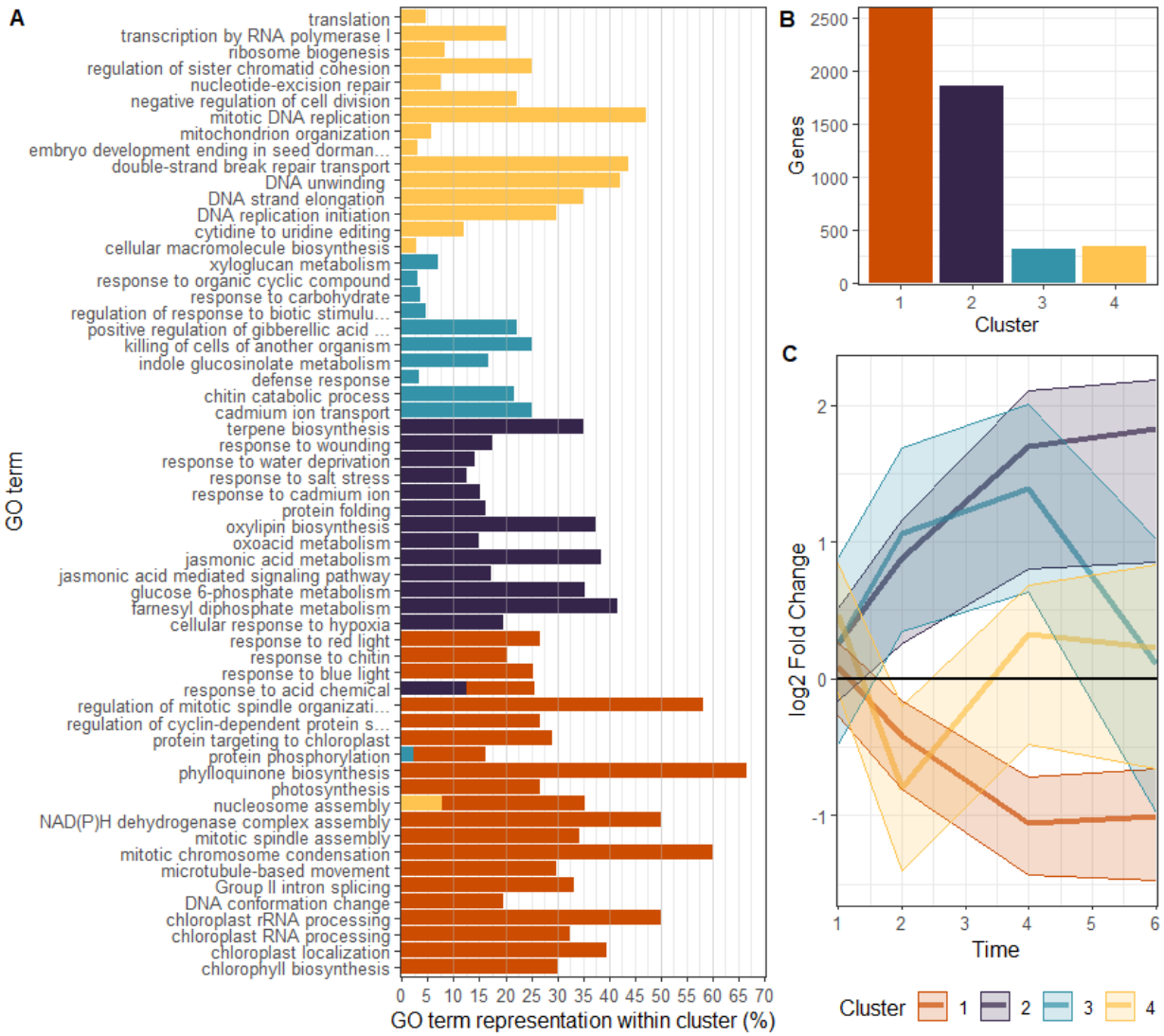


Figure 2 – Cluster analysis and subsequent GO analysis of differentially expressed genes (FDR 1%) in pokeweed during the JA treatment time-course. Hierarchical Euclidean clusters, with four clusters as the target, were generated with the R package dtwclust. Each cluster is represented by a colour identified in the legend. (A) Top significantly differentially enriched GO terms per cluster (p value less than 0.01 using the 'elim', 'weight', and classic methods among the top 25 differentially expressed GO terms). X-axis represents the percentage of each GO term present per cluster compared to all instances of that term, and the y-axis indicates each enriched GO term. (B) Bar graph specifying the number of genes represented per cluster. (C) Ribbon plot of the mean logFC of each cluster, flanked by the interquartile range.

JA gene regulatory network

Many of the downstream effects known to be associated with JA application were reflected in the GO analysis results (Figure 2A). Likewise, many of the products of key genes involved in the JA response pathway that have been identified in Arabidopsis likely have similar physiological effects in pokeweed. To investigate if the gene regulatory patterns observed in Arabidopsis paralleled the expression patterns in pokeweed, a gene regulatory pathway was constructed based on known JA response pathways in Arabidopsis (Kazan and Manners, 2013; Liu and Timko, 2021; Song et al., 2022). Orthologues of these key Arabidopsis genes were identified in pokeweed using a blastp search (Camacho et al., 2009). Only those unique transcripts with the highest percent identity were selected and all those with a minimum E-value of 0.01, minimum bit score of 40, and minimum percent identity of 50 were excluded from consideration. Arrows in Figure 3 represent instances where one gene, or complex of genes, positively regulates another, and the lines ending in a semicircle are instances where one gene negatively regulates another; the white boxes contain the Arabidopsis gene names and the grey boxes contain their associated functions. Alongside this representation of the pathway are heatmaps of the logFC values of pokeweed orthologues of these genes, identified with blastp, at the one-, two-, four-, and six-hour time points, respectively. Among genes with multiple orthologues, only those that were differentially expressed are shown, and in cases where no orthologues were differentially expressed the associated heatmap for that gene depicts a logFC of 0 at all time points (Figure 3). All identified orthologue blast results are available in Supporting Data S4.

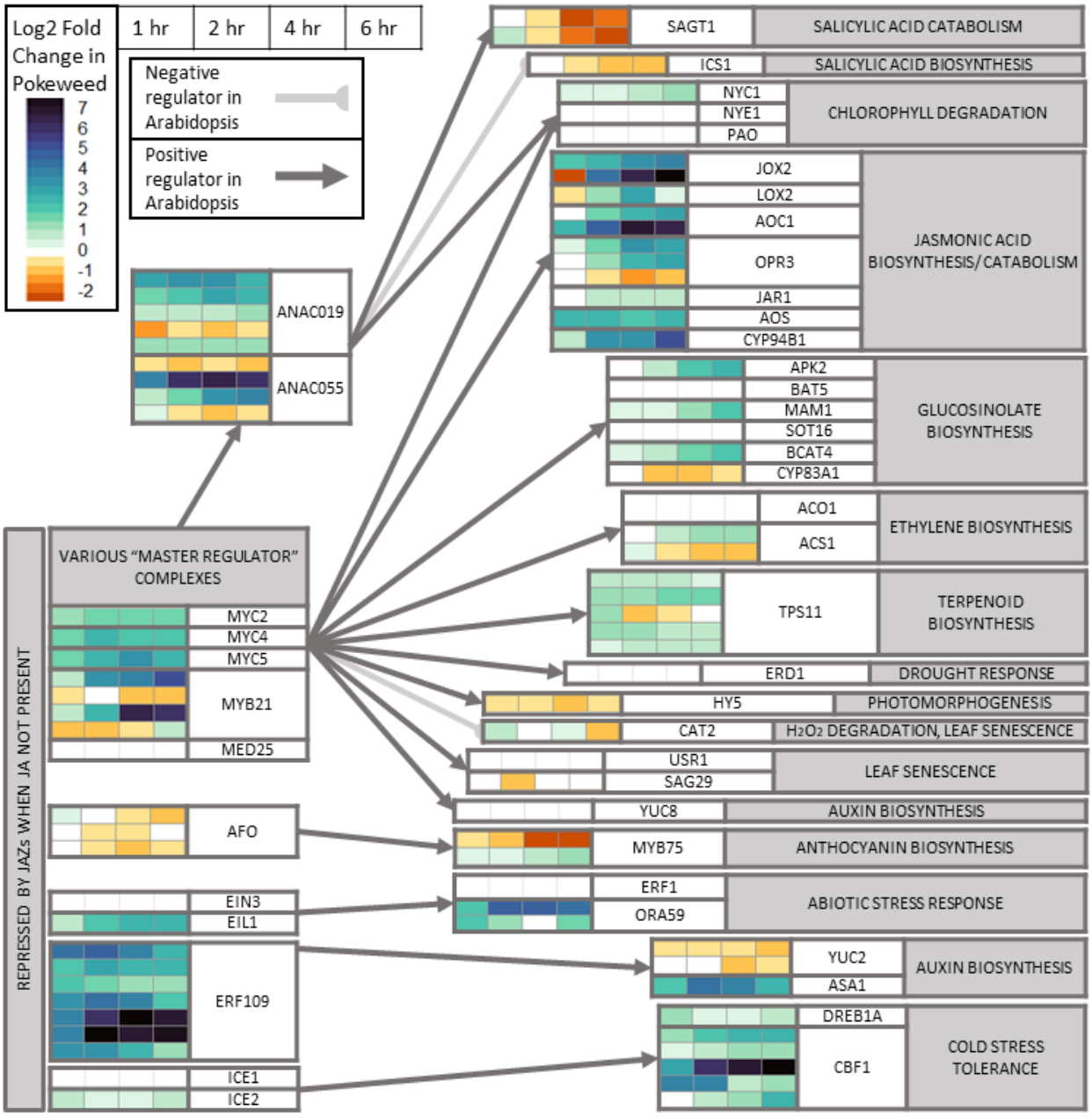


Figure 3 - Known gene regulatory network of Arabidopsis in response to JA with the differential expression patterns of identified pokeweed orthologues. The Araport11 Arabidopsis amino acid sequences for the genes shown were downloaded from TAIR (arabidopsis.org; March 1, 2023) and used as the query sequences in a blastp search against the translated gene sequences from the pokeweed annotations. Genes were considered orthologues if they had an e-value less than 0.01 and percent identity greater than 50. Coloured bars represent the logFC values of each orthologue at the time points 1, 2, 4, and 6 hours (left to right). All orthologues and their associated blast scores are available in Supporting Data S4.

The expression patterns observed in pokeweed for the genes in the JA response pathway were similar in many ways to what has been documented in Arabidopsis. The master regulator of the JA pathway, MYC2, and its relatives MYC4 and MYC5 (Fernández-Calvo et al., 2011) were upregulated at all time points. Likewise, the upstream transcription factors ICE2, EIL1, and ERF109, which are de-repressed by the degradation of JAZs (Kazan and Manners, 2013; Liu and Timko, 2021; Song et al., 2022) were upregulated. These transcription factors control several downstream processes including acclimation to cold (ICE2), regulation of ethylene signaling (EIL1) and cross talk mediation between JA and auxin (ERF109). There were some instances of upstream transcription factor orthologues that were either downregulated or not differentially expressed such as AFO, EIN3, and ICE1, which are involved in flower development, ethylene response and cold tolerance, respectively. Surprisingly there were other cases where some orthologues of upstream transcription factors were strongly upregulated while others for the same gene were strongly downregulated, such as MYB21, ANAC019, ANAC055, and MYB75 (Figure 3). MYB21 and MYB25 are regulators of flavanol and anthocyanin biosynthetic genes, respectively, whereas ANAC019 and ANAC055 are regulators of salicylic acid signaling and its cross talk with JA.

Downstream of these factors, many of their targets were regulated as they would be in Arabidopsis. The orthologues associated with cold stress response (DREB1A, CBF1), JA biosynthesis and catabolism (JOX2, LOX2, AOC1, OPR3, JAR1, AOS, CYP94B1) and terpenoid biosynthesis (TPS11) were upregulated while orthologues associated with SA biosynthesis (ICS1) were downregulated. In other cases, some orthologues were upregulated as expected, such as APK2, MAM1 and BCAT4 involved in glucosinolate biosynthesis, NYC1 involved in chlorophyll

degradation and ASA1 in auxin biosynthesis. Other orthologues were not differentially expressed or were downregulated. For example, genes related to chlorophyll degradation (NYE1, PAO), glucosinolate biosynthesis (BAT5, SOT16, CYP83A1), drought response (ERD1), photomorphogenesis (HY5), leaf senescence (USR1, SAG29), abiotic stress response (ERF1), and auxin biosynthesis (YUC2) were either downregulated or not differentially expressed. Likewise CAT2, involved in metabolism of reactive oxygen species, is normally negatively regulated in Arabidopsis but in pokeweed was upregulated at both the two- and four-hour time points (Figure 3). In sum, orthologues of key genes in the JA response pathway were identified in pokeweed and factors regulating these genes were conserved between pokeweed and Arabidopsis. Some differences existed between the two plants however, with regards to expression levels of genes that control leaf senescence and the extent of secondary metabolite production.

DISCUSSION

The first genome assembly in pokeweed by Neller at al. (2019) contributed greatly to our understanding of the regulation and function of pokeweed genes. However, because the pokeweed genome is more than 50% repetitive regions (Table 1) it was difficult to assemble with short-read data and therefore the result was highly fragmented (Neller et al 2019). The use of PacBio Hifi long read data allowed for the current, and substantially more complete, genome assembly and represents the first publicly available genome assembly for pokeweed. The availability of this genome assembly and annotations will serve as an ongoing resource for future genomics research in pokeweed and, as it is the only publicly available genome assembly in the taxonomic family Phytolaccaceae, other closely related species; presently the species

with a public genome assembly that is most closely related to pokeweed, as they share a taxonomic order, is *Beta vulgaris* (Dohm et al 2013).

Likewise, this is the first time-course RNA-Seq experiment performed in pokeweed and it has revealed numerous insights into the response of pokeweed under stress conditions. We were also interested in comparing the response of pokeweed to JA with Arabidopsis. Two large-scale RNA-Seq time-course experiments measuring effects of methyl jasmonate (MeJA) were completed recently in Arabidopsis, one using gaseous MeJA (Zander et al., 2020) and one spraying liquid MeJA (Hickman et al., 2017). The former identified 7377 differentially expressed genes under MeJA treatment in at least one time point between 0 and 24 hours (Zander et al., 2020); the latter identified 3611 differentially expressed genes in at least one time point within 16 hours of application of MeJA in comparison to an ethanol-sprayed control (Hickman et al., 2017). Our RNA-Seq analysis performed in pokeweed identified approximately 5100 differentially expressed genes (FDR 1%) within six hours of JA treatment (Supporting Data S2), which is similar to that observed in Arabidopsis, despite differences in duration of the time course and specifications for criteria of what was considered a differentially expressed gene in each of the three experiments. Although many general trends were consistent between Arabidopsis and pokeweed, it is clear that they differed in the timing, intensity, and specificity of responses. These trends illustrate that even though both plants respond to JA, differences in their physiology, growth habit and lifespan may contribute to variances in their stress response.

Balance between growth and stress response

Plants under stress generally sacrifice growth to increase defense (Huot et al., 2014; Zander et al., 2020) and GO analysis of differentially expressed genes in pokeweed agree with

these published data. Down-regulated genes in pokeweed were enriched in terms indicating inhibition of photosynthesis and growth (Figure 2A). Analysis of differentially expressed genes in Arabidopsis upon application of MeJA revealed that GO terms relating to growth and development were enriched among downregulated genes (Zander et al., 2020). Similar trends have been found in other species. Following application of JA, growth was reduced in the perennials *Chelidonium majus* (Hashemi et al., 2021), *Hypericum perforatum* (Gadzovska et al., 2007), and *Scrophularia striata* (Sadeghnezhad et al., 2019). With regards to enriched terms associated with upregulated genes in pokeweed, they centered on responses to stress, response to hormones, and secondary metabolite biosynthesis (Figure 2A). Defense-related GO terms were also upregulated, such as flavonoid and terpene biosynthesis genes, in white pine *Picea abies* with application of MeJA (Wilkinson et al., 2022). In Arabidopsis, cluster analysis of differentially expressed genes upon application of MeJA revealed that GO terms such as secondary metabolism, response to wounding, and response to JA were among the upregulated clusters (Zander et al., 2020). Therefore, pokeweed responds to stress by balancing regulation of defense genes with growth and development.

Plants also have means of recycling the products of chlorophyll degradation into useful lipids (Ischebeck et al., 2006) and nitrogen can be re-mobilized from senescing stems and leaves into seeds with high efficiency (Girondé et al., 2015), both of which are beneficial during stress to divert resources to secondary metabolite synthesis and other defense responses. However, leaf senescence and chlorophyll degradation, which are triggered by JA in Arabidopsis (He et al., 2002), appeared to be less regulated by JA in pokeweed. Specifically, none of the top differentially enriched GO terms involved chlorophyll degradation or senescence (Figure 2A),

and orthologues of three primary genes involved in leaf senescence from Arabidopsis did not match their expected expression patterns in pokeweed (Figure 3). In Arabidopsis, USR1 is a ring/U-box transcription factor involved in positively regulating leaf senescence (Zhang et al., 2020b), while SAG29 is a plasma membrane-localized MtN3 protein that accelerates leaf senescence when overexpressed (Seo et al., 2011). USR1 and SAG29 are genes normally upregulated under JA stress but in pokeweed were not differentially expressed aside from the two-hour time point when SAG29 was downregulated. Meanwhile, CAT2 converts the reactive oxygen species H_2O_2 into water and oxygen (von der Mark et al., 2021) which is beneficial for responding to oxidative stress; however, when CAT2 is inhibited by MYC2, leaf senescence can be promoted due to the accumulation of H_2O_2 (Zhang et al., 2020a). Pokeweed showed changing expression of this gene with upregulation seen at one and four hours, no differential expression at two hours, and downregulation at six hours (Figure 3). Together these results suggest that pokeweed relies less on leaf senescence as a defense strategy compared to Arabidopsis. In tandem with this observation, only one of the three main genes involved in chlorophyll degradation known in Arabidopsis, NYC1, was upregulated as expected in pokeweed; the other two, NYE1 and PAO, were not differentially expressed. NYE1 is a Mg-dechelatase involved in chlorophyll degradation (Li et al., 2017), and PAO is involved in converting chlorophyll to colorless nonfluorescent chlorophyll catabolites (Pruzinská et al., 2005). In contrast, NYC1 specifically degrades chlorophyll b, which is beneficial for improving the plant's light-harvesting capability under high-light conditions (Sato et al., 2015). Therefore, it may be that pokeweed is sufficiently able to fine-tune chlorophyll use without resorting to

more generalized chlorophyll degradation and deals with reactive oxygen species as needed through intermittent expression of CAT2 to prevent leaf senescence during stress.

JA gene network and hormone cross talk

This paper represents the fist time that the JA signalling pathway, and its primary associated genes, have been elucidated in pokeweed, and their expression patterns aligned well with their orthologues in Arabidopsis in some cases but less so in others. For instance, the expression patterns of key orthologues in pokeweed consistently aligned with what has been observed in Arabidopsis regarding metabolism of plant hormones JA and salicylic acid (SA), whose signaling antagonism in Arabidopsis has been well established (Thaler et al., 2012; Liu and Timko, 2021). In line with this, pokeweed initially upregulated SAGT1, which converts salicylic acid to an inactive SA glucose conjugate (Thompson et al., 2017), and downregulated ICS1, which generates the precursor to salicylic acid isochorismate (Seguel et al., 2018). Correspondingly, genes involved in JA metabolism were among the most strongly upregulated in the JA pathway (Figure 3). AOS is part of the oxylipin pathway, along with lipoxygenase 2 (LOX2) and allene oxide cyclase (AOC), that is involved in producing JA (He et al., 2002; Bannenberg et al., 2009; Pollmann et al., 2019). OPR3 is then involved in JA synthesis by reducing a certain double bond of the cyclopentenone moiety in 12-oxophytodienoic acid (Maynard et al., 2020). Finally, JAR1 is a jasmonate:amino acid synthetase which is involved in generating jasmonoyl-L-isoleucine (JA-Ile) (Guranowski et al., 2007). Therefore, each major step of the JA biosynthesis pathway was upregulated in pokeweed, from converting lipids to JA precursors, to forming JA itself, and finally converting JA to its active form JA-Ile. With regards to attenuating this JA biosynthesis, JOX2 catalyzes the oxidation of JA to 12OH-JA (Smirnova et

al., 2017), and CYP94B1 is a cytochrome P450 family protein that metabolizes JA-Ile, causing its levels to decrease (Poudel et al., 2016). These expression patterns agree with the GO analysis results (Figure 2A, cluster 2) as terms related to jasmonic acid signaling and biosynthesis were enriched among upregulated genes. This gene pathway indicates that in pokeweed, as in Arabidopsis, application of JA triggers the downregulation of genes involved in SA accumulation and upregulates genes involved in JA accumulation and bioactivity.

Typically, there is considerable cross talk between hormone signalling pathways (Wang et al., 2020; Xu et al., 2020); however, unlike Arabidopsis, key orthologues relating to the biosynthesis of the hormones ethylene and auxin did not tend to be upregulated in pokeweed. Auxin is involved in leaf development (Wang et al., 2011) while ethylene is involved in growth (Munné-Bosch et al., 2018), reproduction, and stress responses (Kieber, 1997) including leaf senescence (Oh et al., 1997). JA and ethylene act synergistically by both regulating key genes involved in the ethylene response pathway (Liu and Timko, 2021). EIN3 and EIL1 are ethylene-responsive genes involved in increasing salt stress tolerance and reducing reactive oxygen species accumulation (Peng et al., 2014). EIL1 was upregulated over the JA time course whereas EIN3, which is involved in inhibition of leaf growth (Munné-Bosch et al., 2018) and leaf senescence (Li et al., 2013), was not differentially expressed in pokeweed. This observation is consistent with the lack of upregulation of primary genes related to senescence.

Genes involved in secondary metabolism

An important response of plants to stress is the production of secondary metabolites. For example, MeJA induces the biosynthesis of a variety of flavonoids and phenylethanoid glycosides in *Scrophularia striata* (Sadeghnezhad et al., 2019), monoterpenes, sesquiterpenes,

and green leaf volatiles in *Polygonum minus* (Rahnamaie-Tajadod et al., 2019), benzophenanthridine alkaloids in *Chelidonium majus* (Hashemi et al., 2021), and isoflavonoids in *Phaseolus vulgaris*, *Glycine max*, and *Vigna radiata* (Gómez et al., 2022). Likewise, after JA application 24 different phenylpropanoids and naphtodianthrones were differentially enriched in *Hypericum perforatum* (Gadzovska et al., 2007) and six phenolics were detected *Fagopyrum esculentum* (Park et al., 2019). Glucosinolates are also required for innate immune response in Arabidopsis (Clay et al., 2009) indicating their more long-term defense capabilities. While the orthologues related to terpenoid and anthocyanin biosynthesis were upregulated as anticipated in pokeweed, expression of genes involved in glucosinolate biosynthesis was inconsistent. Among the orthologues upregulated in pokeweed, APK2 is involved in generating sulfated glucosinolates (Mugford et al., 2009), while MAM1 and BCAT2 are involved in producing glucosinolates from methionine (Kroymann et al., 2001; Schuster et al., 2006). In contrast, BAT5 and SOT16 orthologues were not differentially expressed and CYP83A1 was downregulated beyond the two-hour time point (Figure 3). SOT16 is a sulfotransferase (Klein and Papenbrock, 2009) while BAT5 is involved in aliphatic glucosinolate biosynthesis (Gigolashvili et al., 2009). CYP83A1 is a cytochrome P450 family protein involved in glucosinolate production from methionine (Weis et al., 2014). Genes with GO terms related to indole glucosinolate biosynthesis were enriched in cluster 3, and genes containing GO terms related to the biosynthesis of terpenes and their precursors were enriched in cluster 2 (Figure 2A) indicating that pokeweed is synthesizing glucosinolates, and other secondary metabolites, with other genes. It may be that pokeweed produces a different repertoire of secondary metabolites in response to stress than Arabidopsis which requires a different set of genes. For

example, glucosinolates are known to be produced in pokeweed (Clarke, 2010), pokeberrygenin is a triterpene identified in pokeweed (Kang and Woo, 1980) and phytolaccasaponin B, E and G are the major saponins of pokeweed (Suga et al., 1978). Approximately 85 genes containing a GO term related to glucosinolate, terpenoid, or anthocyanin metabolism were upregulated in at least one time point during the time course, whereas approximately 290 genes containing these GO terms were not upregulated in response to JA. This gene expression pattern suggests that pokeweed may produce more defense-related secondary metabolites than Arabidopsis, and that in many cases their synthesis may be constitutive rather than produced only in response to stress.

The early time course of gene expression changes in response to JA illustrates that pokeweed responds to stress by favouring defense over growth. However, this shift does not come at the expense of leaf senescence or chlorophyll degradation. Rather, pokeweed synthesizes a range of secondary metabolites, many of which may be constitutively expressed suggesting that it allocates resources for survival over the long term compared with Arabidopsis. Pokeweed offers insight into the defense mechanisms of plants beyond those observed in research models and crops, and further study may yield novel approaches to improving the resilience of plants to environmental changes.

EXPERIMENTAL PROCEDURES

Raising plants, sampling and genome sequencing

Pokeweed plants were grown to the 4-5 leaf stage in growth chambers with a 14-hour light/10-hour darkness cycle; chamber lighting was comprised of 75% fluorescent and 25% incandescent bulbs (180 µE/m2/s). Temperature was held at 24°C and 21°C during the light and

darkness periods, respectively, and fan speed was set to 65%. Plants were watered approximately every 2 days and fertilized weekly with NPK 20:20:20 fertilizer.

At the 4-5 leaf stage, leaves were harvested, the mid vein removed, and remaining tissue frozen in liquid N_2. Frozen tissue was shipped on dry ice to Histogenetics (Ossining, NY). They isolated genomic DNA, then performed HiFi standard library preparation with a 15-20kb insert, sequencing with a 30hr movie time on a PacBio Sequel II machine, quality checking, adapter trimming, and HiFi read generation. The sequencing depth was ~25x and the read error rate was 0.1%.

Time-course sample preparation

For the JA time course, 45 pokeweed plants at the four-leaf stage were sprayed with 0.5 mM JA dissolved in 0.5% ethanol (Et). For the control group, another 45 plants were sprayed with 0.5% Et alone. Leaves were harvested and frozen in liquid nitrogen at time points zero, one, two, four, and six hours. To reduce the biological variability within budgetary limitations, two approaches were used. Firstly, three biological replicates were taken per treatment and per time point, and secondly, within each replicate equal proportions of mRNA from three independent plants were pooled, totaling nine plants per time point and per treatment (Supporting Figure S2).

Isolation of total RNA and sequencing

Frozen leaf tissue was ground in liquid N2 with mortar and pestle into a fine powder and total RNA was extracted using organic solvents. Briefly, leaf powder was suspended and vortexed in 1:1 ratio of aqueous buffer and phenol:chloroform:isoamyl alcohol (25:24:1)

equilibrated at pH 5.5. Phases were separated by centrifugation and extraction was repeated once with phenol:chloroform:isoamyl alcohol (25:24:1) and once with chloroform alone. RNA was precipitated in 70% isopropanol, resuspended in water and treated with DNaseI to digest contaminating gDNA. Samples were re-extracted with phenol:chloroform:isoamyl alcohol (25:24:1) and the RNA was precipitated from the aqueous phase in 0.3M NaOAc and 70% ethanol. Following centrifugation, RNA was dissolved in water and quantified using a nanodrop spectrophotometer.

RNA samples were sent to the Centre for Applied Genomics (Toronto, ON) for sequencing. RNA quality was first assessed with a bioanalyzer and mRNA was isolated from each total RNA sample using NEB poly(A) mRNA magnetic isolation module with the NEB Ultra II Directional RNA kit which uses oligo dT beads to capture the mRNA transcripts that have a polyA tail for sequencing. A total of 30 samples were sequenced on a NovaSeq S4 flowcell PE 2x150bp at a depth of ~67-83 million reads per sample.

Genome assembly

Detailed commands used in this section are available on GitHub (https://github.com/kd-lab/Genome_Assembly_Annotation). The quality of the Hifi reads was checked with FastQC (Andrews 2010) and found to be of sufficient quality to proceed without additional adapter trimming. Prior to assembly, a k-mer profile was generated with meryl (Miller et al., 2008) with a k-mer size of 21, and genomescope2 was used to infer genome properties (Ranallo-Benavidez et al., 2020). These results indicated that purging duplicates would be an unnecessary step as this genome was 99.8% homozygous, but that chromosome-level assembly was unlikely because ~55% of the genome length consisted of repeats. To reduce misassemblies because

pokeweed is a tetraploid, a haplotype-resolved assembly was produced with hifiasm (Cheng et al., 2021; Cheng et al., 2022). The alternate assembly was very short and was therefore not used further. The quality of the assembly was checked with BUSCO (Simão et al., 2015), QUAST (Gurevich et al., 2013) using the estimated reference size calculated with genomescope2 (Ranallo-Benavidez et al. 2020), and Merqury (Rhie et al., 2020) using the output from meryl (Miller et al., 2008) generated previously.

Structural gene annotation

Detailed commands used in this section are available on GitHub (https://github.com/kd-lab/Genome_Assembly_Annotation). Repeats were identified with RepeatModeler2 (Flynn et al., 2020) then masked with RepeatMasker (Smit, 2013-2015). This repeat-masked genome assembly file was used as input for all subsequent annotation steps.

Genome annotation was performed with BRAKER2 (Brůna et al., 2021) in two separate runs with default settings, one with protein data only and one with mRNA data only, and then the two were merged into one annotation file as per the recommendations by the BRAKER2 developers. The unaligned protein data used were the 'plants' database from OrthoDB (Zdobnov et al., 2021) and all proteins, both canonical and isoforms, available for the taxonomic order Caryophyllales in the UniProt database (Boutet et al., 2007, retrieved November 2022). The RNA-seq data were from the zero-six hour JA time-course samples described in this study, sequences previously published by our lab (Neller et al., 2019), and sequences downloaded from NCBI (PRJEB21674, One Thousand Plant Transcriptomes Initiative 2019; PRJNA649785, Zhao et al., 2021; PRJNA623405, Jing et al., 2022; PRJNA669370; PRJNA384358). All sequences were spot-checked for quality with FastQC (Andrews, 2010) and any datasets with poor quality

reads were trimmed with Trimmomatic (Bolger et al., 2014) in paired-end mode. The reads were aligned to the reference genome with STAR (Dobin et al., 2013). These files were then sorted with samtools sort using default settings and indexed with samtools index (Danecek et al., 2021).

The program used to merge the two BRAKER2 annotation files was TSEBRA (Gabriel et al., 2021) with default settings. The quality of the annotation was calculated with BUSCO (Simão et al., 2015). Additional annotation files were generated with MAKER (v3.01.04; Campbell et al., 2014) with various settings, and in combinations with the BRAKER2 annotations, but the quality was found to be poorer than the BRAKER2-based annotations so were not used (Supporting Figure S1).

Functional gene annotation

Detailed commands used in this section are available on GitHub (https://github.com/kd-lab/Genome_Assembly_Annotation). The annotation files were converted to transcript sequence files with gffread (Pertea and Pertea, 2020) using default settings and then translated to amino acid sequences with transeq (Madeira et al., 2022). These amino acid fasta sequence files were used as input for the functional annotation prediction programs InterProScan5 (Jones et al., 2014) and eggNOG (Huerta-Cepas et al., 2019). Additionally, blastn (Camacho et al., 2009) was used to identify all high-quality hits for each sequence within the SwissProt database (Boutet et al., 2007) and their associated GO annotations were downloaded from the UniProt database. The gene ontology annotations produced by each of these methods were concatenated, duplicate and erroneous annotations within each gene removed, and the result

exported as a GO annotation file with a custom R script. This GO annotation file is available in Supporting Data S1.

RNA-Seq and differential expression analysis

Detailed commands used in this section are available on GitHub (https://github.com/kd-lab/Genome_Assembly_Annotation). The RNA-seq JA time course reads previously trimmed with Trimmomatic (Bolger et al., 2014) were aligned to the reference genome one sample at a time with STAR (Dobin et al., 2013). The read alignments were sorted first by read name and then by position using default settings with samtools sort (Danecek et al., 2021). Read counting was done with htseq (Putri et al., 2022).

Normalization and differential expression analysis from the counts produced by htseq (Putri et al., 2022) were performed with edgeR (Chen et al., 2016). To reduce false positives, genes with a read count less than 30 were excluded from the analysis, and the trimmed mean of M-values (TMM) method was used for normalization (Robinson and Oshlack, 2010). Raw tagwise estimates, mean dispersion estimates across all genes, and the fitted value of the mean-dispersion trend were calculated using both a linear model and generalized linear model (GLM). To examine likelihood of false discoveries during differential expression analysis for each model, the gene wise biological coefficient of variation (BCV) was plotted against gene abundance for each; the GLM-based approach estimated a slightly larger BCV than the approach based on a linear model so for this reason, and because GLM models are generally more robust than linear models (McCarthy et al., 2012), the GLM-based method was chosen for subsequent analyses. To estimate the differences between replicates and between treatments as a measure of quality control, a multi-dimensional scaling (MDS) plot was generated which

showed, as expected, that there was little difference between biological replicates and increasing difference between treatment groups over the time-course (Supporting Figure S3). Four contrasts were conducted using the formula (JAxh-JA0h)-(Etxh-Et0h) where 'JA' is the jasmonic acid treated group, 'Et' is the control group, 'h' represents hours, and 'x' are the time points 1, 2, 4, and 6. These contrasts were made with the 'glmTreat' function in edgeR with a minimum fold change of 1.5 and the Benjamini-Hochberg method for adjusting p-values. The criteria used to define differentially expressed genes were those that had an adjusted p-value of less than 0.01 in any of the four contrasts.

Cluster analysis and GO analysis

All analyses were performed in R and all code is available on GitHub (https://github.com/kd-lab/Genome_Assembly_Annotation). To get an overview of the number of genes conforming to different upregulation and downregulation patterns over the time-course, an upset plot was generated in R with the UpSetR package (Conway et al. 2017). To ensure that the data conformed to the assumptions of the clustering algorithms, and that outlier genes with unusually high or low fold changes would not impact the effectiveness of clustering, the logFC values from edgeR were centered and scaled using the base R function 'scale' (R Core Team, 2022). Multiple cluster analyses were then performed using various combinations of clustering methods, cluster sizes, distance metrics, and control methods in the package dtwclust (Sarda-Espinosa, 2022). The cluster validity indices (CVIs) Silhouette index (Rousseeuw, 1987), Calinski-Harabasz index, Dunn index, COP index, Davies-Bouldin index (Arbelaitz et al., 2013), Modified Davies-Bouldin index (Kim and Ramakrishna, 2005), and Score Function (Saitta et al., 2007) were calculated for each cluster analysis strategy and the one with

best index/score in multiple CVIs was selected; in this case, hierarchical clustering with Euclidean distance, four clusters, and 'centroid' control performed the best. With the ggplot2 package (Wickham et al., 2019), the number of genes per cluster was plotted as a bar graph, and the average logFC per cluster over time, including interquartile range, was plotted as a ribbon graph.

The GO annotation file generated during functional gene annotation was loaded into R and GO analysis was performed with topGO (Alexa and Rahnenfuhrer, 2022) using Fisher's exact test with the algorithms 'classic', 'elim', and 'weight', excluding GO terms with less than eight annotations in the whole genome, and using the members of each cluster as the test group one at a time. The top 25 enriched terms in each cluster were initially generated, but only those with a p-value of less than 0.01 from all three algorithms were considered enriched within a cluster (Supporting Data S3). These enriched GO terms were plotted as a stacked bar graph with the ggplot2 package (Wickham et al., 2019).

Comparison with Arabidopsis

Detailed commands and R code used in this section are available on GitHub (https://github.com/kd-lab/Genome_Assembly_Annotation). To investigate the early pokeweed JA pathway based on research from Arabidopsis, a detailed JA pathway map was assembled from several published reviews in Arabidopsis (Kazan and Manners, 2013; Liu and Timko, 2021; Song et al., 2022). To obtain the amino acid sequences associated with these genes, the Araport11 representative gene model protein fasta file (Cheng et al., 2017) was downloaded from The Arabidopsis Information Resource (TAIR), (https://www.arabidopsis.org/download/index-

auto.jsp?dir=%2Fdownload_files%2FSequences%2FAraport11_blastsets, March 1, 2023), and the sequences for the genes of interest were selected in R. These sequences were used as the query for blastp (Camacho et al., 2009) against the translated gene models from our pokeweed annotations with a minimum E-value of 0.01, minimum bit score of 40, and minimum percent identity of 50. Candidate orthologues were further refined by removing duplicate pokeweed hits leaving only the one with the highest percent identity and bit score and selecting at most ten hits per Arabidopsis gene. The logFC values of these selected orthologues were plotted as heatmaps with pheatmap (Kolde, 2019) and this information was added to the JA pathway map. The blast results for each Arabidopsis gene and the associated logFC values for each hit are available in Supporting Data S4.

AUTHOR CONTRIBUTIONS

KD and KAH designed the project. KD performed analyses and drafted the manuscript. KD and KAH edited the manuscript. Both authors approved the final version prior to submission.

ACKNOWLEDGEMENTS

This work was supported by a Discovery Grant to K.A.H. from the Natural Sciences and Engineering Research Council of Canada, and a Canada Graduate Scholarship – Master's (CGS M) to K.D.

DATA AVAILABILITY STATEMENT

The annotated genome assembly and all raw sequence data have been submitted to the DDBJ/EMBL/GenBank databases under accession number PRJNA974046.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

SUPPORTING INFORMATION

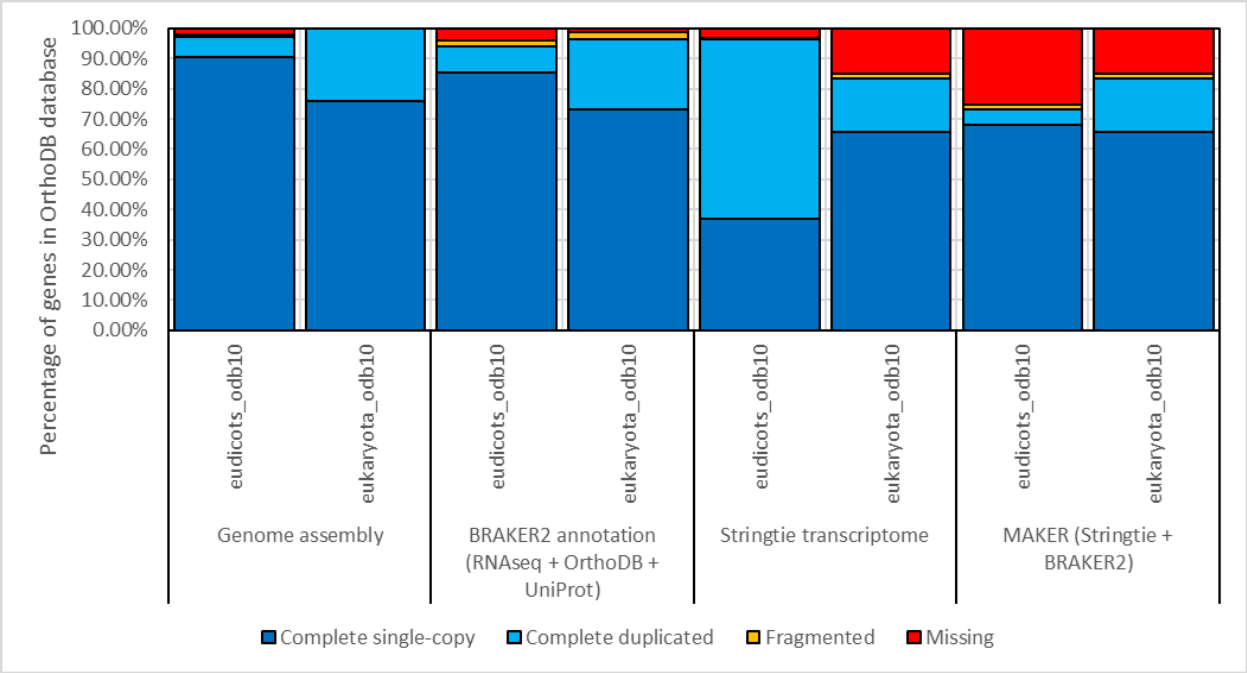


Figure S1. Stacked bar graph of BUSCO scores (from left to right) of pokeweed genome assembly, genome annotation using BRAKER2, reference-guided transcriptome assembly using stringtie, and genome annotation using MAKER. The y-axis represents the percentage of genes from each database, and the colours represent whether the genes were present as a complete single-copy, complete duplicated copy, fragmented copy, or were not present in the tested data.

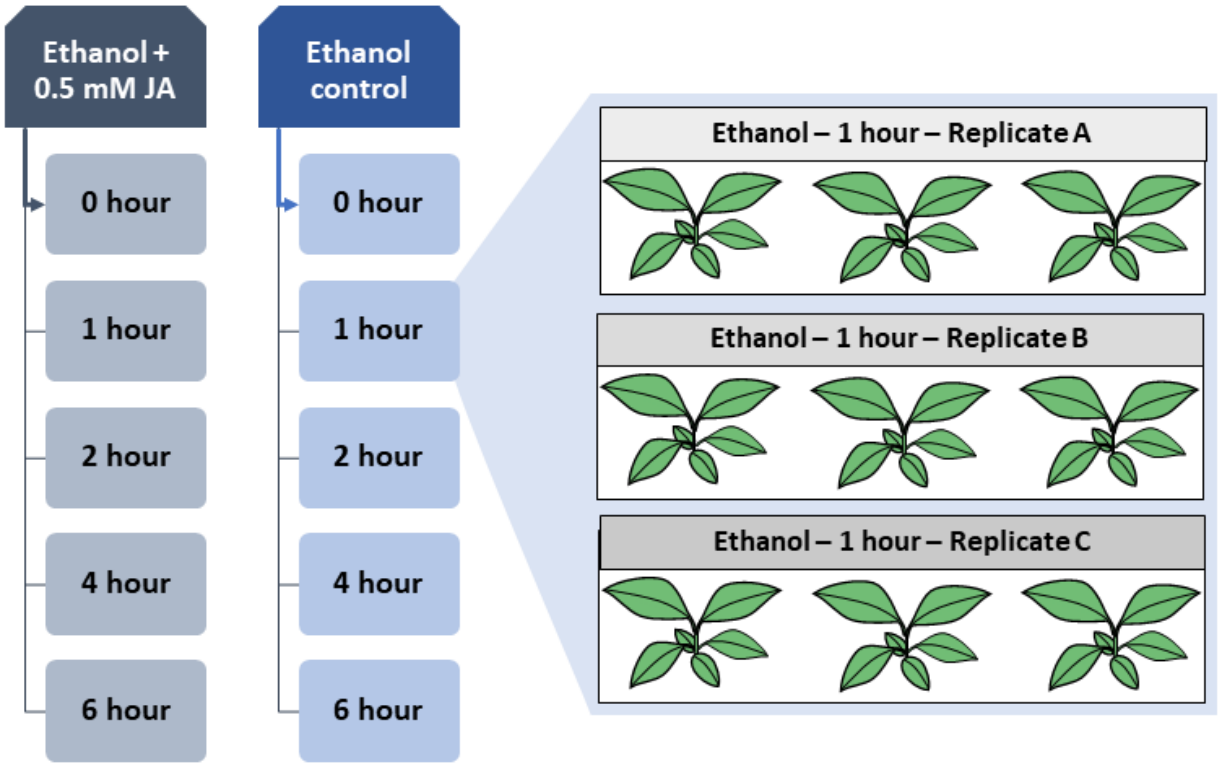


Figure S2. Diagram of experimental design for sample preparation for the JA RNA-Seq time course.

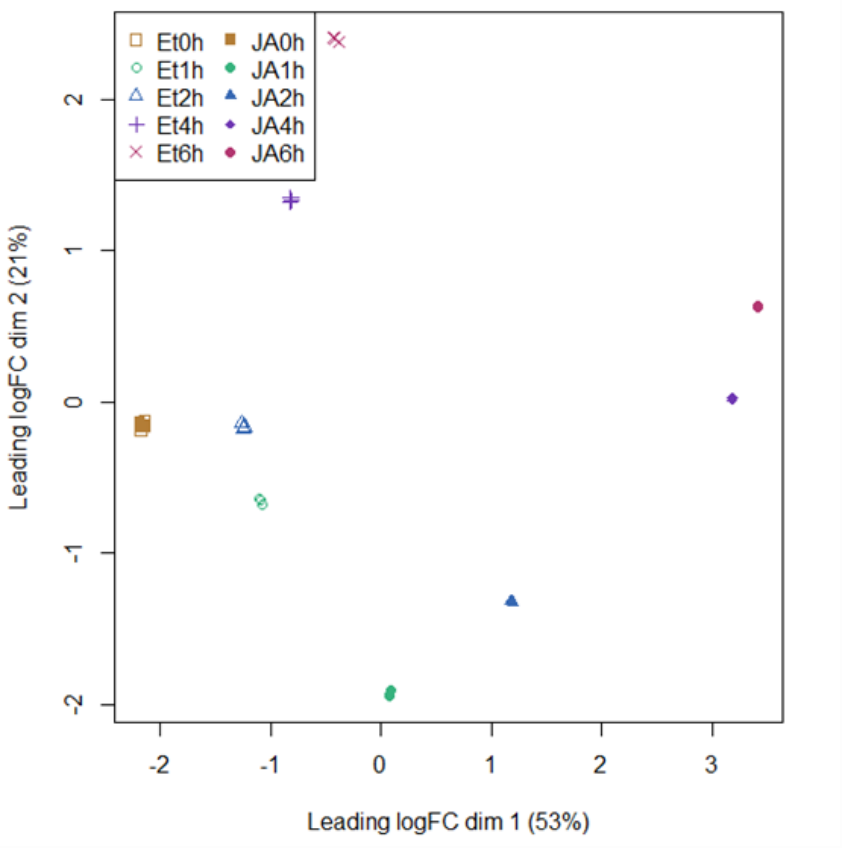


Figure S3. Multi-dimensional scaling plot generated using the edgeR function 'plotMDS'

Data S1. GO annotation file. Column 1 is transcript ID of all annotated pokeweed genes, column 2 is their associated GO terms.

Data S2. All differential expression analysis results including logFC, CPM, and FDR for each contrast calculated.

Data S3. All differentially enriched GO terms, along with their GO IDs, respective groups, and statistical metrics.

Data S4. All identified orthologues, as identified with blastp, and their associated metrics.

REFERENCES

Aeong Oh, S., Park, J. H., In Lee, G., Hee Paek, K., Ki Park, S., &; Gil Nam, H. (1997). Identification of three genetic loci controlling leaf senescence in Arabidopsis thaliana. The Plant Journal, 12(3), 527–535. https://doi.org/10.1046/j.1365-313x.1997.00489.x

Alexa A., Rahnenfuhrer J. (2022). _topGO: Enrichment Analysis for Gene Ontology_. R package version 2.50.0.

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Perez, J. M., Perona, I. (2013). An extensive comparative study of cluster validity indices. Pattern Recognition, 46(1), 243-256.

Bannenberg G., Martínez M., Hamberg M., Castresana C. (2009) Diversity of the enzymatic activity in the lipoxygenase gene family of Arabidopsis thaliana. Lipids.;44(2):85-95. doi: 10.1007/s11745-008-3245-7. Epub 2008 Oct 24. PMID: 18949503.

Boter, M., Golz, J. F., Giménez-Ibañez, S., Fernandez-Barbero, G., Franco-Zorrilla, J. M., Solano, R. (2015). Filamentous flower is a direct target of JAZ3 and modulates responses to jasmonate. The Plant Cell, 27(11), 3160–3174. https://doi.org/10.1105/tpc.15.00220

Bolger A. M., Lohse M., Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30(15):2114-20. doi: 10.1093/bioinformatics/btu170. Epub 2014 Apr 1. PMID: 24695404; PMCID: PMC4103590.

Boutet E., Lieberherr D., Tognolli M., Schneider M., Bairoch A. (2007) UniProtKB/Swiss-Prot. Methods Mol Biol. 406:89-112. doi: 10.1007/978-1-59745-535-0_4. PMID: 18287689.

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., Borodovsky, M. (2021) BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database, NAR Genomics and Bioinformatics, Volume 3, Issue 1, lqaa108, https://doi.org/10.1093/nargab/lqaa108

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T. L. (2009) BLAST+: architecture and applications. BMC Bioinformatics. 10:421. doi: 10.1186/1471-2105-10-421. PMID: 20003500; PMCID: PMC2803857.

Campbell, M. S., Law, M., Holt, C., Stein, J. C., Moghe, G. D., Hufnagel, D. E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C. J., Ware, D., Shiu, S., Childs, K. L., Sun, Y., Jiang, N., Yandell, M., (2014) MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations , Plant Physiology, Volume 164, Issue 2, Pages 513–524, https://doi.org/10.1104/pp.113.230144

Chen, Y., Lun, A. A. T., Smyth, G. .K (2016). “From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline.” F1000Research, 5, 1438. doi: 10.12688/f1000research.8987.2.

Cheng, C. Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., Town, C. D. (2017) Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J. 89(4):789-804. doi: 10.1111/tpj.13415. Epub 2017 Feb 10. PMID: 27862469.

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li H. (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods, 18:170-175. https://doi.org/10.1038/s41592-020-01056-5

Cheng, H., Jarvis, E. D., Fedrigo, O., Koepfli, K. P., Urban, L., Gemmell, N. J., Li, H. (2022) Haplotype-resolved assembly of diploid genomes without parental data. Nature Biotechnology, 40:1332–1335. https://doi.org/10.1038/s41587-022-01261-x

Chini, A., Gimenez-Ibanez, S., Goossens, A., Solano, R. (2016) "Redundancy and specificity in jasmonate signalling" Curr. Opin. Plant Biol. 33, 147–156. doi: 10.1016/j.pbi.2016.07.005

Citores L, Iglesias R, Ferreras JM. (2021) Antiviral Activity of Ribosome-Inactivating Proteins. Toxins (Basel). 13(2):80. doi: 10.3390/toxins13020080.

Clarke, D. B. (2010). Glucosinolates, structures and analysis in food. Analytical Methods, 2(4), 310. https://doi.org/10.1039/b9ay00280d

Clay, N. K., Adio, A. M., Denoux, C., Jander, G., Ausubel, F. M. (2009). Glucosinolate metabolites required for an arabidopsis innate immune response. Science, 323(5910), 95–101. https://doi.org/10.1126/science.1164627

Conway, J. R., Lex, A., Gehlenborg, N. (2017). Upsetr: An R package for the visualization of intersecting sets and their properties. Bioinformatics, 33(18), 2938–2940. https://doi.org/10.1093/bioinformatics/btx364

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., Li H. (2021) Twelve years of SAMtools and BCFtools, GigaScience, Volume 10, Issue 2, giab008, https://doi.org/10.1093/gigascience/giab008

Dobin A., Davis C. A., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M., Gingeras T. R. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29(1):15-21. doi: 10.1093/bioinformatics/bts635. PMID: 23104886; PMCID: PMC3530905.

Dohm, J. C., Minoche, A. E., Holtgräwe, D., Capella-Gutiérrez, S., Zakrzewski, F., Tafer, H., Rupp, O., Sörensen, T. R., Stracke, R., Reinhardt, R., Goesmann, A., Kraft, T., Schulz, B., Stadler, P. F., Schmidt, T., Gabaldón, T., Lehrach, H., Weisshaar, B., Himmelbauer, H. (2013) The genome of the recently domesticated crop plant sugar beet (beta vulgaris). Nature, 505(7484), 546–549. https://doi.org/10.1038/nature12817

Dougherty, K., Hudak, K. A. (2022) Phylogeny and domain architecture of plant ribosome inactivating proteins. Phytochemistry, 202, 113337. https://doi.org/10.1016/j.phytochem.2022.113337

Flynn J. M., Hubley R., Goubert C., Rosen J., Clark A. G., Feschotte C., Smit A. F. (2020) RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl

Acad Sci U S A. 117(17):9451-9457. doi: 10.1073/pnas.1921046117. Epub 2020 Apr 16. PMID: 32300014; PMCID: PMC7196820.

Fernández-Calvo P., Chini A., Fernández-Barbero G., Chico J. M., Gimenez-Ibanez S., Geerinck J., Eeckhout D., Schweizer F., Godoy M., Franco-Zorrilla J. M., Pauwels L., Witters E., Puga M. I., Paz-Ares J., Goossens A., Reymond P., De Jaeger G., Solano R. (2011) The Arabidopsis bHLH transcription factors MYC3 and MYC4 are targets of JAZ repressors and act additively with MYC2 in the activation of jasmonate responses. Plant Cell. 23(2):701-15. doi: 10.1105/tpc.110.080788. PMID: 21335373; PMCID: PMC3077776.

Gadzovska, S., Maury, S., Delaunay, A. et al. (2007) Jasmonic acid elicitation of Hypericum perforatum L. cell suspensions and effects on the production of phenylpropanoids and naphtodianthrones. Plant Cell Tiss Organ Cult 89, 1–13. https://doi.org/10.1007/s11240-007-9203-x

Gigolashvili T., Yatusevich R., Rollwitz I., Humphry M., Gershenzon J., Flügge U. I. (2009) The plastidic bile acid transporter 5 is required for the biosynthesis of methionine-derived glucosinolates in Arabidopsis thaliana. Plant Cell. 21(6):1813-29. doi: 10.1105/tpc.109.066399. PMID: 19542295; PMCID: PMC2714935.

Girondé, A., Etienne, P., Trouverie, J. et al. (2015) The contrasting N management of two oilseed rape genotypes reveals the mechanisms of proteolysis associated with leaf N remobilization and the respective contributions of leaves and stems to N storage and remobilization during seed filling. BMC Plant Biol 15, https://doi.org/10.1186/s12870-015-0437-1

Gómez, K., Quenguan, F., Aristizabal, D., Escobar, G., Quiñones, W., García-Beltrán, O., Durango, D. (2022). Elicitation of isoflavonoids in Colombian edible legume plants with jasmonates and structurally related compounds. Heliyon, 8(2). https://doi.org/10.1016/j.heliyon.2022.e08979

Guranowski A., Miersch O., Staswick P. E., Suza W., Wasternack C. (2007) Substrate specificity and products of side-reactions catalyzed by jasmonate:amino acid synthetase (JAR1). FEBS Lett. 581(5):815-20. doi: 10.1016/j.febslet.2007.01.049. Epub 2007 Feb 2. PMID: 17291501.

Hashemi, S., Naghavi, M., Bakhshandeh, E., Ghorbani, M., Priyanatha, C., Zandi, P. (2021). Effects of abiotic elicitors on expression and accumulation of three candidate benzophenanthridine alkaloids in cultured greater celandine cells. Molecules, 26(5), 1395. https://doi.org/10.3390/molecules26051395

He Y., Fukushige H., Hildebrand D. F., Gan S. (2002) Evidence supporting a role of jasmonic acid in Arabidopsis leaf senescence. Plant Physiol. 128(3):876-84. doi: 10.1104/pp.010843. PMID: 11891244; PMCID: PMC152201.

Hickman R., Van Verk M. C., Van Dijken A. J. H., Mendes M. P., Vroegop-Vos I. A. , Caarls L., Steenbergen M., Van der Nagel I., Wesselink G. J., Jironkin A. , Talbot A., Rhodes J., De Vries M., Schuurink R. C., Denby K., Pieterse C. M. J., Van Wees S. C. M. (2017) Architecture and Dynamics of the Jasmonic Acid Gene Regulatory Network. Plant Cell. 2017 Sep;29(9):2086-2105. doi: 10.1105/tpc.16.00958. Epub 2017 Aug 21. PMID: 28827376; PMCID: PMC5635973.

Hu, Y., Jiang, L., Wang, F., Yu, D. (2013). Jasmonate regulates the inducer of CBF expression–C-repeat binding factor/dre binding FACTOR1 cascade and freezing tolerance in Arabidopsis. The Plant Cell, 25(8), 2907–2924. https://doi.org/10.1105/tpc.113.112631

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., Bork, P. (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses, Nucleic Acids Research, Volume 47, Issue D1, Pages D309–D314, https://doi.org/10.1093/nar/gky1085

Huot, B., Yao, J., Montgomery, B. L., He, S. Y. (2014) Growth–Defense Tradeoffs in Plants: A Balancing Act to Optimize Fitness, Molecular Plant, Volume 7, Issue 8, Pages 1267-1287, ISSN 1674-2052, https://doi.org/10.1093/mp/ssu049.

Ischebeck, T., Zbierzak, A. M., Kanwischer, M., Dörmann, P. (2006). A salvage pathway for Phytol metabolism in Arabidopsis. Journal of Biological Chemistry, 281(5), 2470–2477. https://doi.org/10.1074/jbc.m509222200

Jing M., Zhang H., Wei M., Tang Y., Xia Y., Chen Y., Shen Z., Chen C. (2022) Reactive Oxygen Species Partly Mediate DNA Methylation in Responses to Different Heavy Metals in Pokeweed. Front Plant Sci. 13:845108. doi: 10.3389/fpls.2022.845108. PMID: 35463456; PMCID: PMC9021841.

Jones P., Binns D., Chang H. Y., Fraser M., Li W., McAnulla C., McWilliam H., Maslen J., Mitchell A., Nuka G., Pesseat S., Quinn A. F., Sangrador-Vegas A., Scheremetjew M., Yong S. Y., Lopez R., Hunter S. (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics. 30(9):1236-40. doi: 10.1093/bioinformatics/btu031. Epub 2014 Jan 21. PMID: 24451626; PMCID: PMC3998142.

Kang, S. S., Woo, W. S. (1980). Triterpenes from the berries of Phytolacca Americana. Journal of Natural Products, 43(4), 510–513. https://doi.org/10.1021/np50010a013

Karran, R. A., Hudak, K. A. (2008). Depurination within the intergenic region of Brome mosaic virus RNA3 inhibits viral replication in vitro and in vivo. Nucleic Acids Res. 36, 7230–7239. doi: 10.1093/nar/gkn896

Kazan, K., Manners, J. M. (2013). MYC2: The master in action. Molecular Plant, 6(3), 686–703. https://doi.org/10.1093/mp/sss128

Kieber, J. J. (1997). The ethylene signal transduction pathway in arabidopsis. Journal of Experimental Botany, 48(2), 211–218. https://doi.org/10.1093/jxb/48.2.211

Kim, M., Ramakrishna, R. S. (2005). New indices for cluster validity assessment. Pattern Recognition Letters, 26(15), 2353-2363.

Klein, M., Papenbrock, J. (2009). Kinetics and substrate specificities of desulfo-glucosinolate sulfotransferases in Arabidopsis thaliana. Physiologia Plantarum, 135(2), 140–149. https://doi.org/10.1111/j.1399-3054.2008.01182.x

Kolde R. (2019). _pheatmap: Pretty Heatmaps_. R package version 1.0.12, <https://CRAN.R-project.org/package=pheatmap>.

Kroymann, J., Textor, S., Tokuhisa, J. G., Falk, K. L., Bartram, S., Gershenzon, J., Mitchell-Olds, T. (2001). A gene controlling variation in Arabidopsis glucosinolate composition is part of the methionine chain elongation pathway. Plant Physiology, 127(3), 1077–1088. https://doi.org/10.1104/pp.010416

Li, Z., Peng, J., Wen, X., Guo, H. (2013). Ethylene-insensitive3 is a senescence-associated gene that accelerates age-dependent leaf senescence by directly repressing miR164 transcription in Arabidopsis. The Plant Cell, 25(9), 3311–3328. https://doi.org/10.1105/tpc.113.113340

Li Z., Wu S, Chen J., Wang X., Gao J., Ren G., Kuai B. (2017) NYEs/SGRs-mediated chlorophyll degradation is critical for detoxification during seed maturation in Arabidopsis. Plant J. 92(4):650-661. doi: 10.1111/tpj.13710. Epub 2017 Oct 20. PMID: 28873256.

Liu, X., Peng, K., Wang, A., Lian, C., Shen, Z. (2010). Cadmium accumulation and distribution in populations of Phytolacca americana L. and the role of transpiration. Chemosphere 78, 1136–1141. Doi: 10.1016/j.chemosphere.2009.12.030

Liu, H. Timko, M.P. (2021) Jasmonic Acid Signaling and Molecular Crosstalk with Other Phytohormones. Int. J. Mol. Sci. 22, 2914. https://doi.org/10.3390/ijms22062914

Lodge, J. K., Kaniewski, W. K., Tumer, N. E. (1993). Broad-spectrum virus resistance in transgenic plants expressing pokeweed antiviral protein. Proc. Natl. Acad. Sci. U.S.A. 90, 7089–7093. doi: 10.1073/pnas.90.15.7089

Madeira, F., Pearce, M., Tivey, A. R. N., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A., Lopez, R. (2022) Search and sequence analysis tools services from EMBL-EBI in 2022, Nucleic Acids Research, Volume 50, Issue W1, Pages W276–W279, https://doi.org/10.1093/nar/gkac240

Maynard D., Kumar V., Sproï J., Dietz K. J. (2020) 12-Oxophytodienoic Acid Reductase 3 (OPR3) Functions as NADPH-Dependent α,β-Ketoalkene Reductase in Detoxification and Monodehydroascorbate Reductase in Redox Homeostasis. Plant Cell Physiol. 61(3):584-595. doi: 10.1093/pcp/pcz226. PMID: 31834385.

McCarthy, D. J., Chen, Y., Smyth, G. K. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation, Nucleic Acids Research, Volume 40, Issue 10, Pages 4288–4297, https://doi.org/10.1093/nar/gks042

Miller, J. R., Delcher, A. L., Koren, S., Venter, E., Walenz, B. P., Brownley, A., Johnson, J., Li, K., Mobarry, C., Sutton, G. (2008) Aggressive assembly of pyrosequencing reads with mates, Bioinformatics, vol. 24 (pg. 2818-2824).0

Mugford S. G., Yoshimoto N., Reichelt M., Wirtz M., Hill L., Mugford S. T., Nakazato Y., Noji M., Takahashi H., Kramell R., Gigolashvili T., Flügge U. I., Wasternack C., Gershenzon J., Hell R., Saito K., Kopriva S. (2009) Disruption of adenosine-5'-phosphosulfate kinase in Arabidopsis reduces

levels of sulfated secondary metabolites. Plant Cell. 21(3):910-27. doi: 10.1105/tpc.109.065581. PMID: 19304933; PMCID: PMC2671714.

Munné-Bosch, S., Simancas, B., Müller, M. (2018). Ethylene signaling cross-talk with other hormones in Arabidopsis thaliana exposed to contrasting phosphate availability: Differential effects in roots, leaves and fruits. Journal of plant physiology, 226, 114–122. https://doi.org/10.1016/j.jplph.2018.04.017

Neller, K. C. M., Klenov, A., Guzman, J. C. Hudak, K. A. (2018) Integration of the Pokeweed miRNA and mRNA Transcriptomes Reveals Targeting of Jasmonic Acid-Responsive Genes. Front. Plant Sci. 9:589. doi: 10.3389/fpls.2018.00589

Neller, K. C. M., Klenov, A., Hudak K. A. (2016) The Pokeweed Leaf mRNA Transcriptome and Its Regulation by Jasmonic Acid. Front. Plant Sci. 7:283. doi: 10.3389/fpls.2016.00283

Neller, K. C. M., Diaz, C. A., Platts, A. E., Hudak K. A. (2019) De novo Assembly of the Pokeweed Genome Provides Insight Into Pokeweed Antiviral Protein (PAP) Gene Expression. Front. Plant Sci. 10:1002. doi: 10.3389/fpls.2019.01002

Oh, S. A., Park, J. H., Lee, G. I., Paek, K. H., Park, S. K., Nam, H. G. (1997). Identification of three genetic loci controlling leaf senescence in Arabidopsis thaliana. The Plant Journal, 12(3), 527–535. https://doi.org/10.1046/j.1365-313x.1997.00489.x

One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. Nature 574, 679–685 (2019). https://doi.org/10.1038/s41586-019-1693-2

Park, C. H., Yeo, H. J., Park, Y. E., Chun, S. W., Chung, Y. S., Lee, S. Y., Park, S. U. (2019). Influence of chitosan, salicylic acid and jasmonic acid on phenylpropanoid accumulation in germinated buckwheat (Fagopyrum esculentum moench). Foods, 8(5), 153. https://doi.org/10.3390/foods8050153

Peng, K., Luo, C., You, W., Lian, C., Li, X., and Shen, Z. (2008). Manganese uptake and interactions with cadmium in the hyperaccumulator-Phytolacca americana L. J. Hazard. Mater. 154, 674–681. doi: 10.1016/j.jhazmat.2007.10.080

Peng, J., Li, Z., Wen, X., Li, W., Shi, H., Yang, L., Zhu, H., Guo, H. (2014). Salt-induced stabilization of EIN3/EIL1 confers salinity tolerance by deterring ROS accumulation in Arabidopsis. PLoS Genetics, 10(10), e1004664. https://doi.org/10.1371/journal.pgen.1004664

Pertea G, Pertea M. (2020) GFF Utilities: GffRead and GffCompare. F1000Res. 9:ISCB Comm J-304. doi: 10.12688/f1000research.23297.2. PMID: 32489650; PMCID: PMC7222033.

Pollmann S., Springer A., Rustgi S., von Wettstein D., Kang C., Reinbothe C., Reinbothe S. (2019) Substrate channeling in oxylipin biosynthesis through a protein complex in the plastid envelope of Arabidopsis thaliana. J Exp Bot. 70(5):1483-1495. doi: 10.1093/jxb/erz015. PMID: 30690555; PMCID: PMC6411374.

Poudel, A. N., Zhang, T., Kwasniewski, M., Nakabayashi, R., Saito, K., Koo, A. J. (2016). Mutations in jasmonoyl-L-isoleucine-12-hydroxylases suppress multiple JA-dependent wound responses in Arabidopsis thaliana. Biochimica et biophysica acta, 1861(9 Pt B), 1396–1408. https://doi.org/10.1016/j.bbalip.2016.03.006

Pruzinská A., Tanner G., Aubry S., Anders I., Moser S., Müller T., Ongania K. H., Kräutler B., Youn J. Y., Liljegren S. J., Hörtensteiner S. (2005) Chlorophyll breakdown in senescent Arabidopsis leaves. Characterization of chlorophyll catabolites and of chlorophyll catabolic enzymes involved in the degreening reaction. Plant Physiol. 139(1):52-63. doi: 10.1104/pp.105.065870. Epub 2005 Aug 19. PMID: 16113212; PMCID: PMC1203357.

Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E., Zanini, F. (2022) Analysing high-throughput sequencing data in Python with HTSeq 2.0, Bioinformatics, Volume 38, Issue 10, Pages 2943–2945, https://doi.org/10.1093/bioinformatics/btac166

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rahnamaie-Tajadod, R., Goh, H.-H., Mohd Noor, N. (2019). Methyl jasmonate-induced compositional changes of volatile organic compounds in Polygonum minus leaves. Journal of Plant Physiology, 240, 152994. https://doi.org/10.1016/j.jplph.2019.152994

Ranallo-Benavidez, T. R., Jaron, K. S., Schatz, M.C. (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun 11, 1432 https://doi.org/10.1038/s41467-020-14998-3

Rhie, A., Walenz, B.P., Koren, S. et al. (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol 21, 245. https://doi.org/10.1186/s13059-020-02134-9

Rivero, L., Scholl, R., Holomuzki, N., Crist, D., Grotewold, E., Brkljacic, J. (2014). Handling Arabidopsis Plants: Growth, Preservation of Seeds, Transformation, and Genetic Crosses. In: Sanchez-Serrano, J., Salinas, J. (eds) Arabidopsis Protocols. Methods in Molecular Biology, vol 1062. Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-62703-580-4_1

Robinson, MD, Oshlack, A (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 11, R25.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

Sadeghnezhad, E., Sharifi, M., Zare-Maivan, H., Ahmadian Chashmi, N. (2019). Time-dependent behavior of phenylpropanoid pathway in response to methyl jasmonate in scrophularia striata cell cultures. Plant Cell Reports, 39(2), 227–243. https://doi.org/10.1007/s00299-019-02486-y

Sato R., Ito H., Tanaka A. (2015) Chlorophyll b degradation by chlorophyll b reductase under high-light conditions. Photosynth Res. 126(2-3):249-59. doi: 10.1007/s11120-015-0145-6. PMID: 25896488.

Saitta, S., Raphael, B., Smith, I. F. (2007). A bounded index for cluster validity. In International Workshop on Machine Learning and Data Mining in Pattern Recognition (pp. 174-187). Springer Berlin Heidelberg.

Sarda-Espinosa A (2022). _dtwclust: Time Series Clustering Along with Optimizations for the Dynamic Time Warping Distance_. R package version 5.5.11, <https://CRAN.R-project.org/package=dtwclust>.

Schuster, J., Knill, T., Reichelt, M., Gershenzon, J., Binder, S. (2006). Branched-chain aminotransferase4 is part of the chain elongation pathway in the biosynthesis of methionine-derived glucosinolates in Arabidopsis. The Plant Cell, 18(10), 2664–2679. https://doi.org/10.1105/tpc.105.039339

Seguel, A., Jelenska, J., Herrera-Vásquez, A., Marr, S. K., Joyce, M. B., Gagesch, K. R., Shakoor, N., Jiang, S. C., Fonseca, A., Wildermuth, M. C., Greenberg, J. T., Holuigue L. (2018) PROHIBITIN3 Forms Complexes with ISOCHORISMATE SYNTHASE1 to Regulate Stress-Induced Salicylic Acid Biosynthesis in Arabidopsis. Plant Physiol. 176(3):2515-2531. doi: 10.1104/pp.17.00941. PMID: 29438088; PMCID: PMC5841719.

Seo, P. J., Park, J. M., Kang, S. K., Kim, S. G., Park, C. M. (2011). An Arabidopsis senescence-associated protein SAG29 regulates cell viability under high salinity. Planta, 233(1), 189–200. https://doi.org/10.1007/s00425-010-1293-8

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., Zdobnov, E. M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics, Volume 31, Issue 19, Pages 3210–3212, https://doi.org/10.1093/bioinformatics/btv351

Smirnova E., Marquis V., Poirier L., Aubert Y., Zumsteg J., Ménard R., Miesch L., Heitz T. (2017) Jasmonic Acid Oxidase 2 Hydroxylates Jasmonic Acid and Represses Basal Defense and Resistance Responses against Botrytis cinerea Infection. Mol Plant. 10(9):1159-1173. doi: 10.1016/j.molp.2017.07.010. PMID: 28760569.

Suga, Y. Maruyama, Y., Kawanishi, S., Shoji, J. (1978). Studies on the constituents of phytolaccaceous plants. i. on the structures of phytolaccasaponin B, E and G from the roots of Phytolacca americana L. Chemical and Pharmaceutical Bulletin, 26(2), 520–525. https://doi.org/10.1248/cpb.26.520

Smit, A. F. A., Hubley, R., Green, P. RepeatMasker Open-4.0. 2013-2015 http://www.repeatmasker.org

Song, C., Cao, Y., Dai, J., Li, G., Manzoor, M. A., Chen, C., Deng, H. (2022). The multifaceted roles of MYC2 in plants: Toward transcriptional reprogramming and stress tolerance by jasmonate signaling. Frontiers in Plant Science, 13. https://doi.org/10.3389/fpls.2022.868874

Thaler, J. S., Humphrey, P. T., Whiteman, N. K. (2012). Evolution of jasmonate and salicylate signal crosstalk. Trends in Plant Science, 17(5), 260–270. https://doi.org/10.1016/j.tplants.2012.02.010

Thompson A. M. G., Iancu C. V., Neet K. E., Dean J. V., Choe J. Y. (2017) Differences in salicylic acid glucose conjugations by UGT74F1 and UGT74F2 from Arabidopsis thaliana. Sci Rep. 7:46629. doi: 10.1038/srep46629. PMID: 28425481; PMCID: PMC5397973.

van Delden, S.H., Nazarideljou, M.J. Marcelis, L.F.M. (2020) Nutrient solutions for Arabidopsis thaliana: a study on nutrient solution composition in hydroponics systems. Plant Methods 16, 72. https://doi.org/10.1186/s13007-020-00606-4

von der Mark, C., Ivanov, R., Eutebach, M., Maurino, V. G., Bauer, P., Brumbarova, T. (2021). Reactive oxygen species coordinate the transcriptional responses to iron availability in Arabidopsis. Journal of experimental botany, 72(6), 2181–2195. https://doi.org/10.1093/jxb/eraa522

Wang, W., Xu, B., Wang, H., Li, J., Huang, H., Xu, L. (2011). YUCCA genes are expressed in response to leaf adaxial-abaxial juxtaposition and are required for leaf margin development. Plant Physiology, 157(4), 1805–1819. https://doi.org/10.1104/pp.111.186395

Wang, J., Song, L., Gong, X., Xu, J., & Li, M. (2020). Functions of Jasmonic Acid in Plant Regulation and Response to Abiotic Stress. International journal of molecular sciences, 21(4), 1446. https://doi.org/10.3390/ijms21041446

Wasternack, C., Feussner, I. (2017) The oxylipin pathways: biochemistry and function. Annu. Rev. Plant Biol. 69, 1–24. doi: 10.1146/annurev-arplant-042817-040440

Weis, C., Hildebrandt, U., Hoffmann, T., Hemetsberger, C., Pfeilmeier, S., König, C., Schwab, W., Eichmann, R., Hückelhoven, R. (2014). CYP83A1 is required for metabolic compatibility of Arabidopsis with the adapted powdery mildew fungus Erysiphe cruciferarum. The New Phytologist, 202(4), 1310–1319. https://doi.org/10.1111/nph.12759

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. doi:10.21105/joss.01686.

Wilkinson, S. W., Dalen, L. S., Skrautvol, T. O., Ton, J., Krokene, P., Mageroy, M. H. (2022). Transcriptomic changes during the establishment of long-term methyl jasmonate-induced resistance in Norway spruce. Plant, Cell Environment, 45(6), 1891–1913. https://doi.org/10.1111/pce.14320

Xu, P., Zhao, P. X., Cai, X. T., Mao, J. L., Miao, Z. Q., Xiang C. B. (2020) Integration of Jasmonic Acid and Ethylene Into Auxin Signaling in Root Development. Front. Plant Sci. 11:271. doi: 10.3389/fpls.2020.00271

Zander, M., Lewsey, M. G., Clark, N. M. , Yin, L. , Bartlett, A., Guzmán, J. P. S., Hann, E., Langford, A. E., Jow, B., Wise, A., Nery, J. R., Chen, H., Bar-Joseph, Z., Walley, J. W., Solano, R., Ecker, J. R. (2020) Integrated multi-omics framework of the plant response to jasmonic acid. Nat Plants.

6(3):290-302. doi: 10.1038/s41477-020-0605-7. Erratum in: Nat Plants. 2020 Aug;6(8):1065. PMID: 32170290; PMCID: PMC7094030.

Zdobnov, E. M., Kuznetsov, D., Tegenfeldt, F., Manni, M., Berkeley, M., Kriventseva, E. V. (2021) OrthoDB in 2020: evolutionary and functional annotations of orthologs, Nucleic Acids Research, Volume 49, Issue D1, Pages D389–D393, https://doi.org/10.1093/nar/gkaa1009

Zhabokritsky, A., Kutky, M., Burns, L. A., Karran, R. A., Hudak, K. A. (2011) RNA toxins: mediators of stress adaptation and pathogen defense. Wiley Interdiscip Rev RNA. 2(6):890-903. doi:10.1002/wrna.99.

Zhang, Y., Ji, T. T., Li, T. T., Tian, Y. Y., Wang, L. F., Liu, W. C. (2020a). Jasmonic acid promotes leaf senescence through MYC2-mediated repression of CATALASE2 expression in Arabidopsis. Plant Science, 299, 110604. https://doi.org/10.1016/j.plantsci.2020.110604

Zhang, Z., Xu, M., Guo, Y. (2020b). Ring/U-box protein atusr1 functions in promoting leaf senescence through ja signaling pathway in Arabidopsis. Frontiers in Plant Science, 11. https://doi.org/10.3389/fpls.2020.608589

Zhao, L., Sun, Y., Le, Cui, S. X., Chen, M., Yang, H. M., Liu, H. M., et al. (2011). Cd-induced changes in leaf proteome of the hyperaccumulator plant Phytolacca americana. Chemosphere 85, 56–66. doi: 10.1016/j.chemosphere.2011.06.029

Zhao L., Zhu, Y. H., Wang, M., Ma, L. G., Han, Y. G., Zhang, M. J., Li, X. C., Feng, W. S., Zheng, X. K. (2021) Comparative transcriptome analysis of the hyperaccumulator plant Phytolacca americana in response to cadmium stress. 3 Biotech. 11(7):327. doi: 10.1007/s13205-021-02865-x. PMID: 34194911; PMCID: PMC8197689.

Zhu, Z., An, F., Feng, Y., Li, P., Xue, L., A, M., Jiang, Z., Kim, J. M., To, T. K., Li, W., Zhang, X., Yu, Q., Dong, Z., Chen, W.-Q., Seki, M., Zhou, J.-M., Guo, H. (2011). Derepression of ethylene-stabilized transcription factors (EIN3/EIL1) mediates jasmonate and ethylene signaling synergy in Arabidopsis. Proceedings of the National Academy of Sciences, 108(30), 12539–12544. https://doi.org/10.1073/pnas.1103959108

Zoubenko, O., Uckun, F., Hur, Y. et al. Plant resistance to fungal infection induced by nontoxic pokeweed antiviral protein mutants. Nat Biotechnol 15, 992–996 (1997). https://doi.org/10.1038/nbt1097-992

Chapter 4 - Discussion

4.1 Survey and phylogenetics of RIPs in plants

In Dougherty and Hudak (2022a) our understanding of the structure, function, and evolution of RIPs was expanded from a curated dataset of more than 800 RIPs identified from publicly available sequencing data. From this dataset, RIPs were found to be present in 120 species and had 15 distinct domain configurations. This is more than had ever been identified before, indicating that RIPs are more common and more structurally and functionally diverse than previously known. For example, additional RIP domains might be involved in fine-tuning what and when depurination occurs, or it may be that RIPs have functions other than as a glycosylase. One of the most unexpected results of this paper was that about two thirds of RIPs identified lacked a signal peptide which means these proteins are likely translated on free ribosomes in the cytoplasm. This result is not unheard of in the literature, however. For example, 69% of RIPs identified in rice do not have a signal peptide (Wytynck et al. 2017) and similar trends have been noted in other cereal plants (De Zaeytijd and Van Damme, 2017). It could be, as noted by Wytynck et al. (2017), that some of the proteins identified no longer have depurination activity despite high homology to known RIPs, or that they have a reduced depurination efficiency. Alternatively, it could also be that some of these RIPs are sequestered to other parts of the cell by other means, or that the plant's ribosomes have evolved resistance to depurination by its own cytosolic RIPs. This would need to be determined experimentally.

Many of the proteins we classified as single-domain RIPs had long C-terminal ends which may have a functionality like JIP60, where they are inactive and are proteolytically cleaved

under certain conditions, or it might be that as protein domain databases improve, some of the RIPs with these long C-terminal sequences have additional domains identified that alter how they function. Our research also revealed that most plant species contain more than one RIP, with the Order Poales by far containing the most RIPs per species (Dougherty and Hudak 2022a). This means that the additional copies have more freedom to vary over evolutionary time and potentially provide different kinds of benefits to the plants containing them.

We identified the most highly conserved amino acids through multiple sequence alignment of all identified RIP domains. As expected, the most highly conserved amino acids were located in the active site pocket of the protein, and the two positions associated directly with glycosylase activity were at least 92% conserved. However, there was variability with the conservation of other positions between different RIP groups. Among the proteins identified with novel secondary domains, the two amino acids involved in stabilizing the RNA in the active site are not conserved which indicates that the optimal substrate for this group of proteins might be variable. For example, they might be more ideally suited to stabilizing nucleic acids with a different secondary structure which could make these RIPs better at depurinating RNAs other than rRNA.

To explore how RIPs evolved in plants, we also used our RIP dataset to generate a gene tree. This showed that RIPs began as single-domain proteins and later certain lineages, through several fusion events with other domains, resulted in the emergence of the multi-domain RIPs. Additionally, some of the lectin domain-containing RIPs also lost this domain, reverting them back to type I RIPs. This new gene tree is more complete compared to those that came before (Di Maro et al., 2014; Lapadula et al., 2017) because of the increased amount of sequence data

available. Moreover, previous studies identified RIPs by conducting searches using well-studied RIPs as the query, which created results biased in favor of those most similar to the query proteins and was likely to miss taxonomic groups more distantly related to these well-studied species. The current paper instead identified RIPs by conserved domain; therefore, our method makes it easier to identify novel RIPs and RIPs that have additional unexpected domains or novel long C-terminal sequences.

One limitation of our study is that it relies heavily on the current consensus sequence of a RIP as defined by the conserved domains database in NCBI (Sayers et al. 2022), which is constantly being updated. Therefore, some of the genes identified here as RIPs may not meet future criteria, and others may have been missed. On the other hand, as more and higher quality sequence data are uploaded to this same database the potential to find other novel RIPs will increase. Ultimately, my hope is that this research serves as a basis for additional wet lab testing for RIPs in plant species that had previously not had any identified, or to explore additional RIP variety in plants.

Because the analysis in this paper was built on in-house R code that is complicated to explain in the limited space available in a methods section, we decided to publish an accompanying methods paper describing the code in detail and the rationale behind each step (Dougherty and Hudak, 2022b, Appendix). Researchers are often deterred from reusing public data because of concerns about data quality and/or lack of knowledge about available databases and bioinformatic skills (Denk 2017). Therefore, this paper also serves an important example of how to take advantage of publicly available protein data in a step-by-step manner.

4.2 Pokeweed genome sequencing and annotation

Pokeweed is not a model plant, so bioinformatic analyses performed in this plant rely heavily on a well-assembled and annotated genome. In Dougherty and Hudak (2023) we produced the first publicly available pokeweed genome assembly and annotations. This is a substantial improvement over the pokeweed genome assembly produced by Neller et al. (2019), which was not made publicly available due to the highly fragmented nature of the assembly. This is because, like many plants, pokeweed has a high repeat content, more than 50%. Areas that are the hardest to assemble are those rich in repetitive sequences such as transposable elements, tandem arrays, and ribosomal gene clusters (Sun et al. 2022); when assembling a genome, only the regions where the length of the repeats is shorter than the length of the read can be properly assembled. In Neller et al. (2019) short Illumina reads were used, and despite 80-fold read depth, it was impossible to produce a genome assembly of any significant contiguity.

While three years might seem like a short time between genome assemblies, much has changed in terms of technology and sequencing cost, making the time right for another attempt at a pokeweed genome assembly. We carefully considered which sequencing technology or technologies to utilize, with our priorities to maximize read length while keeping within budget. At first BioNano Genomics optical mapping seemed like the best choice as it was the least expensive and had maps so long that it was possible to scaffold the entire genome to chromosome scale (Deschamps et al. 2018). However, the pokeweed genome produced by Neller et al. (2019) was too fragmented to make use of this technology. Our next choice was between Oxford Nanopore, which currently produces the longest sequencing reads but are

error prone, or PacBio Hifi reads, which are very accurate but not quite as long. Ultimately the PacBio Hifi reads seemed most likely to produce the highest quality result for the most reasonable price.

Protein coding regions will be minimally affected by the use of short reads for assembly as they contain few repeats, however introns, upstream regions, and downstream regions have the potential to be placed on separate contigs. This was the case for a RIP referred to as novel PAP. It was originally identified by Kira Neller (2019), but she could not characterize it fully because the sequence was assembled onto the very end of a contig, calling into question the validity of the sequence as contigs tend to be less reliable towards the ends (PacBio 2023), and because the 5' untranslated region was not assembled within the same contig making the sequence incomplete. In the current version of the pokeweed genome, on the other hand, this gene is located squarely in the middle of a 19 Mb contig and was confirmed to have many large repeat regions in this gene's intron located in the 5' untranslated region. Our current genome assembly and gene annotations are substantially improved which allows us to investigate changes in gene expression in response to stress with greater certainty.

4.3 Early response of pokeweed to stress

This paper has also revealed numerous insights into the ways pokeweed responds to JA, and therefore stress, through the RNA-Seq time-course. There are currently only two other time-course experiments looking at expression changes in plants following JA application, and both were done in Arabidopsis (Hickman et al. 2017; Zander et al. 2020). Therefore, this experiment serves as a useful contrast by highlighting how different species respond to stress.

Furthermore, unlike pokeweed, Arabidopsis is not known to be a particularly hardy plant (Rivero et al. 2014) making the comparison even more insightful because it shows what happens in a plant that can more successfully defend against stress.

One of the main differences between the pokeweed RNA-Seq data and that from the JA time-course done in Arabidopsis by Hickman et al. (2017) is that more differential expression changes were observed in the earlier time points in Arabidopsis whereas in pokeweed the later time points showed greater differential gene expression. However, this may be due to differences in study design. For example, one of the analyses I attempted for this paper that we ultimately decided not to publish was a re-analysis of the raw data from Hickman et al. (2017) and we observed that the quality of the data, not originally reported by the authors, was low and in particular the biological variability between replicates was high in the later time points in the experiment which likely introduced noise to the experiment. These results made direct comparisons between pokeweed and Arabidopsis more subject to misinterpretation, and therefore were not explored further.

The JA signaling pathway constructed in Dougherty and Hudak (2023) from the combined results of several reviews on the subject (Kazan and Manners, 2013; Liu and Timko, 2021; Song et al., 2022) allows the reader to see a key subset of the differential expression changes in pokeweed in the context of a known pathway. While a similar JA response pathway was constructed in Norway spruce (*Picea abies*) (Wilkinson et al. 2022), the JA signaling pathway constructed in Dougherty and Hudak (2023) is the first time such a pathway has been elucidated in pokeweed and, to our knowledge, the only such pathway looking at early transcriptomic changes. Although the JA pathway was expected to be similar to Arabidopsis,

there are some notable differences. For example, in Arabidopsis there is considerable cross talk between the different hormonal signalling pathways (Wang et al. 2020, Xu et al. 2020) however in pokeweed this was not consistently the case. Typically, JA and ethylene act synergistically to regulate key genes (Liu and Timko 2021) including those involved with growth (Munné-Bosch et al 2018), reproduction, and stress responses (Kieber 1997) such as leaf senescence (Aeong et al 1997). EIL1, which is an ethylene-responsive abiotic response gene (Peng et al 2014), was upregulated during the pokeweed JA time course whereas EIN3, which is another ethylene-responsive gene involved in inhibition of leaf growth (Munné-Bosch et al 2018) and leaf senescence (Li et al 2013), was not differentially expressed in pokeweed. On the other hand, the well-established antagonism between jasmonic acid and salicylic acid (SA) known in Arabidopsis (Thaler et al., 2012; Liu and Timko, 2021) was observed in pokeweed as JA biosynthesis genes were upregulated, genes related to SA biosynthesis were downregulated, and genes related to SA inactivation were upregulated (Dougherty and Hudak 2023).

The results of gene ontology (GO) analysis on the gene groups in pokeweed agree with published data in Arabidopsis (Zander et al. 2020) that plants under stress generally sacrifice growth to respond to stress (Huot et al. 2014), but they differ in the timing, intensity, and specificity of these responses. For example, leaf senescence and chlorophyll degradation, which are triggered by JA in Arabidopsis (He et al 2002), were not upregulated in pokeweed or enriched among GO terms. One possible explanation for the differences in defense response between pokeweed and Arabidopsis is that the former is a perennial and the latter is an annual plant. When Arabidopsis is under stress, it can use an escape strategy by diverting resources away from leaves and into seeds. Pokeweed, on the other hand, would more likely instead use

strategies to deal with stress in place and therefore perhaps relies less on leaf senescence as a defense mechanism compared to Arabidopsis. Instead, pokeweed seems to deploy a larger arsenal of secondary metabolites in response to stress, the majority of which may be constitutively expressed (Dougherty and Hudak 2023).

4.4 How do plants defend themselves?

This research explores plant defense on two different but complementary levels. Firstly, the defense response of a RIP-containing plant was elucidated by application of JA in a time-course experiment. The benefit of this research is that it serves as a necessary counterpoint to currently available defense response research in the model plant Arabidopsis. Secondly, the JA-responsive defense protein, the RIP, was explored in greater detail in all plant species and serves to emphasize how widespread and diverse this protein group is. Together, they help to answer the global question, how do plants defend themselves?

4.5 Future work

Both papers have made publicly available a wealth of data for future experiments. The pokeweed genome assembly and annotations serve as a continued resource for any researcher, and there are many applications for this data. For example, any experiment that involves aligning reads to a reference genome, such as RNA-Seq, ChIP-Seq, or ATAC-seq, rely on a quality reference genome for accurate results. One experiment of interest to this lab is a modified version of the PARE assay (German et al. 2009) designed to identify depurinated sites from RNA. This would illuminate the impact of PAP-I on pokeweed RNA at baseline levels compared to, for example, the elevated levels observed during JA application. The quality of this genome

also improves the ability to design sequences specific to certain regions of the genome, such as primers, micro RNAs, or guide RNAs, without the exhaustive trial and error that would be required without such a resource. For example, developing a strain of pokeweed with the RIP genes knocked out, or whose expression could be controlled, would be extremely useful for future research on the function of RIPs. Furthermore, there are many in-silico analyses that can be performed on this dataset such as promoter studies or closer looks at specific proteins. For example, it was revealed that all RIPs are present on the same contig within 10 Mb of each other, and all except for novel PAP and PAP-II are located in a region of about 225 Kb. This smaller region also contains many partial RIP sequences and repeats, indicating that the expansion of RIP genes in pokeweed is the result of tandem duplication. Finally, the usefulness of this genome extends beyond pokeweed as it also represents the only publicly available genome assembly in the taxonomic family Phytolaccaceae. While sequence similarity diminishes the less related two species are, it can serve as a place to start in an underrepresented taxon.

There is also a great deal to uncover about how pokeweed responds to stress from the RNA-seq experiment that was not included in the JA time course paper. For example, we did not discuss the gene expression patterns of the pokeweed RIPs because we believed that this topic warranted its own publication along with additional wet lab validation and the information about genomic RIP placement. For instance, PAP-I was identified to be upregulated in response to JA at 24 hours (Neller et al. 2019) and the results of this time-course agree with the current JA time-course results as PAP-I was upregulated at all time points. It is my hope that

other researchers studying non-model plants in response to stress will be able to use the data from this time-course as a point of comparison.

The pokeweed phylogeny paper identifies many uncharacterized or under-characterized proteins in species that could be investigated with wet lab experiments. For example, it would be interesting to see confirmation of the depurination activity of RIPs identified in plant species where RIPs were previously unknown with a depurination assay. It would also be interesting to see how some of the non-lectin secondary domains affect protein binding or enzymatic activity. Insights into the functionality of these proteins could further our understanding of how plants defend themselves.

References

Ali, M.S. and Baek, K.H. (2020) Jasmonic Acid Signaling Pathway in Response to Abiotic Stresses in Plants. Int. J. Mol. Sci. doi:10.3390/ijms21020621

Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S. O., Gundlach, H., Hale, I., Mascher, M., Spannagl, M., Wiebe, K., Jordan, K. W., Golan, G., Deek, J., Ben-Zvi, B., Ben-Zvi, G., Himmelbach, A., MacLachlan, R. P., Sharpe, A. G., Fritz, A., Ben-David, R., … Distelfeld, A. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science, 357(6346), 93–97. https://doi.org/10.1126/science.aan0032

Bannenberg, G., Martínez, M., Hamberg, M., Castresana, C. (2009) Diversity of the enzymatic activity in the lipoxygenase gene family of Arabidopsis thaliana. Lipids.;44(2):85-95. doi: 10.1007/s11745-008-3245-7. PMID: 18949503.

Baranwal, V. K., Tumer, N. E., Kapoor, H. C. (2002) Depurination of ribosomal RNA and inhibition of viral RNA translation by an antiviral protein of Celosia cristata. Indian J. Exp. Biol. 40, 1195–1197.

Barbieri, L., Aron, G. M., Irvin, J. D., Stirpe, F. (1982) Purification and partial characterization of another form of the antiviral protein from the seeds of Phytolacca americana L. (pokeweed). Biochem. J. 203, 55–59.

Barbieri, L., Bolognesi, A., Cenini, P., Falasca, A. I., Minghetti, A., Garofano, L., Guicciardi, A., Lappi, D., Miller, S. P., Stirpe, F. (1989) Ribosome-inactivating proteins from plant cells in culture. The Biochemical journal, 257(3), 801–807. https://doi.org/10.1042/bj2570801

Barbieri L., Valbonesi P., Bonora E., Gorini P., Bolognesi A., Stirpe F. (1997) Polynucleotide: adenosine glycosidase activity of ribosome-inactivating proteins: effect on DNA, RNA and poly(A), Nucleic Acids Res., 25 pp. 518-522, 10.1093/nar/25.3.518

Bass, H. W., Webster, C., OBrian, G. R., Roberts, J. K., Boston, R. S. (1992) A maize ribosome-inactivating protein is controlled by the transcriptional activator Opaque-2. The Plant cell, 4(2), 225–234. https://doi.org/10.1105/tpc.4.2.225

Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., Dekker, J. (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. Methods, 58(3), 268–276. https://doi.org/10.1016/j.ymeth.2012.05.001

Bolognesi, A., Barbieri, L., Abbondanza, A., Falasca, A. I., Carnicelli, D., Battelli, M. G., Stirpe, F. (1990) Purification and properties of new ribosome-inactivating proteins with RNA N-glycosidase activity. Biochimica et biophysica acta, 1087(3), 293–302. https://doi.org/10.1016/0167-4781(90)90002-j

Bolognesi, A., Bortolotti, M., Maiello, S., Battelli, M. G., Polito, L. (2016) Ribosome-inactivating proteins from plants: a historical overview. Molecules 21:1627. doi: 10.3390/molecules21121627

Boter, M., Ruíz-Rivero, O., Abdeen, A., Prat, S. (2004) Conserved MYC transcription factors play a key role in jasmonate signaling both in tomato and Arabidopsis. Genes Dev. 18(13):1577-91 doi:10.1101/gad.297704.

Browse J. (2005) Jasmonate: an oxylipin signal with many roles in plants. Vitam Horm 72: 431–456

Carnicelli, D., Brigotti, M., Montanaro, L., Sperti, S. (1992) Differential requirement of ATP and extra-ribosomal proteins for ribosome inactivation by eight RNA N-glycosidases. Biochemical and biophysical research communications, 182(2), 579–582. https://doi.org/10.1016/0006-291x(92)91771-h

Chaudhry, B. Müller-Uri, F., Cameron-Mills, V., Gough, S., Simpson, D., Skriver, K., Mundy, J. (1994) The barley 60 kDa jasmonate-induced protein (JIP60) is a novel ribosome-inactivating protein. Plant J., 6 pp. 815-824, 10.1046/j.1365-313x.1994.6060815.x

Chen, Y. A., Wen, Y. C., and Chang, W. C. (2012) AtPAN, an integrated system for reconstructing transcriptional regulatory networks in Arabidopsis thaliana. BMC Genomics, 13; 85

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., Li, H. (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature methods, 18(2), 170–175. https://doi.org/10.1038/s41592-020-01056-5

Chini, A., Fonseca, S., Fernández, G., Adie, B., Chico, J. M., Lorenzo, O., García-Casado, G., López-Vidriero, I., Lozano, F. M., Ponce, M. R., Micol, J. L., Solano, R. (2007) The JAZ family of repressors is the missing link in jasmonate signalling. Nature, 448(7154), 666–671. https://doi.org/10.1038/nature06006

Chini, A., Gimenez-Ibanez, S., Goossens, A., Solano, R. (2016) Redundancy and specificity in jasmonate signalling. Curr. Opin. Plant Biol. 33, 147–156. doi: 10.1016/j.pbi.2016.07.005

Choudhary, N., Kapoor, H. C., and Lodha, M. L. (2008) Cloning and expression of antiviral/ribosome-inactivating protein from Bougainvillea xbuttiana. J. Biosci. 33, 91–101. doi: 10.1007/s12038-008-0025-8

Citores, L., Iglesias, R., Ferreras, J. M. (2021) Antiviral Activity of Ribosome-Inactivating Proteins. Toxins (Basel). 13(2):80. doi: 10.3390/toxins13020080.

Endo, Y., Tsurugi, K., Lambert, J. M. (1988) The site of action of six different ribosome-inactivating proteins from plants on eukaryotic ribosomes: the RNA N-glycosidase activity of the proteins. Biochem. Biophys. Res. Commun. 150, 1032–1036. doi:10.1016/0006-291X(88)90733-4.

Dallal, J. A., Irvin, J. D. (1978) Enzymatic inactivation of eukaryotic ribosomes by the pokeweed antiviral protein. FEBS Lett. 89, 257–259. doi: 10.1016/0014-5793(78)80230-0

Day, P. J., Lord, J. M., Roberts, L. M. (1998) The deoxyribonuclease activity attributed to ribosome-inactivating proteins is due to contamination. European journal of biochemistry, 258(2), 540–545. https://doi.org/10.1046/j.1432-1327.1998.2580540.x

Deschamps, S., Zhang, Y., Llaca, V., Ye, L., Sanyal, A., King, M., May, G., Lin, H. (2018) A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. Nature communications, 9(1), 4844. https://doi.org/10.1038/s41467-018-07271-1

De Zaeytijd, J., Van Damme, E. J. (2017) Extensive Evolution of Cereal Ribosome-Inactivating Proteins Translates into Unique Structural Features, Activation Mechanisms, and Physiological Roles. Toxins, 9(4), 123. https://doi.org/10.3390/toxins9040123

Di Maro, A., Citores, L., Russo, R., Iglesias, R., Ferreras, J. M. (2014) Sequence comparison and phylogenetic analysis by the Maximum Likelihood method of ribosome-inactivating proteins from angiosperms. Plant molecular biology, 85(6), 575–588. https://doi.org/10.1007/s11103-014-0204-y

Dougherty, K., Hudak, K. A. (2022a) Phylogeny and domain architecture of plant ribosome inactivating proteins. Phytochemistry. 2022 Oct; 202:113337. doi: 10.1016/j.phytochem.2022.113337. PMID: 35934106

Dougherty, K., Hudak, K. A. (2022b) Computational curation and analysis of publicly available protein sequence data from a single protein family. MethodsX, 9, 101846. https://doi.org/10.1016/j.mex.2022.101846

Duggar, B. M., Armstrong, J. K. (1925) The effect of treating the virus of tobacco mosaic with the juice of various plants. Ann. Mo. Bot. Gard. 12:359 366.

Endo, Y., Mitsui, K., Motizuki, M., Tsurugi, K. (1987) The mechanism of action of ricin and related toxic lectins on eukaryotic ribosomes. The site and the characteristics of the modification in 28 S ribosomal RNA caused by the toxins. J Biol Chem. 262, 5908–5912

Endo, Y., Tsurugim, K. (1987) RNA N-glycosidase activity of ricin A-chain. Mechanism of action of the toxic lectin ricin on eukaryotic ribosomes. J Biol Chem. 262, 8128-8130, 10.1016/s0021-9258(18)47538-2

Foa-Tomasi, L., Campadelli-Fiume, G., Barbieri, L., Stirpe F. (1982) Effect of ribosome-inactivating proteins on virus infected cells. Inhibition of virus multiplication and of protein synthesis, Arch. Virol., pp. 323-332, 10.1007/BF01315062

Frankel, A., Welsh, P., Richardson, J., Robertus, J. D. (1990) Role of arginine 180 and glutamic acid 177 of ricin toxin A chain in enzymatic inactivation of ribosomes. Mol. Cell Biol., pp. 6257-6263, 10.1128/mcb.10.12.6257-6263.1990

Gandhi, R., Manzoor, M., Hudak, K. A. (2008) Depurination of Brome mosaic virus RNA3 in vivo results in translation-dependent accelerated degradation of the viral RNA. J. Biol. Chem., 283 pp. 32218-32228, 10.1074/jbc.M803785200

German, M. A., Luo, S., Schroth, G., Meyers, B. C., Green, P. J. (2009) Construction of Parallel Analysis of RNA Ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. Nat Protoc. 4(3):356-62. doi: 10.1038/nprot.2009.8. PMID: 19247285.

Gessner, S. L., Irvin, J. D. (1980) Inhibition of elongation factor 2-dependent translocation by the pokeweed antiviral protein and ricin. J. Biol. Chem. 255, 3251–53.

Gu, Y. J., Xia, Z. X. (2000) Crystal structures of the complexes of trichosanthin with four substrate analogs and catalytic mechanism of RNA N-glycosidase. Proteins, 39 pp. 37-46, PMID: 10737925

Gui, S., Peng, J., Wang, X., Wu, Z., Cao, R., Salse, J., Zhang, H., Zhu, Z., Xia, Q., Quan, Z., Shu, L., Ke, W., Ding, Y. (2018) Improving Nelumbo nucifera genome assemblies using high-resolution genetic maps and BioNano genome mapping reveals ancient chromosome rearrangements. The Plant journal for cell and molecular biology, 94(4), 721–734. https://doi.org/10.1111/tpj.13894

Guranowski, A., Miersch, O., Staswick, P. E., Suza, W., Wasternack, C. (2007) Substrate specificity and products of side-reactions catalyzed by jasmonate:amino acid synthetase (JAR1). FEBS Lett. 581(5):815-20. doi: 10.1016/j.febslet.2007.01.049. PMID: 17291501.

He, Y. W., Guo, C. X., Pan, Y. F., Peng, C., Weng, Z. H. (2008) Inhibition of hepatitis B virus replication by pokeweed antiviral protein in vitro. World J. Gastroenterol. 14, 1592–1597. doi:10.3748/wjg.14.1592.

He, Y., Fukushige, H., Hildebrand, D. F., Gan, S. (2002) Evidence supporting a role of jasmonic acid in Arabidopsis leaf senescence. Plant Physiol. 128(3):876-84. doi: 10.1104/pp.010843. PMID: 11891244; PMCID: PMC152201.

Heather, J. M., Chain, B. (2016) The sequence of sequencers: The history of sequencing DNA. Genomics, 107(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Heim, M. A., Jakoby, M., Werber, M., Martin, C., Weisshaar, B., Bailey, P.C. (2003) The basic helix-loop-helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. Mol Biol Evol. 20(5):735-747. doi:10.1093/molbev/msg088

Hey, T. D., Hartley, M., Walsh, T. A. (1995) Maize ribosome-inactivating protein (b-32). Homologs in related species, effects on maize ribosomes, and modulation of activity by pro-peptide deletions. Plant physiology, 107(4), 1323–1332. https://doi.org/10.1104/pp.107.4.1323

Hickman, R., Van Verk, M. C., Van Dijken, A. J. H., Mendes, M. P., Vroegop-Vos, I. A. , Caarls, L., Steenbergen, M., Van der Nagel, I., Wesselink, G. J., Jironkin, A. , Talbot, A., Rhodes, J., De Vries, M., Schuurink, R. C., Denby, K., Pieterse, C. M. J., Van Wees, S. C. M. (2017) Architecture and Dynamics of the Jasmonic Acid Gene Regulatory Network. Plant Cell. 29(9):2086-2105. doi:10.1105/tpc.16.00958. Epub 2017 Aug 21. PMID: 28827376; PMCID: PMC5635973.

Hon, T., Mars, K., Young, G., Tsai, Y. C., Karalius, J. W., Landolin, J. M., Maurer, N., Kudrna, D., Hardigan, M. A., Steiner, C. C., Knapp, S. J., Ware, D., Shapiro, B., Peluso, P., Rank, D. R. (2020) Highly accurate long-read HiFi sequencing data for five complex genomes. Scientific data, 7(1), 399. https://doi.org/10.1038/s41597-020-00743-4

Honjo, E., Dong, D., Motoshima, H., Watanabe, K. (2002). Genomic clones encoding two isoforms of Pokeweed antiviral protein in seeds (PAP-S1 and S2) and the N-glycosidase activities

of their recombinant proteins on ribosomes and DNA in comparison with other isoforms. J. Biochem. 131, 225–31. doi:10.1093/oxfordjournals.jbchem.a003092.

Hunt, M., Newbold, C., Berriman, M., Otto, T. D. (2014) A comprehensive evaluation of assembly scaffolding tools. Genome biology, 15(3), R42. https://doi.org/10.1186/gb-2014-15-3-r42

Iordanov, M. S., Pribnow, D., Magun, J. L., Dinh, T. H., Pearson, J. A., Chen, S. L., Magun, B. E. (1997). Ribotoxic stress response: activation of the stress-activated protein kinase JNK1 by inhibitors of the peptidyl transferase reaction and by sequence-specific RNA damage to the alpha-sarcin/ricin loop in the 28S rRNA. Molecular and cellular biology, 17(6), 3373–3381. https://doi.org/10.1128/MCB.17.6.3373

Irvin, J. (1975) Purification and partial characterization of the antiviral protein from Phytolacca americana which inhibits eukaryotic protein synthesis. Arch Biochem Biophys 169, 522–528

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., Snutch, T. P., … Loose, M. (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature biotechnology, 36(4), 338–345. https://doi.org/10.1038/nbt.4060

Jiang, S. Y., Bhalla, R., Ramamoorthy, R., Luan, H. F., Nori Venkatesh, P., Cai, M., Ramachandran, S. (2012) Over-expression of OSRIP18 increases drought and salt tolerance in transgenic rice plants. Transgenic Res., 21 pp. 785-795, 10.1007/s11248-011-9568-9

Jiang, S. Y., Ramamoorthy, R., Bhalla, R., Luan, H. F., Venkatesh, P. N., Cai, M., Ramachandran, S. (2008). Genome-wide survey of the RIP domain family in Oryza sativa and their expression profiles under various abiotic and biotic stresses. Plant molecular biology, 67(6), 603–614. https://doi.org/10.1007/s11103-008-9342-4

Kassanis, B., Kleczkowskl, A. (1948) The isolation and some properties of a virus-inhibiting protein from Phytolaeca esculenta. J. gen. Microbiol. 2: 143-153.

Karran, R. A., Hudak, K. A. (2008) Depurination within the intergenic region of Brome mosaic virus RNA3 inhibits viral replication in vitro and in vivo. Nucleic Acids Res. 36, 7230–7239. doi:10.1093/nar/gkn896

Kataoka, J., Habuka, N., Masuta, C., Miyano, M., Koiwai, A. (1992) Isolation and analysis of a genomic clone encoding a pokeweed antiviral protein. Plant Mol. Biol. 20, 879–886.

Kawade K., Masuda K. (2009) Transcriptional control of two ribosome inactivating protein genes expressed in spinach (Spinacia oleracea) embryos, Plant Physiol. Biochem., 47 pp. 327-334, 10.1016/j.plaphy.2008.12.020

Kazan, K., Manners, J., (2008) Jasmonate Signaling: Toward an Integrated View. Plant Physiology, 146 1459-1468; DOI: 10.1104/pp.107.115717

Kazan, K., Manners, J. M. (2013) MYC2: the master in action. Molecular plant, 6(3), 686–703. https://doi.org/10.1093/mp/sss128

Krivdova, G., Hudak, K. A. (2015) Pokeweed antiviral protein restores levels of cellular APOBEC3G during HIV-1 infection by depurinating Vif mRNA. Antiviral Res. 122, 51–54. doi:10.1016/j.antiviral.2015.08.007

Kurinov, I. V., Uckun, F. M. (2003) High resolution X-ray structure of potent anti-HIV pokeweed antiviral protein-III. Biochem. Pharmacol. 65, 1709–1717. doi:10.1016/S0006-2952(03)00144-8.

Lapadula, W. J., Ayub, M. J. (2017) Ribosome Inactivating Proteins from an evolutionary perspective. Toxicon : official journal of the International Society on Toxinology, 136, 6–14. https://doi.org/10.1016/j.toxicon.2017.06.012

Lapadula, W. J., Marcet, P. L., Mascotti, M. L., Sanchez-Puerta, M. V., Juri Ayub, M. (2017) Metazoan ribosome inactivating protein encoding genes acquired by horizontal gene transfer. Scientific Reports, 7(1). https://doi.org/10.1038/s41598-017-01859-1

Li, H. G., Xu, S. Z., Wu, S., Yan, L., Li, J. H., Wong, R. N., Shi, Q. L., Dong, Y. C. (1999) Role of Arg163 in the N-glycosidase activity of neo-trichosanthin. Protein Eng., 12 pp. 999-1004, 10.1093/protein/12.11.999

Liu, X., Peng, K., Wang, A., Lian, C., Shen, Z. (2010) Cadmium accumulation and distribution in populations of Phytolacca americana L. and the role of transpiration. Chemosphere 78, 1136–1141. Doi: 10.1016/j.chemosphere.2009.12.030

Liu, R. S., Yang, J. H., Liu, W. Y. (2002) Isolation and enzymatic characterization of lamjapin, the first ribosome-inactivating protein from cryptogamic algal plant (Laminaria japonica A). European journal of biochemistry, 269(19), 4746–4752. https://doi.org/10.1046/j.1432-1033.2002.03165.x

Lodge, J. K., Kaniewski, W. K., Tumer, N. E. (1993) Broad-spectrum virus resistance in transgenic plants expressing pokeweed antiviral protein. Proc. Natl. Acad. Sci. U. S. A. 90, 7089–7093. doi:10.1073/pnas.90.15.7089

Mansouri, S., Choudhary, G., Sarzala, P. M., Ratner, L., Hudak, K. A. (2009) Suppression of human T-cell leukemia virus I gene expression by pokeweed antiviral protein. J. Biol. Chem. 284, 31453–31462. doi:10.1074/jbc.M109.046235

Maynard, D., Kumar, V., Sproï, J., Dietz, K. J. (2020) 12-Oxophytodienoic Acid Reductase 3 (OPR3) Functions as NADPH-Dependent α,β-Ketoalkene Reductase in Detoxification and Monodehydroascorbate Reductase in Redox Homeostasis. Plant Cell Physiol. 61(3):584-595. doi: 10.1093/pcp/pcz226. PMID: 31834385.

McConn, M., Browse, J. (1996) The critical requirement for linolenic acid is pollen development, not photosynthesis, in an Arabidopsis mutant. Plant Cell 8: 403–416

Mitros, T., Session, A. M., James, B. T., Wu, G. A., Belaffif, M. B., Clark, L. V., Shu, S., Dong, H., Barling, A., Holmes, J. R., Mattick, J. E., Bredeson, J. V., Liu, S., Farrar, K., Głowacka, K., Jeżowski,

S., Barry, K., Chae, W. B., Juvik, J. A., Gifford, J., … Rokhsar, D. S. (2020) Genome biology of the paleotetraploid perennial biomass crop Miscanthus. Nature communications, 11(1), 5442. https://doi.org/10.1038/s41467-020-18923-6

Montanaro, L., Sperti, S., Mattioli, A., Testoni, G., Stirpe, F. (1975) Inhibition by ricin of protein synthesis in vitro. Inhibition of the binding of elongation factor 2 and of adenosine diphosphate-ribosylated elongation factor 2 to ribosomes. Biochem J. 146, 127–31

Monzingo, A. F., Robertus, J. D. (1992) X-ray analysis of substrate analogues in the ricin-A chain active site. J. Mol. Biol., 227 pp. 1136-11311 45, 10.1016/0022-2836(92)90526-p

Neller, K. C. M., Klenov, A., Hudak, K. A. (2016) The Pokeweed Leaf mRNA Transcriptome and Its Regulation by Jasmonic Acid. Front. Plant Sci. 7. doi:10.3389/fpls.2016.00283.

Neller, K. C. M., Klenov, A., Guzman, J. C. Hudak, K. A. (2018) Integration of the Pokeweed miRNA and mRNA Transcriptomes Reveals Targeting of Jasmonic Acid-Responsive Genes. Front. Plant Sci. 9:589. doi: 10.3389/fpls.2018.00589

Neller, K. C. M., Diaz, C. A., Platts, A. E., Hudak K. A. (2019) De novo Assembly of the Pokeweed Genome Provides Insight Into Pokeweed Antiviral Protein (PAP) Gene Expression. Front. Plant Sci. 10:1002. doi: 10.3389/fpls.2019.01002

Obrig, T. G., Irwn, J. D. Hardesty, B. (1973) The effect of an antiviral peptide on the ribosomal reactions of the peptide elongation enzymes, EF I and EF II. Arch. Biochem. Biophys. 155:278 289.

Osborn, R.W., Hartley, M.R., (1990) Dual effects of the ricin A chain on protein synthesis in rabbit reticulocyte lysate. Inhibition of initiation and translocation. Eur. J. Biochem., 193 pp. 401-407, 10.1111/j.1432-1033.1990.tb19353.x

PacBio. (2023, February 24). Genomes vs. Gennnnes: The difference between contigs and scaffolds in Genome Assemblies. PacBio. https://www.pacb.com/blog/genomes-vs-gennnnes-difference-contigs-scaffolds-genome-assemblies/

Papaloucas, M., Papaloucas, C., Stergioulas, A. (2008) Ricin and the assassination of Georgi Markov. Pakistan journal of biological sciences : PJBS, 11(19), 2370–2371. https://doi.org/10.3923/pjbs.2008.2370.2371

Peng, J., Li, Z., Wen, X., Li, W., Shi, H., Yang, L., Zhu, H., Guo, H. (2014) Salt-induced stabilization of EIN3/EIL1 confers salinity tolerance by deterring ROS accumulation in Arabidopsis. PLoS genetics, 10(10), e1004664. https://doi.org/10.1371/journal.pgen.1004664

Peumans, W. J., Van Damme, E. (2010) Evolution of plant ribosome-inactivating proteins. Toxic Plant Proteins, Springer, pp. 1-26, 10.1007/978-3-642-12176-0_1

Peumans, W. J., Shang, C., Van Damme, E. J. (2014) Updated model of the molecular evolution of rip genes. Ribosome-Inactivating Proteins, 134–150. https://doi.org/10.1002/9781118847237.ch9

Pollmann, S., Springer, A., Rustgi, S., von Wettstein, D., Kang, C., Reinbothe, C., Reinbothe, S. (2019) Substrate channeling in oxylipin biosynthesis through a protein complex in the plastid envelope of Arabidopsis thaliana. J Exp Bot. 70(5):1483-1495. doi: 10.1093/jxb/erz015. PMID: 30690555; PMCID: PMC6411374.

Potato Genome Sequencing Consortium, Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G., Yang, S., Li, R., Wang, J., Orjeda, G., Guzman, F., Torres, M., Lozano, R., Ponce, O., Martinez, D., De la Cruz, G., Chakrabarti, S. K., Patil, V. U., … Visser, R. G. (2011). Genome sequence and analysis of the tuber crop potato. Nature, 475(7355), 189–195. https://doi.org/10.1038/nature10158

Poudel, A. N., Zhang, T., Kwasniewski, M., Nakabayashi, R., Saito, K., Koo, A. J. (2016) Mutations in jasmonoyl-L-isoleucine-12-hydroxylases suppress multiple JA-dependent wound responses in Arabidopsis thaliana. Biochimica et biophysica acta, 1861(9 Pt B), 1396–1408. https://doi.org/10.1016/j.bbalip.2016.03.006

Thines, B., Katsir, L., Melotto, M. Niu, Y., Mandaokar, A., Liu, G., Nomura, K., He, S. Y., Howe, G. A., Browse, J. (2007) JAZ repressor proteins are targets of the SCFCOI1 complex during jasmonate signalling. Nature 448, 661–665. https://doi.org/10.1038/nature05960

Rajamohan, F., Venkatachalam, T. K., Irvin, J. D., Uckun, F. M. (1999) Pokeweed antiviral protein isoforms PAP-I, PAP-II, and PAP-III depurinate RNA of human immunodeficiency virus (HIV)-1. Biochem. Biophys. Res. Commun. 260, 453–58. doi:10.1006/bbrc.1999.0922.

Ready, M. P., Brown, D. T., Robertus, J. D. (1986) Extracellular localization of pokeweed antiviral protein. Proc. Natl. Acad. Sci. U.S.A., 83 pp. 5053-5056, 10.1073/pnas.83.14.5053

Rivero, L., Scholl, R., Holomuzki, N., Crist, D., Grotewold, E., Brkljacic, J. (2014) Handling Arabidopsis Plants: Growth, Preservation of Seeds, Transformation, and Genetic Crosses. Arabidopsis Protocols. Methods in Molecular Biology, vol 1062. Humana Press. https://doi.org/10.1007/978-1-62703-580-4_1

Robinson, K. S., Toh, G. A., Rozario, P., Chua, R., Bauernfried, S., Sun, Z., Firdaus, M. J., Bayat, S., Nadkarni, R., Poh, Z. S., Tham, K. C., Harapas, C. R., Lim, C. K., Chu, W., Tay, C. W. S., Tan, K. Y., Zhao, T., Bonnard, C., Sobota, R., Connolly, J. E., … Zhong, F. L. (2022) ZAKα-driven ribotoxic stress response activates the human NLRP1 inflammasome. Science, 377(6603), 328–335. https://doi.org/10.1126/science.abl6324

Roy, S., Sadhana, P., Begum, M., Kumar, S., Lodha, M. L., Kapoor, H. C. (2006) Purification, characterization and cloning of antiviral/ribosome inactivating protein from Amaranthus tricolor leaves. Phytochemistry 67, 1865–1873. doi: 10.1016/j.phytochem.2006.06.011

Rustgi, S., Pollmann, S., Buhr, F., Springer, A., Reinbothe, C., von Wettstein, D., Reinbothe, S. (2014) Jip60-mediated, jasmonate- and senescence-induced molecular switch in translation toward stress and defense protein synthesis. Proc. Natl. Acad. Sci., 111 pp. 14181-14186, 10.1073/pnas.1415690111

Sanger, F., Nicklen, S., Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America, 74(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., … Sherry, S. T. (2022) Database resources of the national center for biotechnology information. Nucleic acids research, 50(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112

Schilmiller, A. L., Koo, A. J. K., Howe, G. A. (2006) Functional diversification of acyl-coenzyme A oxidases in jasmonic acid biosynthesis and action. Plant Physiology, 143(2), 812–824. https://doi.org/10.1104/pp.106.092916

Schrot, J., Weng, A., Melzig, M. F. (2015) Ribosome-inactivating and related proteins. Toxins, 7(5), 1556–1615. https://doi.org/10.3390/toxins7051556

Sipahioğlu, H. M., Kaya, İ., Usta, M.,Ünal, M., Özcan, D., Özer, M. (2017) Pokeweed (Phytolacca americana L.) antiviral protein inhibits Zucchini yellow mosaic virus infection in a dose-dependent manner in squash plants. Turk. J. Agric. For. 41, 256–262. doi: 10.3906/tar-1612-30

Slatko, B. E., Gardner, A. F., Ausubel, F. M. (2018) Overview of Next-Generation Sequencing Technologies. Current protocols in molecular biology, 122(1), e59. https://doi.org/10.1002/cpmb.59

Smirnova, E., Marquis, V., Poirier, L., Aubert, Y., Zumsteg, J., Ménard, R., Miesch, L., Heitz, T. (2017) Jasmonic Acid Oxidase 2 Hydroxylates Jasmonic Acid and Represses Basal Defense and Resistance Responses against Botrytis cinerea Infection. Mol Plant. 10(9):1159-1173. doi: 10.1016/j.molp.2017.07.010. PMID: 28760569.

Srivastava, A., Trivedi, S., Krishna, S. K., Verma, H. N., Prasad, V. (2009) Suppression of papaya ringspot virus infection in Carica papaya with CAP-34, a systemic antiviral resistance inducing protein from Clerodendrum aculeatum. Eur. J. Plant Pathol. 123, 241–246. doi: 10.1007/s10658-008-9358-2

Stillmark, H. (1888) Über Ricin, ein giftiges Ferment aus den Samen von Ricinus comm. L. und einigen anderen Euphorbiaceen [About ricin, a poisonous ferment [i.e., enzyme] from the seeds of Ricinus communis L. and some other Euphorbiaceae] (M.D. thesis) (in German). Dorpat, Estonia: University of Dorpat.

Stintzi A., Browse J. (2000) The Arabidopsis male-sterile mutant, opr3, lacks the 12-oxophytodienoic acid reductase required for jasmonate synthesis. Proc Natl Acad Sci USA 97: 10625–10630

Stirpe, F., Barbieri, L. (1986) Ribosome-inactivating proteins up to date. FEBS letters, 195(1-2), 1–8. https://doi.org/10.1016/0014-5793(86)80118-1

Stirpe, F., Olsnes, S., Pihl, A. (1980) Gelonin, a new inhibitor of protein synthesis, nontoxic to intact cells. Isolation, characterization, and preparation of cytotoxic complexes with concanavalin A. The Journal of biological chemistry, 255(14), 6947–6953.

Stirpe, F. (2004) Ribosome-inactivating proteins. Toxicon, 44(4), 371–383. https://doi.org/10.1016/j.toxicon.2004.05.004

Sun, Y., Shang, L., Zhu, Q. H., Fan, L., Guo, L. (2022) Twenty years of plant genome sequencing: achievements and challenges. Trends in plant science, 27(4), 391–401. https://doi.org/10.1016/j.tplants.2021.10.006

Tang, H., Lyons, E.,Town, C. D. (2015) Optical mapping in plant comparative genomics. GigaScience, 4, 3. https://doi.org/10.1186/s13742-015-0044-y

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815. https://doi.org/10.1038/35048692

Valbonesi, P., Barbieri, L., Bolognesi, A., Bonora, E., Polito, L., Stirpe, F. (1999) Preparation of highly purified momordin II without ribonuclease activity. Life sciences, 65(14), 1485–1491. https://doi.org/10.1016/s0024-3205(99)00389-6

Verma, H. N., Srivastava, S., Varsha-Kumar, D. (1996) Induction of systemic resistance in plants against viruses by a basic protein from Clerodendrum aculeatum leaves. Phytopathology 86, 485–492. doi: 10.1094/Phyto-86-485

Vivanco, J. M., Tumer, N. E. (2003) Translation inhibition of capped and uncapped viral RNAs mediated by ribosome-inactivating proteins. Phytopathology 93, 588–595. doi: 10.1094/PHYTO.2003.93.5.588

Walsh, T. A., Morgan, A. E., Hey, T. D. (1991) Characterization and molecular cloning of a proenzyme form of a ribosome-inactivating protein from maize. Novel mechanism of proenzyme activation by proteolytic removal of a 2.8-kilodalton internal peptide segment, J. Biol. Chem., 266 pp. 23422-23427, 10.1016/S0021-9258(18)54513-0

Wasternack, C., Feussner, I. (2017) The oxylipin pathways: biochemistry and function. Annu. Rev. Plant Biol. 69, 1–24. doi: 10.1146/annurev-arplant-042817-040440

Watanabe, K., Kawasaki, T., Sako, N., Funatsu, G. (1997) Actions of pokeweed antiviral protein on virus-infected protoplasts. Biosci. Biotechnol. Biochem., 61 pp. 994-997, 10.1271/bbb.61.994

Wytynck, P., Lambin, J., Chen, S., Demirel Asci, S., Verbeke, I., De Zaeytijd, J., Subramanyam, K., Van Damme, E. J. M. (2021) Effect of RIP Overexpression on Abiotic Stress Tolerance and Development of Rice. International journal of molecular sciences, 22(3), 1434. https://doi.org/10.3390/ijms22031434

Wytynck, P., Rougé, P., Van Damme, E. J. M. (2017) Genome-wide screening of Oryza sativa ssp. japonica and indica reveals a complex family of proteins with ribosome-inactivating protein domains. Phytochemistry, 143, 87–97. https://doi.org/10.1016/j.phytochem.2017.07.009

Wyatt, S. D., Shepherd, R. J. (1969) Isolation and characterization of a virus inhibitor from Phytolacca americana. Phytopathology 59: 1787-1794.

Xiao, T., Zhou, W. (2020) The third generation sequencing: the advanced approach to genetic diseases. Translational pediatrics, 9(2), 163–173. https://doi.org/10.21037/tp.2020.03.06

Xu, L., Liu, F., Lechner, E., Genschik, P., Crosby, W. L., Ma, H., Peng, W., Huang, D., Xie, D. (2002) The SCF(COI1) ubiquitin-ligase complexes are required for jasmonate response in Arabidopsis. The Plant cell, 14(8), 1919–1935. https://doi.org/10.1105/tpc.003368

Yang, T., Meng, Y., Chen, L. J., Lin, H. H., Xi, D. H. (2016) The roles of alpha-momorcharin and jasmonic acid in modulating the response of Momordica charantia to Cucumber mosaic virus. Front. Microbiol. 7:1796. doi: 10.3389/fmicb.2016.01796

Youle, R. J., Huang, A. H. (1976) Protein Bodies from the Endosperm of Castor Bean: Subfractionation, Protein Components, Lectins, and Changes during Germination. Plant physiology, 58(6), 703–709. https://doi.org/10.1104/pp.58.6.703

Zander, M., Lewsey, M. G., Clark, N. M. , Yin, L. , Bartlett, A., Guzmán, J. P. S., Hann, E., Langford, A. E., Jow, B., Wise, A., Nery, J. R., Chen, H., Bar-Joseph, Z., Walley, J. W., Solano, R., Ecker, J. R. (2020) Integrated multi-omics framework of the plant response to jasmonic acid. Nat Plants. 6(3):290-302. doi: 10.1038/s41477-020-0605-7. Erratum in: Nat Plants. 6(8):1065. PMID: 32170290; PMCID: PMC7094030

Zhabokritsky, A., Mansouri, S., Hudak, K.A. (2014) Pokeweed antiviral protein alters splicing of HIV-1 RNAs, resulting in reduced virus production. RNA, 20 pp. 1238-1247, 10.1261/rna.043141.113

Zhao, L., Sun, Y. L., Cui, S. X., Chen, M., Yang, H. M., Liu, H. M., Chai, T. Y., Huang, F. (2011) Cd-induced changes in leaf proteome of the hyperaccumulator plant Phytolacca americana. Chemosphere, 85(1), 56–66. https://doi.org/10.1016/j.chemosphere.2011.06.029

Zhu, F., Zhang, P., Meng, Y. F., Xu, F., Zhang, D. W., Cheng, J., Lin, H. H., Xi, D. H. (2013) Alpha-momorcharin, a RIP produced by bitter melon, enhances defense response in tobacco plants against diverse plant viruses and shows antifungal activity in vitro. Planta, 237(1), 77–88. https://doi.org/10.1007/s00425-012-1746-3

Zhu, F., Zhou, Y. K., Ji, Z. L., Chen, X. R. (2018) The plant ribosome-inactivating proteins play important roles in defense against pathogens and insect pest attacks. Frontiers in Plant Science, 9. https://doi.org/10.3389/fpls.2018.00146

Zoubenko, O., Uckun, F., Hur, Y., Chet, I., Tumer, N. (1997) Plant resistance to fungal infection induced by nontoxic pokeweed antiviral protein mutants. Nat. Biotechnol. 15, 992–996. doi:10.1038/nbt1097-992.

Appendix – Computational curation and analysis of publicly available protein sequence data from a single protein family

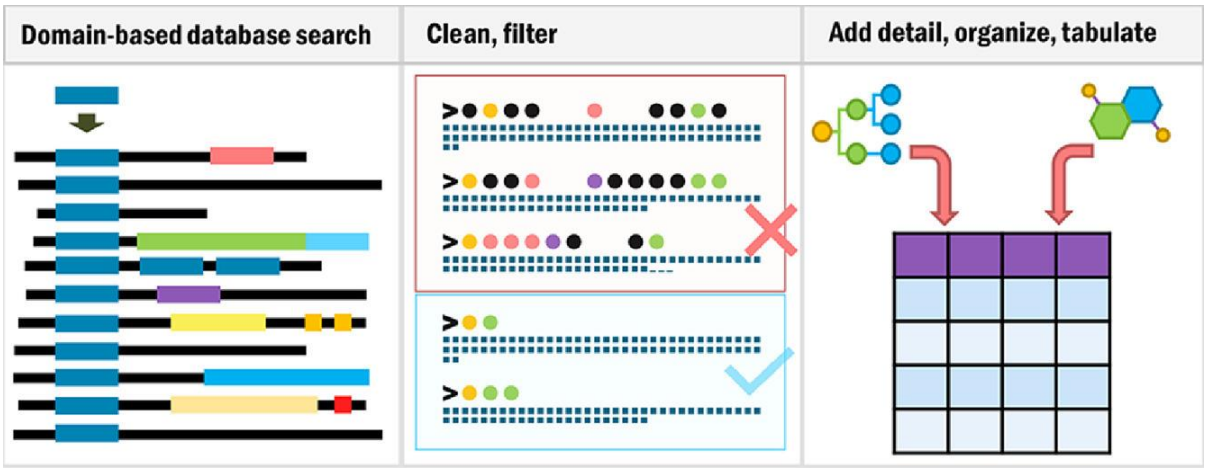
Dougherty, K., & Hudak, K. A. (2022). Computational curation and analysis of publicly available protein sequence data from a single protein family. MethodsX, 9, 101846. https://doi.org/10.1016/j.mex.2022.101846

Abstract

The wealth of sequence data available on public databases is increasing at an exponential rate, and while tremendous efforts are being made to make access to these resources easier, these data can be challenging for researchers to reuse because submissions are made from numerous laboratories with different biological objectives, resulting in inconsistent naming conventions and sequence content. Researchers can manually inspect each sequence and curate a dataset by hand but automating some of these steps will reduce this burden. This paper is a step-by-step guide describing how to identify all proteins containing a specific domain with the Conserved Protein Domain Architecture Retrieval Tool, download all associated amino acid sequences from NCBI Entrez, tabulate, and clean the data. I will also describe how to extract the full taxonomic information and computationally predict some physicochemical properties of the proteins based on amino acid sequence. The resulting data are applicable to a wide range of bioinformatic analyses where publicly available data are utilized.

- Step-by-step guide to gathering, cleaning, and parsing data from publicly available databases for computational analysis, plus supplementation of taxonomic data and physicochemical characteristics from sequence data.
- This strategy allows for reuse of existing large-scale publicly available data for different downstream applications to answer novel biological questions.

Graphical abstract



Abbreviations:

RIP (Ribosome inactivating protein)

Method name

Text manipulation for mined biological data

Specifications table

Subject Area;	Bioinformatics
More specific subject area;	Preparation of protein domain-based mined data for phylogenetic and computational analysis
Method name;	Text manipulation for mined biological data
Name and reference of original method;	No original method used
Resource availability;	● RStudio ● The following R packages: ○ seqinr v4.2-8, RRID:SCR_022678 ○ Biostrings v2.62.0, RRID:SCR_016949 ○ tidyverse v1.3.1, RRID:SCR_019186 ○ taxize v0.9.99, RRID:SCR_022677 ○ Peptides v2.4.4, RRID:SCR_022675 ● Desktop computer capable of running RStudio (2 core / 2G (RAM) / 200 G (Disk)) ● Any web browser, internet access ○ Conserved Domain Database, RRID:SCR_002077 ○ NCBI protein, RRID:SCR_003257

1. Identify all protein sequences containing the domain of interest

The example used here was the input data for the analyses described in Dougherty and Hudak [[3]]. The Conserved Domains section of NCBI (https://www.ncbi.nlm.nih.gov/cdd/) contains a database of protein domains collected from a variety of external databases. Here you can search for your domain of interest; for this example the ribosome inactivating protein (RIP domain) will be used (Fig. 1). When you select your domain of interest you will be redirected to a page which outlines details about the domain, including protein structure, related domain families, and representative sequences (Fig. 2). Under the drop-down window called “Links” select “Architectures” to be redirected to the conserved domain architecture retrieval tool [[4], [9]].

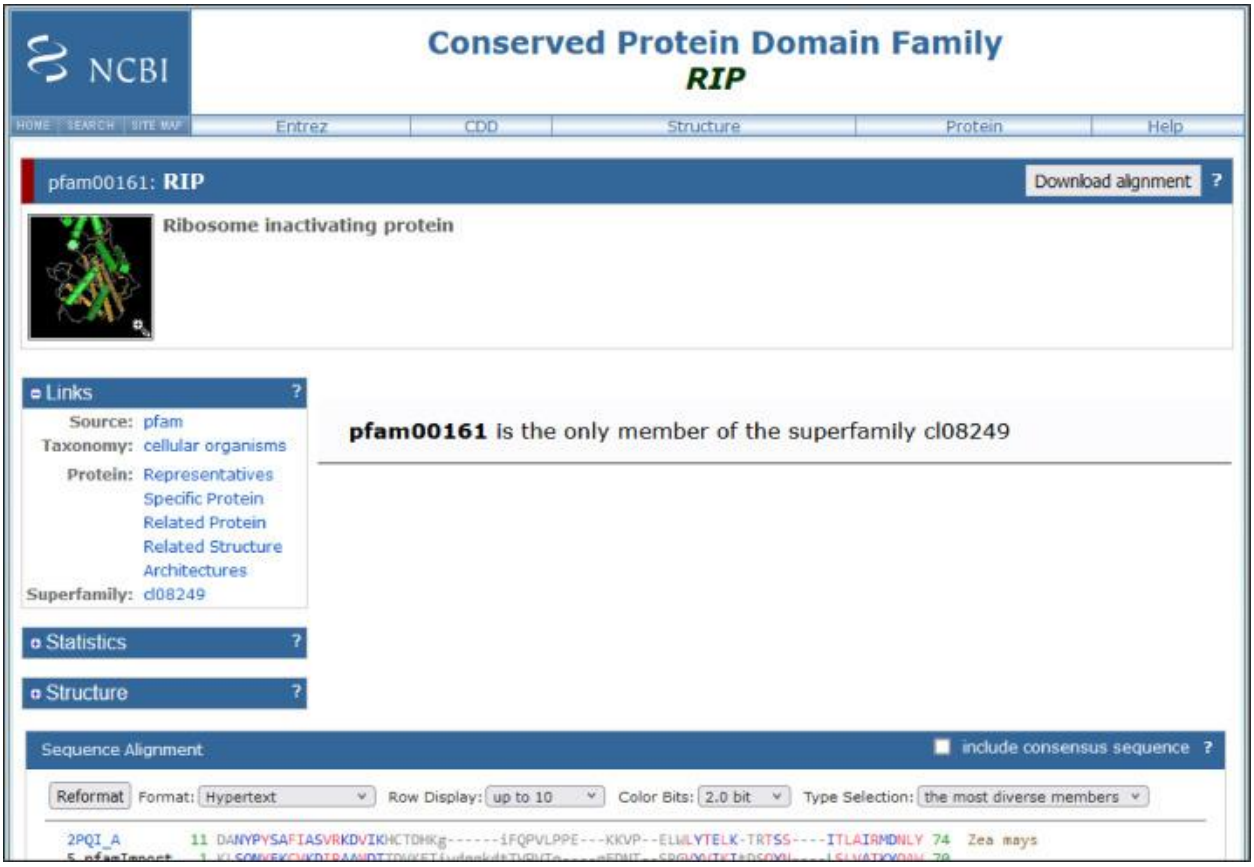


Fig. 1 Screenshot of the Conserved Domains entry for pfam00161: RIP (https://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=395109).

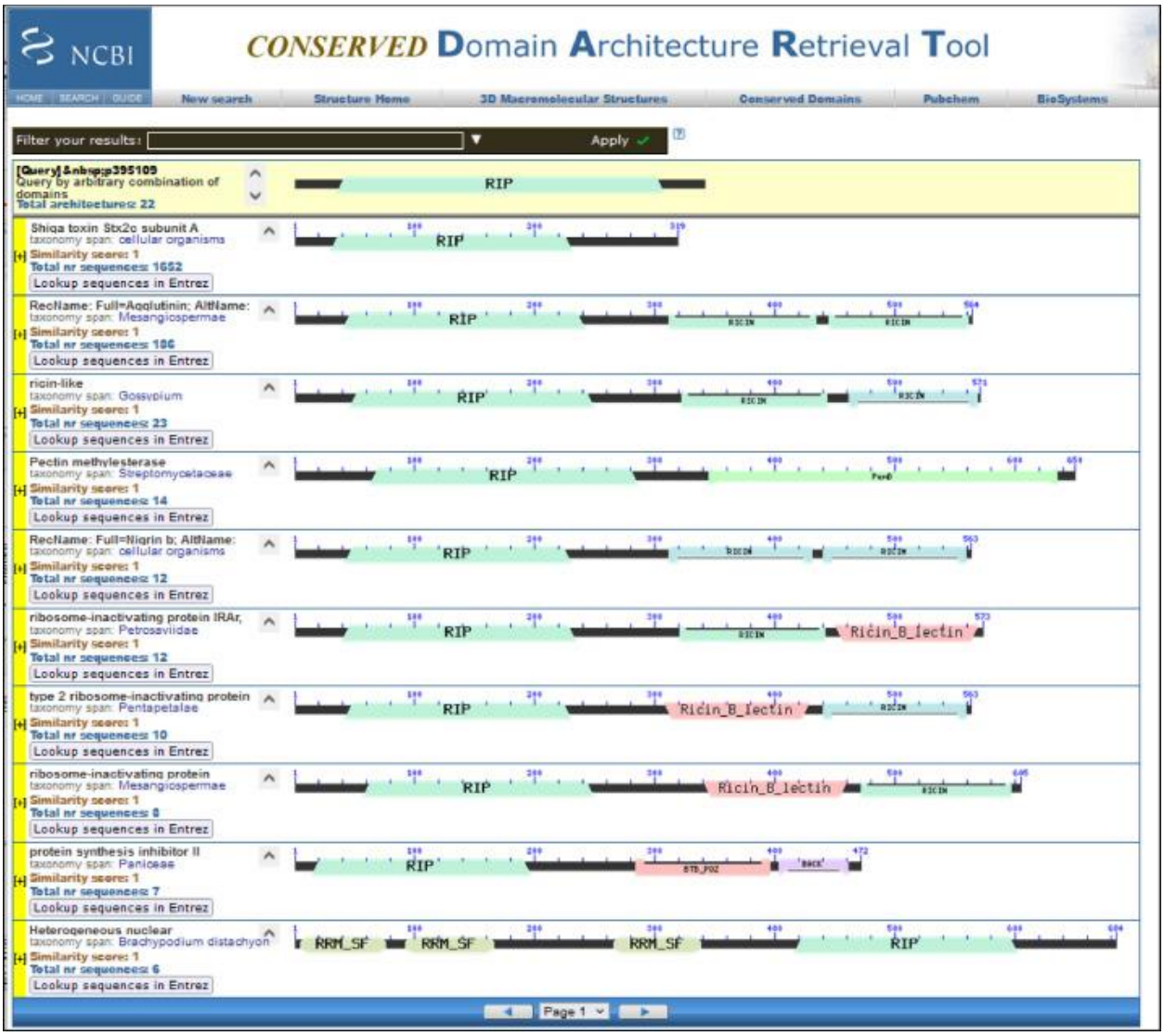


Fig. 2 Screenshot of the results page on the Conserved Domain Architecture Retrieval Tool using ‘pfam00161: RIP’ as the query.

Here you will see a graphical view of all the proteins in NCBI with annotations for your query domain, and any other domains that have been annotated as well; they will be separated into combinations of domains. These results can be filtered by taxonomy from the drop-down menu at the top. Under “Filter by taxonomy” select “NCBI taxonomy tree”, select your taxonomic group of interest (in this case plants), select “Include” at the bottom, and click “Apply” at the top to apply the changes.

To access the amino acid sequence data of the identified proteins, navigate to the domain configuration of interest and click “Lookup sequences in Entrez”. This will redirect you to the search results in the Proteins section of NCBI [[7]]. Download all sequences by selecting “Send to:” > “File” > “FASTA” > “Create file” in the top right corner. If you are interested in investigating more than one domain configuration, as is the case in this example, go back to the previous page and repeat this process for each domain configuration, then copy and paste the sequences into a single FASTA file. The raw data used in this example are available in Supplementary Data 1.

2. Clean and tabulate data in R

The following code blocks are all in the R programming language and were written in RStudio as a markdown file. This file, along with its accompanying HTML output which includes the results of each intermediate step, is available in Supplementary Data 2 and 3, respectively.

2.1 Load in data

Open RStudio and load the required packages: seqinr [[2]], Biostrings [[6]], Peptides [[5]], tidyverse [[8]], and taxize [[1]].

```
library(seqinr) # Biological Sequences Retrieval and Analysis, CRAN v4.2-8

library(Biostrings) # Efficient manipulation of biological strings, Bioconductor v2.62.0

library(Peptides) # Calculate Indices and Theoretical Physicochemical Properties of Protein Sequences, CRAN v2.4.4

library(tidyverse) # Many useful packages for data manipulation and plotting, CRAN v1.3.1

library(taxize) # Taxonomic Information from Around the Web, CRAN v0.9.99
```

2.2 Import the FASTA file, convert to table

```
fasta1 <- readAAStringSet("sequence.fasta", use.names=TRUE)
dataset_fasta1 <- data.frame(names(fasta1), paste(fasta1))
colnames(dataset_fasta1) <- c("Name","Sequence")
# Count how many sequences
print(paste0("Number of sequences before filtering: " ,
nrow(dataset_fasta1)))
```

2.3 Filter by character patterns

Most sequences will have flags in the FASTA description line indicating if a sequence is incomplete or low quality; therefore, you can remove these sequences with specific keyword searches. The commands shown below are not exhaustive but instead show some examples of potential keywords that can be used for protein data. Partial sequences can be further filtered by removing sequences that do not start with a methionine. The results of this code are visualized in Fig. 3.

Name	Sequence
CAA65328.1 antiviral protein [Volkameria aculeata]	MKASLVMTMIFGLGVLHMFEFARAQTPAIFHVGGATISIYTTFINT...
sp|P24478.2|RIPS_TRIKI RecName: Full=Ribosome-inactivati...	MIRFLVFSLLILTLFLTAPAVEGDVSFRLSGATSSSYGVFISNLRKALP...
AAD09240.1 ribosome-inactivating protein amaranthin [Am...	MIMLIIMITTVVKQSEAQQYRTVGFELHKENSPNGYANFLRRLRS...
AAL15442.1 anti-viral protein PAP [Phytolacca acinosa]	MKSMLVVTISVWLILAPTSTWAVNTIIYNVGSTTISKYATFLDNLR...
NP_001295744.1 ribosome-inactivating protein cucurmosin-...	MKGGKMNLSIMVAAWFCWSCIIFGWASAREIVCPFSSNQNYKA...
BAB83662.1 RA39p [Oryza sativa]	MVKPAAVLLLLYLPLLATPTRIGLSRNPFVPPPNSVPTIDRTEMDVS...
sp|Q03464.1|RIPA_PHYAM RecName: Full=Antiviral protein a...	MKMMVVVVVMMLSWLILKPPSTWAINTITFDVGNATINKYATF...
sp|P56626.2|RIP1_TRIAN RecName: Full=Type I ribosome-in...	MALSFFFLAISLGSPTAIGDVSFDLSTATKKSYSSFITQLRDALPTQG...
AAM89504.1 type 1 ribosome-inactivating protein musarmi...	MAASTGMHRLIIFMLIIAAAAGQGFLTVQFTETLDAVTLNRATYTA...
AAL61546.1 ribosome inactivating protein type 1 precursor [...	MHVHLINHKSFSCSAQQMKVLKQEGGKMKLMLMVMILAWLIL...

Fig. 3 Output of 'head(dataset_fasta1, *n*=10)' from the code block in Section 2.4.

```
dataset_fasta1 <-
dataset_fasta1[!str_detect(dataset_fasta1$Name,"partial"),]

dataset_fasta1 <-
dataset_fasta1[!str_detect(dataset_fasta1$Name,"[Cc]hain"),]

dataset_fasta1 <-
dataset_fasta1[!str_detect(dataset_fasta1$Name,"fragment"),]

dataset_fasta1 <-
dataset_fasta1[!str_detect(dataset_fasta1$Name,"LOW "),]

dataset_fasta1 <-
dataset_fasta1[!str_detect(dataset_fasta1$Name,"truncated"),]

dataset_fasta1 <-
dataset_fasta1[!str_detect(dataset_fasta1$Name,"protein
product"),]

# Select only sequences that start with methionine

dataset_fasta1 <-
dataset_fasta1[str_detect(dataset_fasta1$Sequence,"^M"),]

# Remove gaps/stop codons (can cause errors in other programs)

dataset_fasta1$Sequence <- gsub("\\-", "",
dataset_fasta1$Sequence)

# Count how many sequences survived this filtering process

print(paste0("Number of sequences after this filtering step: ",
nrow(dataset_fasta1)))

# Inspect the table (Table 1)

head(dataset_fasta1, n=10)
```

2.4 Identify missing species instances

Some entries will not have the standard notation for species name, which are surrounded by square brackets. Find sequences without species names within the table, then use the accession number to find the species of origin on NCBI and add them manually to the FASTA file. Then you can reload the updated FASTA file into R and continue to the next step. If

the output of the following command is empty, then there are no sequences with missing species names and no action is required.

```
dataset_fasta1[!str_detect(dataset_fasta1$Name,"\\["),]
```

2.5 Extract species names

Extract the species names by selecting the characters between the square brackets in the 'Names' column. It may also be useful to replace or delete 'special characters' such as periods and spaces, as they can cause errors for other programs in future analyses.

```
gene_tax1 <- sub(".*\\[([^][]+)].*", "\\1", dataset_fasta1$Name)
# Replace the spaces with underscores
gene_tax1 <- gsub(" ","_",gene_tax1)
# Remove the periods
gene_tax1 <- gsub("\\.","",gene_tax1)
# Inspect
head(gene_tax1, n=20)
```

2.6 Clean gene IDs

Clean gene IDs by removing everything except the accession number. Not all submissions will follow the same naming conventions, but all information about a sequence can be retrieved with the accession number so it is the only piece that is necessary to keep. Again, this code is not exhaustive but merely shows some examples of what can be done; be sure to inspect your sequence names to see what kinds of details you need to consider.

```
gene_ID1 <- dataset_fasta1$Name

# Remove any lowercase letters plus a vertical bar present
before the accession number (eg. sp|P22851)

gene_ID1 <- str_remove(gene_ID1, "^[a-z]+\\|")

# Keep only the accession number, plus the version

# This is denoted by a combination of capital letters and
numbers and sometimes underscores, followed by a period then a
single number

gene_ID1 <- str_extract(gene_ID1, "^[A-Z0-9_]+\\.[0-9]")

# Optional: remove version number

gene_ID1 <- str_remove(gene_ID1, "\\.[0-9]")

head(gene_ID1, n=20)
```

2.7 Add the accession number and species name to separate columns of the original table

The results of this code are visualized in Fig. 4.

Gene_ID	Gene_tax	Sequence
CAA65328	Volkameria_aculeata	MKASLVMTMIFGLGVLHMFEFARAQTPAIFHVGGATISIYTTFINT...
P24478	Trichosanthes_kirilowii	MIRFLVFSLLILTLFLTAPAVEGDVSFRLSGATSSSYGVFISNLRKALP...
AAD09240	Amaranthus_viridis	MIMLIIMITTVVKQSEAQQYRTVGFELHKENSPNGYANFLRRLRS...
AAL15442	Phytolacca_acinosa	MKSMLVVTISVWLILAPTSTWAVNTIIYNVGSTTISKYATFLDNLR...
NP_001295744	Jatropha_curcas	MKGGKMNLSIMVAAWFCWSCIIFGWASAREIVCPFSSNQNYKA...
BAB83662	Oryza_sativa	MVKPAAVLLLLYLPLLATPTRIGLSRNPFVPPPNSVPTIDRTEMDVS...
Q03464	Phytolacca_americana	MKMMVVVVVMMLSWLILKPPSTWAINTITFDVGNATINKYATF...
P56626	Trichosanthes_anguina	MALSFFFLAISLGSPTAIGDVSFDLSTATKKSYSSFITQLRDALPTQG...
AAM89504	Muscari_armeniacum	MAASTGMHRLIIFMLIIAAAAGQGFLTVQFTETLDAVTLNRATYTA...
AAL61546	Phytolacca_americana	MHVHLINHKSFSCSAQQMKVLKQEGGKMKLMLMVMILAWLIL...

Fig. 4 Output of 'head(dataset_fasta1, *n*=10)' from the code block in Section 2.7.

The results of this code are visualized in Table 2.

```
dataset_fasta1$Gene_tax <- gene_tax1

dataset_fasta1$Gene_ID <- gene_ID1

# Remove the old ‘Names’ column

dataset_fasta1 <- dataset_fasta1[,c("Gene_ID", "Gene_tax",
"Sequence")]

# Inspect (Table 2)

head(dataset_fasta1, n=10)
```

2.8 Check that there are no empty cells in the table

This command will return no results if all cells contain data. If any results are missing an accession number, you can use the amino acid sequence to search your raw data FASTA file and see if this number is missing or if some part of the code caused it to be lost. Row 41 in this example is missing the accession number (Fig. 5), which corresponds to line 484 of the raw FASTA file (Supplementary Data 1). The accession number provided there lacks the version number, which means that the code used in Section 2.6 above for extracting this information found no match with the expected pattern. Because the accession number is present in the FASTA file, the missing data can be added into the table.

Gene_ID	Gene_tax	Sequence
NA	Trichosanthes_kirilowii	MIRFLVLSLLILTLFLTTPAVEGDCSFRLSGATSSSYGVFISNLRKALP...

Fig. 5. Output of ‘dataset_fasta1[is.na(dataset_fasta1),]’ in Section 2.8.

```
# Check for missing data (Table 3)

dataset_fasta1[is.na(dataset_fasta1),]

# Find out which rows are affected (output to console in this
case will be: 41)

which(is.na(dataset_fasta1))

# Add missing accession number

dataset_fasta1$Gene_ID[41] <- "2019502A"

# Check again (should be an empty data frame)

dataset_fasta1[is.na(dataset_fasta1),]
```

2.9 Check for duplicate sequences

Check that there are no duplicate sequences by calculating the pairwise percentage identity of all sequences. This is necessary because there are instances where different researchers submitted the sequence of the same gene to NCBI at different times, but the sequences were not 100% identical. The following code will iterate through each sequence and do a pairwise comparison with every other sequence, tabulate the results, and save the entries with a sequence identity over 99% into a new table.

Note, the speed of this process will greatly vary depending on the number of sequences searched and the computational power allocated to R. The dataset used in this example contained approximately 820 sequences and took several minutes to run. The results of this code are visualized in Fig. 6.

Percent_identity	gene_id_query	gene_id_test	gene_tax_query	gene_tax_test	sequence_test
100.00000	CAA65328	CAA65328	Volkameria_aculeata	Volkameria_aculeata	MKASLVMTMIFGL...
100.00000	P24478	P24478	Trichosanthes_kirilowii	Trichosanthes_kirilowii	MIRFLVFSLLILTLFLT...
100.00000	AAD09240	AAD09240	Amaranthus_viridis	Amaranthus_viridis	MIMLIIMITTVVKQS...
100.00000	AAL15442	AAL15442	Phytolacca_acinosa	Phytolacca_acinosa	MKSMLVVTISVWLI...
100.00000	NP_001295744	NP_001295744	Jatropha_curcas	Jatropha_curcas	MKGGKMNLSIMVA...

Fig. 6 Output of 'head(table_pairwise_I, *n*=5)' from the code block in Section 2.9.

```
end <- length(dataset_fasta1$Gene_ID)

count <- 1:end

table_pairwise_I <- data.frame(gene1 = character(),
gene2 = character(), Percent_identity = double())

for (i in count){

pairwise <-
pairwiseAlignment(pattern = dataset_fasta1$Sequence[i:end],
subject = dataset_fasta1$Sequence[i])

pi <- data.frame(Percent_identity=pid(pairwise),
gene_id_query=dataset_fasta1$Gene_ID[i],
gene_id_test=dataset_fasta1$Gene_ID[i:end],
gene_tax_query=dataset_fasta1$Gene_tax[i],
gene_tax_test=dataset_fasta1$Gene_tax[i:end],
sequence_test=dataset_fasta1$Sequence[i:end])

table_pairwise_I <- rbind(table_pairwise_I,
pi[pi$Percent_identity > 99,])

}

# Inspect output (Table 4)

head(table_pairwise_I, n=5)
```

2.10 Make a table of the duplicates

The output will be all pairwise comparisons in your dataset, including those between other species and to itself. If you are dealing with multiple species, this may result in the identification of orthologs rather than actual duplicates, so these should be excluded. The

results of this code are visualized in Fig. 7, and the csv file saved at this step is available under Supplementary Data 4.

Percent_identity	gene_id_query	gene_id_test	gene_tax_query	gene_tax_test	sequence_test
99.31973	Q03464	AAN16078	Phytolacca_americ...	Phytolacca_americana	MKMMVVVVVMMLSW...
99.65870	P24476	CAA41953	Dianthus_caryoph...	Dianthus_caryophyllus	MKIYLVAAIAWILFQSSS...
99.64286	Q00531	AAB33361	Hordeum_vulgare	Hordeum_vulgare	MALDKVAPIVIVTPFNV...
99.65035	AAB35194	P24817	Momordica_chara...	Momordica_charantia	MVKCLLLSFLIIAIFIGVPT...
99.30070	AAB35194	ABG37691	Momordica_chara...	Momordica_charantia	MVVCLLLSFLIIAIFIGVPT...

Fig. 7 Output of 'head(table_pairwise_I2, *n*=5)' from the code block in Section 2.10.

```
# Remove the entries where the query is the same as the test

table_pairwise_I2 <-
table_pairwise_I[table_pairwise_I$gene_id_query !=
table_pairwise_I$gene_id_test,]

# Remove ones where the query and test are from different
species

table_pairwise_I2 <-
table_pairwise_I2[table_pairwise_I2$gene_tax_query ==
table_pairwise_I2$gene_tax_test,]

# Save results to a file, for reference (Supplementary Data 4)

write.csv(table_pairwise_I2,
"pairwise_percent_identity_over_99.csv",row.names = FALSE)

# Inspect output (Table 5)

head(table_pairwise_I2, n=5)
```

2.11 Remove duplicates

Remove all test sequences that matched with over 99% similarity between two sequences in the same species. The results of this code are visualized in Fig. 8.

Gene_ID	Gene_tax	Sequence
CAA65328	Volkameria_aculeata	MKASLVMTMIFGLGVLHMFEFARAQTPAIFHVGGATISIYTTFINTL...
P24478	Trichosanthes_kirilowii	MIRFLVFSLLILTLFLTAPAVEGDVSFRLSGATSSSYGVFISNLRKALP...
AAD09240	Amaranthus_viridis	MIMLIIMITTVVKQSEAQQYRTVGFELHKENSPNGYANFLRRLRS...
AAL15442	Phytolacca_acinosa	MKSMLVVTISVWLILAPTSTWAVNTIIYNVGSTTISKYATFLDNLR...
NP_0012957...	Jatropha_curcas	MKGGKMNLSIMVAAWFCWSCIIFGWASAREIVCPFSSNQNYKA...
BAB83662	Oryza_sativa	MVKPAAVLLLLYLPLLATPTRIGLSRNPFVPPPNSVPTIDRTEMDVS...
P56626	Trichosanthes_anguina	MALSFFFLAISLGSPTAIGDVSFDLSTATKKSYSSFITQLRDALPTQG...
AAM89504	Muscari_armeniacum	MAASTGMHRLIIFMLIIAAAAGQGFLTVQFTETLDAVTLNRATYTA...
AAL61546	Phytolacca_americana	MHVHLINHKSFSCSAQQMKVLKQEGGKMKLMLMVMILAWLIL...
AAB67746	Amaranthus_tricolor	MKKVLGGGTWVWWCMIMLIIMITTVVKQSEAQQYRTVGFELHK...

Fig. 8 Output of 'head(dataset_fasta1, *n*=10)' from the code block in Section 2.11.

```
dataset_fasta1 <- dataset_fasta1[! dataset_fasta1$Gene_ID %in%
table_pairwise_I2$gene_id_query, ]

# See how many sequences survived through to this stage of the
filtering process

print(paste0("Number after filtering: ", nrow(dataset_fasta1)))

# Inspect table (Table 6)

head(dataset_fasta1, n=10)
```

2.12 Save cleaned and filtered data as a FASTA file

The file generated from this code is available in Supplementary Data 5.

```
write.fasta(strsplit(dataset_fasta1$Sequence,""),
paste(dataset_fasta1$Gene_ID, datset_fasta1$Gene_tax, sep="-"),
"filtered_sequences.fasta", open="w", as.string=F)
```

3. Add physicochemical properties and detailed taxonomic information for each sequence

3.1 Tabulate species representation

If you are working with a large dataset from a variety of species, as is the case in this example, it is useful to tabulate the species representation and how many proteins are associated with each species. This can be repeated later for any taxonomic level by replacing 'Gene_tax' with the column name of the taxonomic level of interest. The results of this code are visualized in Fig. 9.

Species	Number_of_sequences
Abrus_precatorius	19
Abrus_pulchellus_subsp_tenuiflorus	4
Aegilops_tauschii_subsp_strangulata	2
Amaranthus_tricolor	1
Amaranthus_viridis	2
Ananas_comosus	3
Artemisia_annua	10
Benincasa_hispida	4
Beta_vulgaris_subsp_vulgaris	22
Brachypodium_distachyon	18

Fig. 9 Output of 'head(table_summary, *n*=10)' from the code block in Section 3.1.

```
table_summary <- as.data.frame(table(dataset_fasta1$Gene_tax))

colnames(table_summary) <- c("Species", "Number_of_sequences")

# How many species are represented in this dataset?

print(paste0("Total number of species: ",
length(table_summary$Species)))

# Inspect table (Table 7)

head(table_summary, n=10)
```

3.2 Make a table of the full taxonomy of each species based on the NCBI taxonomy classification

Note that this will take several minutes to run as retrieving the data for each species takes a couple of seconds. The results of this code are visualized in Fig. 10, and the csv file generated from this code is available under Supplementary Data 6.

Species	Number_of_sequences	Genus	Family	Order	Class	Phylum
Abrus_precatorius	19	Abrus	Fabaceae	Fabales	Magnoliopsida	Streptophyta
Abrus_pulchellus_subsp_tenuiflorus	4	Abrus	Fabaceae	Fabales	Magnoliopsida	Streptophyta
Aegilops_tauschii_subsp_strangulata	2	Aegilops	Poaceae	Poales	Magnoliopsida	Streptophyta
Amaranthus_tricolor	1	Amaranthus	Amaranthaceae	Caryophyllales	Magnoliopsida	Streptophyta
Amaranthus_viridis	2	Amaranthus	Amaranthaceae	Caryophyllales	Magnoliopsida	Streptophyta

Fig. 10 Output of 'head (taxonomy_summary, *n*=5)' from the code block in Section 3.2.

```
# Convert from data type ‘factor’ to ‘character’

table_summary$Species <- as.character(table_summary$Species)

nspecies <- length(table_summary$Species)

# Make empty data frame

full_tax <- data.frame(Species=table_summary$Species,
Genus=character(nspecies), Family=character(nspecies),
Order=character(nspecies), Class=character(nspecies),
Phylum=character(nspecies))

# Fill in data frame for each protein

for (i in full_tax$Species){

full_tax$Genus[full_tax$Species == i] <- tax_name(i,
get = "genus", db = "ncbi")$genus

full_tax$Family[full_tax$Species == i] <- tax_name(i,
get = "family", db = "ncbi")$family

full_tax$Order[full_tax$Species == i] <- tax_name(i,
get = "order", db = "ncbi")$order

full_tax$Class[full_tax$Species == i] <- tax_name(i,
get = "class", db = "ncbi")$class

full_tax$Phylum[full_tax$Species == i] <- tax_name(i,
get = "phylum", db = "ncbi")$phylum

}

taxonomy_summary <- merge(table_summary,full_tax, by= "Species")

# Inspect (Table 8)

head(taxonomy_summary, n=5)

# Save results to a file, for reference (Supplementary Data 6)

write.csv(full_tax, file = "detailed_taxonomy.csv",
row.names = FALSE)
```

3.3 Calculate physicochemical properties

Computationally infer physicochemical properties for each amino acid sequence:

aliphatic index, Bowman potential protein interaction index, theoretical net charge,

hydrophobicity index, instability index, molecular weight, monoisotopic mass over charge ratio, and isoelectric point. This package can calculate more properties than what is shown here, so this is just an example of some of them. Note: if there are unusual characters in your sequence (e.g., B, U, X, Z, *, or any number) then this code will produce an error. You can remove these sequences in the same way you removed those that did not start with a methionine. Alternatively, you can replace the amino acid with another character or with nothing (i.e., empty quotes) the same way as was done to remove special characters from the taxonomic names.

```
dataset_fasta1$aliphatic_index <-
aIndex(dataset_fasta1$Sequence)

dataset_fasta1$Boman_Potential_Protein_Interaction_index <-
boman(dataset_fasta1$Sequence)

dataset_fasta1$theoretical_net_charge <-
charge(dataset_fasta1$Sequence, pH=7, pKscale="Lehninger")

dataset_fasta1$hydrophobicity_index <-
hydrophobicity(dataset_fasta1$Sequence, scale="KyteDoolittle")

dataset_fasta1$instability_index <-
instaIndex(dataset_fasta1$Sequence)

dataset_fasta1$molecular_weight  <- mw(dataset_fasta1$Sequence)

dataset_fasta1$monoisotopic_mass_over_charge_ratio <-
mz(dataset_fasta1$Sequence)

dataset_fasta1$isoelectic_point <- pI(dataset_fasta1$Sequence,
pKscale="EMBOSS")

dataset_fasta1[order(dataset_fasta1$Gene_ID),]

head(dataset_fasta1, n=10)
```

3.4 Combine results into a single table

The csv file generated from this code is available under Supplementary Data 7, and the text file is available under Supplementary Data 8.

```
for (i in dataset_fasta1$Gene_tax){

dataset_fasta1$Genus[dataset_fasta1$Gene_tax == i] <-
taxonomy_summary$Genus[taxonomy_summary$Species == i]

dataset_fasta1$Family[dataset_fasta1$Gene_tax == i] <-
taxonomy_summary$Family[taxonomy_summary$Species == i]

dataset_fasta1$Order[dataset_fasta1$Gene_tax == i] <-
taxonomy_summary$Order[taxonomy_summary$Species == i]

dataset_fasta1$Class[dataset_fasta1$Gene_tax == i] <-
taxonomy_summary$Class[taxonomy_summary$Species == i]

dataset_fasta1$Phylum[dataset_fasta1$Gene_tax == i] <-
taxonomy_summary$Phylum[taxonomy_summary$Species == i]

}

# Save full table (Supplementary Data 7)

write.csv(dataset_fasta1, file = "tabulated_cleaned_data.csv",
row.names = FALSE)

# Save accession numbers only (Supplementary Data 8)

write.table(dataset_fasta1$Gene_ID, “accessions.txt”,
quote=FALSE, row.names=FALSE, col.names=FALSE)

# Inspect

head(dataset_fasta1, n=20)
```

The final output of this process is included in Dougherty and Hudak [[3]], Supplementary Data 1 and Supplementary Data 2.

4. Manual inspection of sequences

The protein sequence data on NCBI come from a variety of sources with different experimental purposes. Therefore, it may be necessary to assess the quality of all sequences and filter any that do not meet the standards of your experiment. While many of these steps have been done in R, some manual inspection is still advised by reading the GenPept entries of each sequence. Some useful details available on these pages are 1: whether sequences are genomic in origin or clones from cDNA, and 2: whether sequences are annotated as mature peptides. To view the GenPept pages of only the sequences that passed previous filtering steps you can use Batch Entrez (https://www.ncbi.nlm.nih.gov/sites/batchentrez, [[7]]). Upload the text file containing the NCBI accession numbers (accessions.txt), select “Protein” and select “Retrieve”. You will be redirected to a page indicating how many records were successfully retrieved. Click the link “Retrieve records”, and you will be redirected again to the Proteins database on NCBI where you can inspect each sequence or download the GenPept files.

5. Method validation

Because each step is performed in RStudio, and because the cleaning and reorganizing of data are done in stages, it is straightforward to inspect the data at each step to ensure that the changes being made are expected. This inspection was done in the case of the example used here; all relevant data were retained and all data from incomplete sequences, low quality sequences, and duplicates were removed. In addition, all irrelevant information from the FASTA description lines was removed and the filtered sequences were successfully written to a new FASTA file and their description lines contained only the NCBI accession number and the species

name. These cleaned data have many potential bioinformatic applications; for further detail on the subsequent analyses used with this dataset see Dougherty and Hudak [[3]].

Data availability

All code and data are available in supplementary materials.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

FUNDING: This work was supported by a Discovery Grant to K.A.H. from the Natural Sciences and Engineering Research Council of Canada, and a Canada Graduate Scholarship – Master's (CGS M) to K.D.

Footnotes

Related research article: K. Dougherty, K.A. Hudak Phylogeny and domain architecture of plant ribosome inactivating proteins Phytochemistry, 202 (2022), pp. 113337, 10.1016/j.phytochem.2022.113337

References

[1] S.A. Chamberlain, E. Szöcs, Taxize: taxonomic search and retrieval in R, F10 0 0 Res. 2 (2013) 191, doi: 10.12688/f10 0 0research. 2-191.v1 .

[2] D. Charif, J.R. Lobry, Seqin{R} 1.0-2: a contributed package to the {R} project for statistical computing devoted to biological sequences retrieval and analysis, Struct. Approaches Seq. Evol. (2007) 207–232, doi: 10.1007/978- 3- 540- 35306- 5 _ 10 .

[3] K. Dougherty, K.A. Hudak, Phylogeny and domain architecture of plant ribosome inactivating proteins, Phytochemistry 202 (2022) 113337, doi: 10.1016/j.phytochem.2022.113337 .

[4] L.Y. Geer, M. Domrachev, D.J. Lipman, S.H. Bryant, CDART: protein homology by domain architecture, Genome Res. 12 (2002) 1619–1623, doi: 10.1101/gr.278202 .

[5] D. Osorio, P. Rondon-Villarreal, R. Torres, Peptides: a package for data mining of antimicrobial peptides, R. J. 7 (2015) 4–14, doi: 10.32614/RJ-2015-001 .

[6] H. Pagès, P. Aboyoun, R. Gentleman, S. DebRoy. Biostrings: efficient manipulation of biological strings R package version

2.62.0. (2021) https://bioconductor.org/packages/Biostrings .

[7] E.W. Sayers, E.E Bolton, J.R. Brister, K. Canese, J. Chan, D.C. Comeau, R. Connor, K. Funk, C. Kelly, S. Kim, T. Madej, A. Marchler- Bauer, C. Lanczycki, S. Lathrop, Z. Lu, F. Thibaud-Nissen, T. Murphy, L. Phan, Y. Skripchenko, T. Tse, J. Wang, R. Williams, B.W. Trawick, K.D. Pruitt, S.T. Sherry, Database resources of the national center for biotechnology information, Nucleic Acids Res. 50 (2022), doi: 10.1093/nar/gkab1112 .

[8] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D'Agostino McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester,

M. Kuhn, T.L. Pedersen, E. Miller, S.M. Bache, K. Müller, J. Ooms, D. Robinson, D.P. Seidel, V. Spinu, K. Takahashi, D. Vaughan,

C. Wilke, K. Woo, H. Yutani, Welcome to the tidyverse, J. Open Source Softw. 4 (2019) 1686, doi: 10.21105/joss.01686 .

[9] M. Yang, M.K. Derbyshire, R.A . Yamashita, A . Marchler-Bauer, NCBI's conserved domain database and tools for protein domain analysis, Curr. Protoc. Bioinform. 69 (2020), doi: 10.1002/cpbi.90 .