The Role of Aphid Cathepsin B in Virulence: Linking Oral Secretions to Host Protein Interactions

George Alfie Seddon-Roberts

A thesis submitted for the degree of Doctor of Philosophy
University of East Anglia
John Innes Centre
September 2025

Abstract

The green peach aphid, *Myzus persicae*, is an extremely polyphagous hemipteran herbivore able to colonise plants from over 40 plant families. This insect causes significant agricultural losses, both directly by causing mechanical damage during feeding, and indirectly via transmission of over 100 plant viruses. One key to the extreme generalism of *M. persicae* is its deployment of oral secretions containing effector molecules that modulate basal plant processes.

This thesis focuses on identifying the origins of orally secreted *M. persicae* proteins. I developed organ-specific RNA-seq methods, and found members of a recently expanded clade of CathB cysteine proteases are most highly expressed in the *M. persicae* foregut, rather than the salivary glands. Previous studies by my host lab have shown that several CathB proteins, including CathB6, are effectors that promote aphid fecundity by recruiting the plant immune regulator EDS1 and its partners to p-bodies. This effect is inhibited by the *Arabidopsis thaliana* alpha-crystallin domain-containing protein ACD28.9.

I show that members of the *Myzus*-expanded CathB clade exhibit different affinities for *A. thaliana* ACD28.9 in yeast. CathB12, the closest paralog of CathB6, does not bind ACD28.9, despite only two diverging residues versus the ACD28.9-binding region of CathB6. Reciprocal residue swaps between CathB6 and CathB12 established that these residues are necessary for ACD28.9-driven rescue of CathB from p-bodies.

I also show that ACD28.9 belongs to a clade of ACD proteins significantly expanded in the order Brassicales, and that other *A. thaliana* ACD proteins in this clade also bind CathB6. Previously, work by my host lab determined that *M. persicae* CathB effector genes upregulate in aphids on two plant species in this order (*Brassica rapa* and *A. thaliana*), and downregulate on Solanales.

In summary, this work provides evidence that aphid CathB effectors and plant ACD proteins might be engaged in an evolutionary arms race.

Table of contents

List of Figures

List of Tables

Common Abbreviations

ACD – Alpha-crystallin domain

BFP – Blue fluorescent protein

CathB – Cathepsin B

ETI – Effector-triggered immunity

GFP – Green fluorescent protein

HAMP – Herbivore-associated molecular pattern

HMM – Hidden Markov model

HSP – Heat shock protein

lncRNA – Long non-coding RNA

NLR – Nucleotide binding leucine-rich repeat

PAMP – Pathogen-associated molecular pattern

PTI – Pattern-triggered immunity

RNA-seq – RNA sequencing

sHSP – Small heat shock protein

SLI1 – Sieve element-lining chaperone 1 (Kloth *et al.*, 2017)

VCS - Varicose

Acknowledgements

My PhD has been by far the most challenging experience of my life. Fortunately, I stand on the shoulders of giants, who have all been instrumental in the completion of this thesis. Though words cannot accurately express my gratitude to everyone who has shaped my life on this journey, I would like to acknowledge the professional and personal support that I have been fortunate enough to receive during my PhD.

Firstly, I would like to thank my supervisory team, who have provided me with support in times at which the demands of research felt insurmountable. Saskia, thank you for your continued support, and especially for your lessons about ingenuity and resilience. Richard, thank you for your kind words of motivation during review meetings, which repeatedly renewed my drive at exactly the times I needed them. Sam, thank you for sharing your eternal fountain of wisdom with me, and for being such a reliable source of support and kindness during challenging times. To Tom and Matteo, thank you for sharing your skills and experience with me.

To members of the Hogenhout Lab, past and present, thank you all for being sources of inspiration throughout my time in the lab. Your collective commitment to research, as well as your propensity to go above and beyond to support me when I had questions, is something which I hope I can mirror and share in the future. Qun, thank you for all of the work you have done over the years which formed the basis of several aspects of my PhD, it has been an honour to work alongside you. To the various residents of Office 124 (the NOS office), thank you for making my work environment so vibrant, supportive, and intellectually stimulating. To Josh, Reuben, and Amber, working through a degree with you has been an absolute pleasure, and I can't thank you enough.

Thank you to the various JIC platforms that have been so incredibly helpful through my PhD. Thank you to Victor and Anna from the Entomology platform, to Eva and Sergio from the Bioimaging platform, Saleha and Daniella at the TILLING platform, and everyone at the Horticultural Services platform. I quite literally couldn't have done it without your consistent technical support.

I would also like to thank Dr Elizabeth Duncan, without whom none of this would have even happened. Your continued encouragement to broaden the horizons of my research interests during my undergrad gave me the confidence to apply to this PhD, and has served as a constant motivator through its duration.

While these last four years have taught me huge amounts about insect-plant interactions, they have taught me infinitely more about myself, and a key factor in that has been my blind fortune to have found the most amazing friends that anyone could ever hope for. To Becky, Jess, Kate, Lu, Mia, and Sofia (listed in alphabetical order, no favourites here), I can't think of anyone I would rather have spent the last four years with, and our countless lunches together, crafty Thursdays and Tuesday pub quizzes are something I will hold dear forever.

Thank you to those who helped me nourish my mind, body, and spirit during the most taxing moments of this experience. To Conni, thank you for helping me to harness my resilience and perseverance, and balance when I needed it most, I truly appreciate it. To Kaori, thank you for giving me the space to process my struggles, and helping me tackle difficulties with kindness and care.

To everyone who welcomed me during my time in Cardiff, I would like to say a huge thank you (diolch!). To Kirsty Lloyd and the entire team at Amgueddfa Cymru, your support during my internship has cemented my goals for my career going forward, and I'm so glad I was able to experience museum life for the first time with you all. To Josh, thank you for coming into my life when I was navigating a new city, with a level of optimism and sunshine that has shaped the way I have seen myself, and the world, ever since.

To Toby, of all of my PhD-based victories, being your friend is the one I am proudest of. You have been a constant lighthouse in the storm, and the knowledge that your kindness and support is only ever a phone call away has been integral in navigating the most difficult moments. I mean it when I say I could not have done this without you.

[redacted], it wouldn't feel right writing an acknowledgements section and not including you. You motivated me to run the marathon that was writing this thesis, and have allowed me to grow beyond even what I thought possible. It hasn't always been easy, but I think we've done it. [redacted].

And of course, to my family, thank you for everything. Thank you for laying the foundations, thank you for showing me from an early age that I was capable of anything I set my mind to, thank you for being a soft place to land when it felt like the ground was caving in. You have made me the person I am today, and for that, no amounts of words in a thesis would ever reflect just how grateful I am.

Funding Acknowledgement

This work was supported by the Norwich Research Park Biosciences Doctoral Training Partnership, funded by UK Research and Innovation (UKRI) Biotechnology and Biological Sciences Research Council (BBSRC) NRP Doctoral Training Programme Studentship BB/T008717/1.

Chapter 1 – General Introduction

1.1: Herbivorous insects and their interactions with plants

1.1.1: Insects are a hugely diverse class of invertebrates.

The Insecta are an incredibly vast and diverse class of invertebrates that are ubiquitous across terrestrial and freshwater biomes. With the earliest known insects being thought to have existed around 410 million years ago (mya) (Engel, 2015), the insect clade has gone on to be the most abundant class of animals, with an estimated 5.5 million species (Stork *et al.*, 2015).

Insect species utilise a wide range of life history strategies, with species commonly divided into two groups: the holometabola and the hemimetabola. Holometabolous insects constitute approximately 85% of insect species (Engel, 2015) and are characterised by the presence of larvae which undergo complete metamorphosis before reaching maturity, with a sedentary pupal stage prior to adulthood. This group includes hyper-diverse groups of insects such as those from the orders Coleoptera (beetles), Hymenoptera (bees, wasps, sawflies and ants), Lepidoptera (butterflies and moths) and Diptera (flies, mosquitoes and midges), which combined contain >800,000 named species (Peters *et al.*, 2014). Hemimetabolous insects (aphids, mealybugs, whiteflies, psyllids, leafhoppers, planthoppers, treehoppers, spittle bugs/froghoppers and true bugs), however, feature 'miniaturised adult' forms known as nymphs, and merely undergo a series of moults to grow to reach maturity. It is thought that holometabolism is the derived developmental strategy, with the earliest holometabolous insects initially appearing 350 mya (Misof *et al.*, 2014).

The order Hemiptera is an incredibly diverse order of hemimetabolous insects, which includes groups such as aphids, plant hoppers and cicadas. Comprised of over 90,000 species (Mayhew, 2007), and sharing a common ancestor approximately 290 mya, the order contains insects with a wide range of life-history strategies. Though species in this order are predominantly terrestrial, there are species, such as the giant water bugs in the family Belostomatidae, which are adapted to life in freshwater habitats. Similarly, though most hemipterans are herbivorous, there are predatory species, such as assassin bugs in the family Reduviidae, as well as those that feed on blood, including bed bugs in the genus *Cimex*.

1.1.2: Insect-plant interactions are widespread across the insect phylogeny.

Insect species also vary dramatically in the extent to which they interact with plant species. These associations can vary in a number of ways and can range in outcome from mutualistic to parasitic. Through these interactions, insects perform invaluable ecosystem services, but may also negatively affect plant performance. Species across the insect phylogeny play pivotal roles in the pollination of many land plant species, with some studies suggesting as many as 90% of land plant species are pollinated by insects (Hoshiba and Sasaki, 2008). This interaction is typically performed by larger insect species, most notably Hymenoptera, and therein the honeybee, *Apis mellifera*, which performs 34% of pollination services for agriculture in the United Kingdom (Breeze *et al.*, 2011).

However, not all insect-plant interactions are so mutualistic. Plants constitute the diet of a huge number of insect species, and as such, insects cause extensive damage to plant species. Dietary diversity is rich within herbivorous insects, which have evolutionarily established a wide range of feeding guilds, each utilising highly specialised mouthparts to access different plant tissues (Lehnert *et al.*, 2025). Chewing insects, such as larvae of the lepidopteran insects, cause large amounts of tissue damage to plants during feeding. Piercing insects, including plant-feeding members of the order Hemiptera, insert specialised stylets into the plant to access the liquid inside the phloem or xylem. Insects in the piercing-sucking feeding guild generally cause less tissue damage, though feeding by some leafhoppers and planthoppers can lead to hopperburn, which has the potential to destroy whole plants or fields (Backus *et al.*, 2005). Herbivory by sap feeding insects also opens host plants to an increased risk of insect-vectored plant diseases (Fereres and Moreno, 2009).

1.1.3: Insect herbivory has significant cultural and commercial impacts.

The negative implications of insect herbivory are of particular importance when it comes to their impacts on agriculture. Crop pests have the potential to cause significant yield losses, with an estimated 20-40% of potential yield being lost as a result of pests and diseases (Douglas, 2018). This has potentially devastating ramifications upon both profit and food security, with damages arising from mechanical damages to crop species during feeding, or from the transmission of

plant viruses by a number of crop pests. Insects, particularly hemipteran pest species, act as vectors for the majority of known plant viruses (Hogenhout *et al.*, 2008), thereby compounding the negative effect caused by physical damage by insect herbivores on crop yields. This agricultural concern is further heightened when the effects of global climate change on insect herbivores are considered, with increasing temperature leading to earlier emergence of insect pests, as well as expanding the range in which insects can survive (Skendžić *et al.*, 2021). Increases in connectivity in international trade and travel also increases the chances of the introduction of invasive insect pests, exposing crops to novel challenges at the hands of newly introduced insect species (Tobin, 2018). As such, the reliance on ecologically damaging chemical pesticides such as neonicotinoids has become so strong that the UK government has allowed for the use of pesticides containing the entomotoxic compound thiamethoxam to combat virus yellows of sugar beet in England.

However, there is significant pressure to move away from broad spectrum chemical pesticides, which have proven to be increasingly damaging to insect biodiversity, indiscriminately posing risks to 'beneficial' insect species such as pollinators in addition to the pest insects which they are intended to target. Therefore, the need for highly specific methods of damage control is of growing concern.

One promising alternative is the exploitation of naturally occurring resistance to herbivores that exists within and across plant species. Plants exhibit resistance to insect herbivory through a diverse array of chemical, structural, and physiological traits that reduce herbivore performance, preference, or colonisation success. These defences may be constitutive or inducible, and include the production of specialised metabolites, the activation of defence signalling pathways, and physical barriers that impair feeding. Resistant cultivars can suppress pest populations and reduce crop damage while limiting the need for chemical intervention, mitigating issues such as pesticide resistance evolution and environmental contamination (Stout, 2013).

The identification of resistance-associated traits present in elite cultivars, wild relatives, and landraces of plant species creates opportunities for their incorporation into agricultural varieties through conventional breeding, or

molecular approaches (Smith, 2005). The viability of identification and utilisation of these resistance factors is only increasing as more modern genotyping and phenotyping approaches become available to researchers. However, the long-term viability of resistance depends on understanding the mechanisms by which insect-plant interactions occur at the interface of herbivory. Therefore, understanding the molecular and signalling processes that underpin early herbivore recognition and defence activation is essential for the rational design and breeding of resilient crop varieties. In this context, insight into insect-plant interactions provides the foundation for developing targeted, sustainable alternatives to broad-spectrum pesticidal intervention.

1.1.4: Aphid species display a variety of morphological and reproductive phenotypic plasticities.

Aphids are a family of hemipteran insects in the sub-order Sternorrhynca. Most aphid species are widely considered economically significant crop pests and can cause significant reductions in yields from a number of food crops (Van Emden and Harrington, 2007). Their success as herbivores can be attributed to several physiological adaptations which enable them to maximise the efficiency of their host colonisation, feeding establishment, and subsequent reproduction.

Many aphid species have evolved to utilise a unique life history strategy termed cyclical parthenogenesis (Figure 1.1), which involves the rapid production of genetically identical progeny by clonally reproducing asexual females. During this phase of the aphid life cycle, which typically occurs during the spring and summer when day length is long, progeny are born viviparously, and in most cases, already containing their own developing progeny. In response to waning day length at the beginning of autumn, aphids undergo a reproductive mode switch, in which sexually reproductive progeny are produced (Simon *et al.*, 2010). These sexual aphids mate and produce eggs which are capable of surviving less favourable conditions over winter. This also maintains genetic variation, with the sexually reproductive phase diminishing mutations which accumulate during the parthenogenetic generations of an aphid colony (Simon *et al.*, 2003). Despite cyclical parthenogenesis being the standard for aphid species, approximately 3% of species are reported to reproduce exclusively through female parthenogenesis, losing the sexually reproductive phase (Simon *et al.*, 2002).

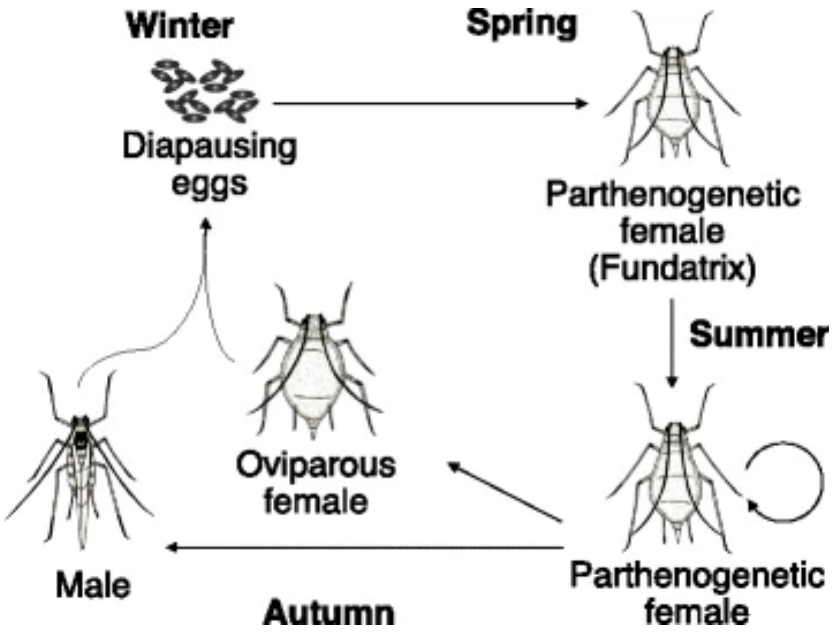


Figure 1.1: Aphids exhibit cyclical parthenogenesis involving both asexual and sexual morphs. In spring, parthenogenetic females hatch from eggs and begin reproducing clonally to create several further generations of parthenogenetic females through spring and summer. In autumn, in response to shortening day lengths, parthenogenetic females begin to give rise to sexually reproducing males and females, the latter of which lay eggs which are capable of overwintering. These eggs will give rise to the next year's first generation of parthenogenetic females. Reproduced from (Murano *et al.*, 2018) under Creative Commons License.

As the population of aphids on a plant host tends towards carrying capacity, and density reaches a critical level, aphids are capable of producing winged morphs which function to disperse the population, as well as to select and colonise suitable new host plants (Mehrparvar *et al.*, 2013). These winged morphs, though genetically identical to the founders of the population, show an upregulation in a number of genes playing a variety of roles in dispersal and nutrition (Hu *et al.*, 2019).

1.1.5: Aphids are highly adapted for colonisation of plants and phloem feeding.

Aphids, like other hemipteran insects, have highly specialised mouthparts which have evolved to effectively access the vascular tissue of host plants. In the case of aphids these specialised mouthparts, termed stylets, are used to establish feeding sites in the phloem of a host. Stylets are comprised of two maxillary stylets which are encapsulated by mandibular stylets (Figure 1.2A). Proximally, the maxillary stylets house two canals, one for the deposition of saliva—which originates at the salivary glands—and the other for the uptake of phloem sap during feeding. These two canals connect in the distal region of the stylets, forming a common duct. The aphids use these elongated mouthparts to pierce the

epidermis of a plant, and then navigate through the layers of tissue, until reaching the phloem (Figure 1.2B). The distal extremity of the aphid stylet also contains a structure called the acrostyle, which lies at the tip of the stylet and has been implicated in the binding of plant viruses such as *Cauliflower Mosaic Virus* (CaMV), and as such might be involved in the transmission of this and other non-persistent plant viruses (Uzest *et al.*, 2010, Webster *et al.*, 2017).

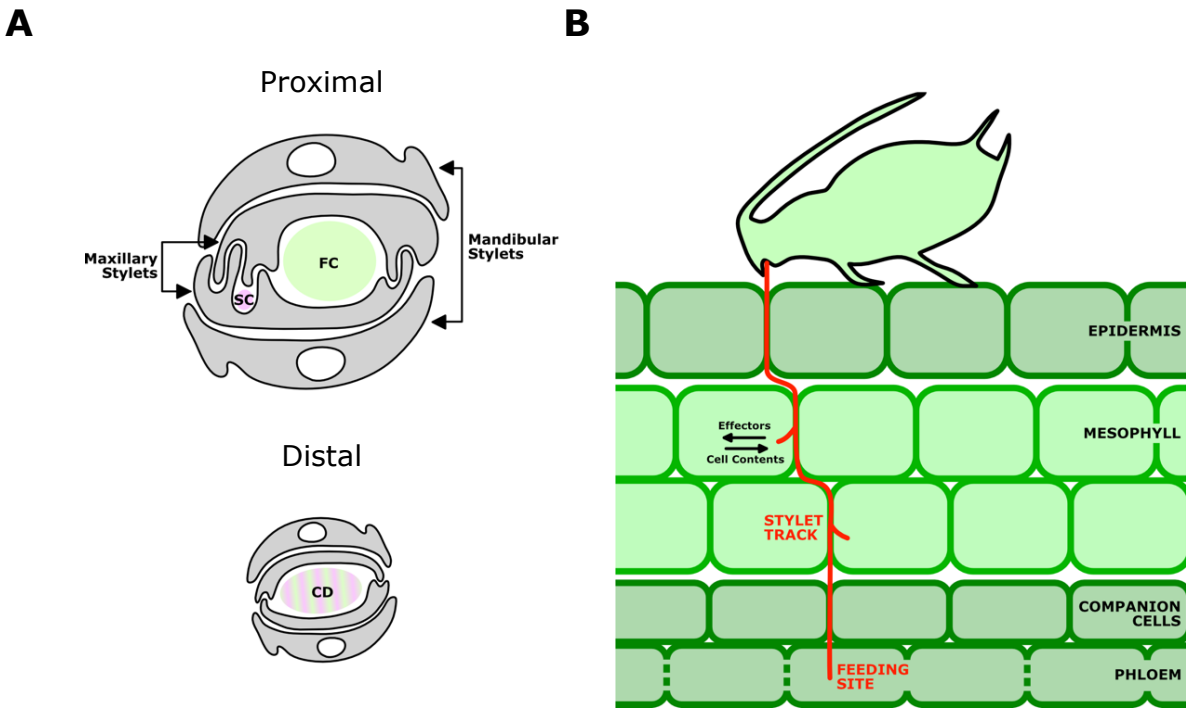


Figure 1.2: Aphids use specialised mouthparts to navigate plant tissues before accessing the vasculature.
Aphids use needle-like stylets to pierce the host plant epidermis and navigate through the layers of tissue. **A:** Cross-section of the aphid stylet at proximal and distal regions. Proximally, the food and salivary canals are separated, but merge to form a common duct at the distal regions of the stylet. FC: Food Canal; SC: Salivary Canal; CD: Common Duct. **B:** Illustration showing the path of the aphid stylet through plant tissues. Aphids sample cellular contents in the mesophyll before reaching the phloem where a feeding site is established.

The identification and establishment of feeding sites by aphids consists broadly of three distinct behavioural phases (Tjallingii, 2006). The first phase is a probing phase, in which aphids pierce the epidermis and begin to assess plant cells for quality, using metrics such as pH and sucrose concentration in order to identify high quality hosts (Hewer *et al.*, 2011). This is the phase during which most aphid-vectored, non-persistent viruses are transmitted into hosts, potentially due to their rapid disassociation from the acrostyle (Uzest *et al.*, 2007, Moreno *et al.*,

2005). After identifying a suitable host, the aphid begins to deposit saliva into the host in a phase termed the salivary phase, which aims to avoid or inhibit perception by the host immune system. If the previous steps are successful, and a consistent feeding site can be established within the phloem, the aphid begins ingesting sap in the phloem phase. This phase is typically the point at which persistent plant viruses, which are often, but not always, phloem restricted, are transmitted by the aphid.

Aphids have been shown to secrete two distinct types of saliva during the probing, salivary, and phloem feeding phases. These two saliva types, termed gelling saliva and watery saliva, have been shown to be synthesised in two different areas within the salivary glands. The aphid salivary glands are comprised of a bilobed principal gland containing 42 large secretory cells, and two accessory glands, which contain a further 4 secretory cells (Mutti *et al.*, 2008).

Gelling saliva is secreted during the migration of the aphid stylet through the apoplast prior to reaching the phloem, and has been shown to form a sheath around the stylet track (Will and Vilcinskas, 2015). It is possible that this acts as a molecular cloak for the feeding aphid, preventing plant immune activity in response to the mechanical damage or aphid-derived molecules.

Watery saliva is secreted during the penetration of sieve elements and subsequent ingestion of phloem sap (Will *et al.*, 2007) and likely has a role in inhibiting plant immunity during active feeding. Deposition of watery saliva during the phloem feeding phase has been shown to decrease the extent of sieve element occlusion (Will *et al.*, 2009), potentially acting to maintain a constant supply of sap, as well as potentially enable the migration of deposited virulence factors (Chen *et al.*, 2020).

These two distinct forms of aphid saliva are thought to have different physiological origins within the salivary glands. Watery saliva has been shown to be secreted from cells in the accessory glands, and gelling saliva originating in the principal gland (Moreno *et al.*, 2011).

The dietary restriction of feeding on plant vascular tissues gives rise to a notable nutritional challenge, in which phloem sap constitutes a carbohydrate-rich diet, but cannot provide aphids with the complete spectrum of essential amino acids.

As such, hemipterous insects have fostered a number of close endosymbiotic associations with bacterial species, which function to complement the diet of the insect, in an association so mutualistic that it has shaped the evolution of both partners. These endosymbionts, the best studied being aphid endosymbiont *Buchnera aphidicola*, are able to synthesise a number of essential amino acids which aphids themselves are not able to synthesise *de novo* (Wilson *et al.*, 2010), nor would they be able to acquire directly from phloem sap. This endosymbiosis is thought to have arisen between 250 and 150 million years ago following a single infection event (Baumann *et al.*, 1998), and is so well established that it has given rise to a specialised organ within aphid species which acts as the interface for nutritional complementarity between the aphid and its endosymbiont (Nakabachi *et al.*, 2005). The provision of nutrients such as vitamin B2 (Nakabachi and Ishikawa, 1999) to aphids by *Buchnera* has been proven to eliminate the deficit in nutrition resulting from phloem specialisation, allowing for optimal utilisation of dietary resources by *Acyrthosiphon pisum* (Akman Gündüz and Douglas, 2009).

Aphids exhibit numerous different dietary strategies, with species ranging from host-restricted specialists (monophagous), to generalists with broad host spectra (polyphagous). There is thought to be no close links between taxonomy and an aphid species' ability to colonise numerous plant species, with host generalism appearing almost indiscriminately across the aphid clade (Loxdale and Balog, 2018). These diets may also vary seasonally, with some species displaying host alternation between summer and winter (Orantes *et al.*, 2012).

1.1.6: Aphid species are valuable model systems for studying insect-plant interactions.

Aphids have become an increasingly utilised model system, due to their ability to shed light on insect-plant interactions, virus transmission, and phenotypic plasticity. However, the broad spectrum of their potential insights is not the sole reason that aphids make a particularly tractable model system. For example, aphid species are uniquely suited for use as a model system due to their ability to clonally reproduce. This parthenogenesis allows for large numbers of genetically identical individuals, produced in relatively short generation times, for use in extraction and sequencing of DNA and RNA. The cyclical nature of the parthenogenesis employed by aphids also allows for the study of genes involved

in sexual reproduction, with sexual mode shifts being well studied in aphid species (Ishikawa *et al.*, 2012).

The wide variety of life history strategies across the aphid clade is also a useful way for researchers to examine the evolution of various traits between relatively closely related species. For example, aphids in the *Myzus/Brachycaudus* clade exhibit a distinct variability in the complexity of their life cycle. Within this clade of approximately 50 species, there are examples of host-alternation between both woody and herbaceous plants. It is thought that host-alternation is an evolutionarily derived life cycle, which may have arisen multiple times since the most recent common ancestor of the clade (Emmanuelle *et al.*, 2010).

The production of alate (winged) individuals by aphid species also allows us to examine the genetic bases of sensory perception and host quality analysis by colonising aphids, as well as the genes that are involved in the development of wings. In the brown citrus aphid, *Aphis citricidus*, alate morphs showed high levels of overexpression of 279 genes when compared to apterous (wingless) morphs (Shang *et al.*, 2021). Knockdown of a number of these genes involved in lipid and glycogen metabolism led to improperly developed wings, suggesting that these genes played essential roles in wing development.

1.1.7: Several aphids including M. persicae have well annotated genomes and transcriptomic data.

The completed genome of the specialist pea aphid *A. pisum* was published in 2010 (The International Aphid Genomics Consortium, 2010), 10 years after the first insect genome (Adams *et al.*, 2000). This genome allowed researchers to examine the genetic relationship between the pea aphid and the symbiont *B. aphidicola*, as well as the function of a large number of genes in the maintenance of extreme phenotypic plasticity in response to environmental stressors. Since the publication of the pea aphid genome, an increasing number of genomes of aphids across the clade have been released, with 21 genome assemblies available as of 2021 (Shigenobu and Yorimoto, 2022, Biello *et al.*, 2021,

2020a, Mathers *et al.*, 2020b, Mathers *et al.*, 2021), eight of which are chromosome-level assemblies. More recently, a further 18 genome assemblies, including 8 chromosome-level assemblies were published for species across the

subfamily Aphidinae (Mathers *et al.*, 2023). This more comprehensive coverage of species closely related to *M. persicae* allows for more powerful evolutionary analyses between species in this clade.

Genomic resources for *M. persicae* are relatively abundant when comparing to other aphid species. A full chromosome-scale genome assembly has been generated for *M. persicae* (Clone O), the predominant genotype in the UK (Mathers *et al.*, 2017), with 27,663 protein coding genes annotated (Mathers *et al.*, 2021). There have also been studies using these genome assemblies to examine the methylome of *M. persicae* (Clone O), exploring how the methylation landscape differs between various morphs and sexes (Mathers *et al.*, 2019). Global clonal populations of *M. persicae* have also been examined, allowing for the analysis of evolution of various traits, such as insecticide resistance (Singh *et al.*, 2021, Wouters, 2021). The availability of such a breadth of genomic resources for *M. persicae* only improves its standing as a model system suitable for the analysis of a range of biological questions.

The creation of an increasing number of genomes for species spanning the aphid clade has allowed for a more comprehensive understanding of the functions of a variety of aphid genes, and their transcriptional responses to aphid life history events and stress. For example, transcriptomic analyses have been used to assess the regulation of gene transcription in response to insecticide exposure, during which susceptible genotypes of *M. persicae* displayed upregulation of 187 genes in response to exposure to an anticholinesterase insecticide (Silva *et al.*, 2012). Upregulated genes had function in metabolism and detoxification of the insecticide, suggesting a preferential upregulation of the genes in response to the insecticide. It has also been shown that aphids upregulate genes in response to endosymbiont cues, with *A. pisum* hosting facultative endosymbiont *Serratia symbiotica* expressing a Ca^{2+}-binding protein-like gene (*ApHRC*) in the salivary gland at a higher rate than those without the endosymbiont (Wang *et al.*, 2020). This increased feeding duration and assisted in the colonisation of new host plants by *A. pisum.*

There is also evidence that aphid species are able to modulate gene expression based on the specific plant host which they are colonising. A study exploring gene expression in different host races in the specialist aphid *A. pisum* indicated that

expression patterns are different between races, with increased levels of differential expression between more distantly related races (Eyres *et al.*, 2016). This study also explored adjustment to novel host plants, and identified upregulation of genes following host transition, including upregulation of some chemosensory and salivary genes. Additionally, comparison of the transcriptomes of *M. persicae* on hosts spanning the angiosperm phylogeny identified 171 genes that were differentially expressed during host adjustment (Mathers *et al.*, 2017). This study focused on a line of clones derived from a single clonally reproducing adult, which is able to survive on plants across this study, contrasting the limited survival of *A. pisum* races on novel host plants. Another study of *M. persicae* colonising diverse plant species showed that there are a number of genes that are differentially expressed dependent on the host plant (Chen *et al.*, 2020).

With the recent advancements in transcriptomic and proteomic technologies, an increasing amount of research is being used to establish the contents and origins of the saliva of herbivorous insects. Salivary proteins have been identified in whitefly *Bemisia tabaci* (Huang *et al.*, 2021), hoppers *Nilaparvata lugens, Sogatella furcifera, Laodelphax striatellus* (Huang *et al.*, 2018) and *Empoasca flavescens* (Pan *et al.*, 2024), as well as aphid species *Pseudoregma bambucicol* (Zhang *et al.*, 2023), *A. pisum* (Carolan *et al.*, 2009) and *M. persicae* (Guo *et al.*, 2020, Liu *et al.*, 2024).

1.1.8: The need for increased RNA-seq resolution in aphids.

To date, the resolution of RNA-seq data from *M. persicae* has been relatively broad. The small size and high sugar content of aphids has meant that previously generated transcriptomic resources from *M. persicae* have relied on the generation of RNA-seq libraries from whole adult bodies, and have, as such, overlooked the spatial heterogeneity of gene expression within the body. For example, winged and wingless aphids share an identical genome, and instead rely on differential expression of genes to perform their specialised functions. As such, it is expected that winged aphids, which primarily function as dispersal units for aphid populations, will overexpress genes in both the antennae and legs, which both have functions in volatile perception. However, a wingless aphid on an already established host plant is required to sustain itself long enough to reproduce. In order to do this, wingless aphids must limit host plant defence using effectors

produced in the salivary gland. Therefore, it would be expected that wingless aphids would preferentially express genes involved in plant host manipulation.

Thus, to fully elucidate the granularity of transcriptomic regulation of candidate effectors in *M. persicae*, the resolution of RNA-seq libraries must be improved. As such, RNA-seq libraries generated from isolated organs within *Myzus persicae* will improve the capacity to track the organ-specific expression of genes of interest, in this case candidate effectors. The generation of such tissue specific RNA-seq datasets will allow such hypotheses to be tested.

1.2: Plant immunity

Plants face a consistent risk of invasion by a broad range of pests and pathogens from multiple kingdoms of life. Bacteria, viruses, fungi and animals including insects are capable of parasitising plant hosts and causing a breadth of physiologically and developmentally deleterious phenotypes. Plants are unable to respond behaviourally to such threats and so rely on an intricate and effective immune response in order to ameliorate these risks. Pathogen- and pest-resistant plants deploy this response across several phases, each of which relies on the perception of a number of self and non-self molecular cues, and leading to the use of specialised metabolites which have evolved to limit mechanical and biological damage caused by invaders.

1.2.1: Plants deploy several pathways in response to pathogens and pests.

There are a number of distinct signalling cascades which are deployed by invaded plants in order to ameliorate damage caused by pathogens and pests. The specific pathways utilised during plant disease depend largely on the infection strategy of the invading pathogen or pest. Microbial plant pathogens can largely be categorised as either necrotrophic or biotrophic. Necrotrophic pathogens, which destroy plant tissues in order to access their nutrients, typically lead to jasmonic acid triggered immunity. This immune pathway triggers the production of many defensive pathogenesis related proteins, as well as a number of secondary metabolites that are damaging to pathogens (Campos *et al.*, 2014). On the other hand, biotrophic pathogens which access essential nutrients from living plant tissues, trigger immunity based on the phytohormone salicylic acid, which

culminates in systemic acquired resistance, involving an increase in defensive capacity in an entire plant following a localised infection (Pieterse *et al.*, 2012).

And as with microbial infection, the immune pathways employed by plants in response to damage from insect herbivores is thought to differ based on the feeding mechanisms utilised by the insect itself (Gosset *et al.*, 2009). This is supported by the fact that the expression patterns of various immune pathway regulatory genes are distinct in each feeding guild e.g. chewing versus piercing insects (Howe and Jander, 2008). Under feeding by phloem-feeding insects, plants express genes involved in the salicylic acid pathway more highly than genes involved in the jasmonic acid pathway (Thompson and Goggin, 2006), the latter of which are only expressed at low levels, and may be a result of physical tissue damage by probing stylets, rather than an active immune response against insect herbivory. This mirrors plants' response to infection by biotrophic pathogens.

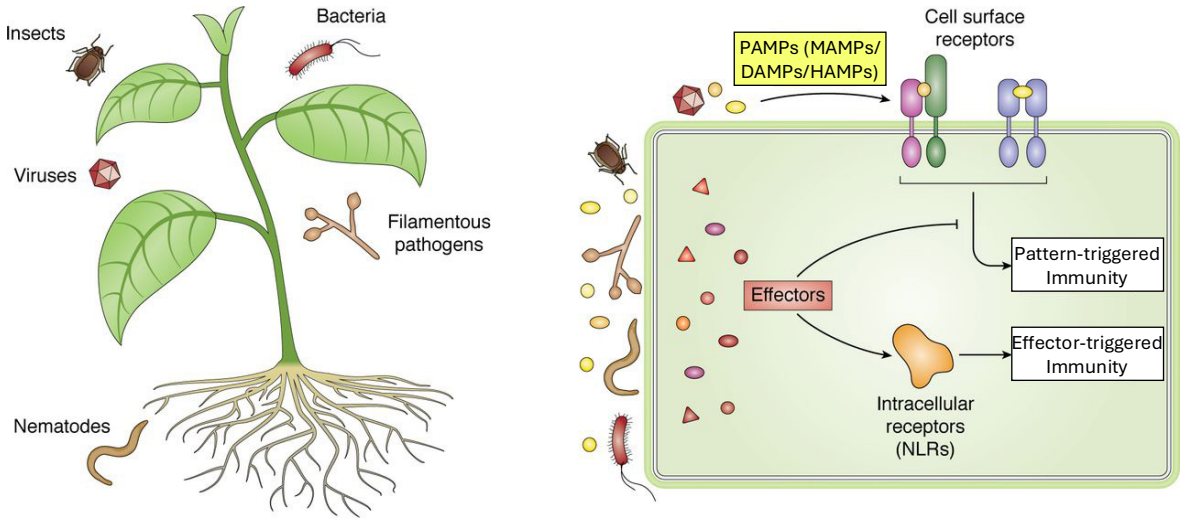


Figure 1.3: Plants launch a complex immune response during invasion by pathogens and pests. Illustration summarising the molecular response mounted by plants during pathogen invasion. Cell surface receptors recognise molecular patterns associated with pathogens and pests, which triggers a cascade of responses termed pattern-triggered immunity (PTI). In response, well-adapted pathogens and pests can deposit intracellular effectors to suppress PTI. However, these effectors may be recognised by intracellular receptors, which leads to effector-triggered immunity (ETI). Modified from (Bentham *et al.*, 2020) under a CC-BY License.

1.2.2: Plants detect pathogen and pest invasion in a process termed Pattern Triggered Immunity (PTI).

The first layer of defence in a plant-pathogen interaction is provided by the capacity of a host to identify molecules that are associated with pathogens. These molecules, termed microbe associated molecular patterns (MAMPs), or herbivore

associated molecular patterns (HAMPs) in the case of insect herbivory, are recognised by a plants' cell surface pattern recognition receptors (PRRs) (Mostafa *et al.*, 2022) (Figure 1.3). This perception of PAMPs by PRRs leads to a cascade of immune functions categorised as pattern triggered immunity (PTI). This process includes the accumulation of reactive oxygen species (ROS), sometimes referred to as a ROS burst, which relies on the induction of NADPH oxidases and calcium-dependent protein kinases, the latter of which is dependent on calcium ion (Ca^{2+}) flux into the plant cell (Jones *et al.*, 2024). Typically, this immune response is sufficient to suppress pathogen damage and further invasion, leading to a resistant phenotype (Zhang and Zhou, 2010).

Ligands for the induction of PTI are wide ranging and can be derived from several self- and non-self- components. Notable PTI elicitors include bacterial flagellin, including a truncated form termed flg22 (Chinchilla *et al.*, 2006); the fungal polymer chitin, a molecule which also makes up the exoskeletons of arthropods including insects (Wan *et al.*, 2008); and oligogalacturonides (Gravino *et al.*, 2017, Gravino *et al.*, 2025), plant cell wall derived oligomers which signal mechanical damage to plant cell walls, and are termed damage associated molecular patterns (DAMPs) (Benedetti *et al.*, 2015). When looking at the molecular interfaces involved, aphid infestation largely mirrors plant infection by microbial pathogens (Musaqaf *et al.*, 2023). Plants can identify molecular patterns associated with insect herbivory, though the identity of such patterns is less comprehensively characterised than in microbial pathogens.

HAMPs from a number of herbivorous insects across the chewing and piercing-sucking feeding guilds have been identified and characterised, with effects on a wide range of plant hosts (Snoeck *et al.*, 2022). While no aphid-derived HAMPs have been functionally characterised to date, there are a number of opportunities for plants to identify aphid invasion using PTI. The plant protein BAK1 is thought to interact with PRRs which recognise insect-derived HAMPs (Jaouannet *et al.*, 2014). The first comes from the presence of fungal-chitin responsive pattern recognition. Chitin, which is a structural component of the exoskeleton and mouthparts of insect pests, is also a key component of fungal cell walls, and has been identified as a fungal MAMP (Shibuya and Minami, 2001). It is thought that chitin induced PTI requires specific sizes of chitin fragments, which it is not yet

established that hemipterans produce during moulting (Kaloshian and Walling, 2016).

Another HAMP that has been identified in association with aphid infestation is *B. aphidicola* chaperonin protein growth requirement protein E, large subunit (GroEL). This protein has been shown to trigger plant immune responses following infiltration into plant tissues and leads to ROS burst and upregulation of PTI marker genes (Chaudhary *et al.*, 2014). *M. persicae* also shows reduced fecundity on plants expressing GroEL, indicating that such plants might have an increased resistance to aphid infection.

Beyond aphid species, there is also evidence that microbes associated with the honeydew of the brown planthopper *N. lugens* have been shown to elicit a number of immune responses in rice (*Oryza sativa*), including accumulation of phytoalexins and release of volatile organic compounds (Wari *et al.*, 2019). Elimination of honeydew microbiota using dietary antibiotics led to planthoppers which elicited a reduced degree of plant defence during feeding.

1.2.3: Adapted pathogens and pests can suppress PTI by deploying effectors.

Well-adapted pests and pathogens can supress the resistive effects of PTI, by deploying effector molecules (Figure 1.3). Additionally, it has been reported that aphids, like microbes, are able to make use of plant defence suppressing effectors in order to successfully colonise their hosts. While the mode of action of aphid effectors is less definitively understood, there have been numerous effectors which have been shown to improve aphid fitness on their selected hosts.

Unlike PAMPs, effectors are typically dispersed intracellularly, and as such the first step in plant perception of, and resistance to, pathogen effectors, are cytosolic nucleotide-binding leucine-rich repeat molecules (NLRs) (Pruitt *et al.*, 2021). Upon recognition of effector molecules by NLRs, there is a spike in reactive oxygen species within the infected host plant cell, eventually leading to apoptosis, and thereby inhibiting transmission of pathogens through a plant (Jones and Dangl, 2006). While the majority of described effectors are proteinaceous, there is evidence that other molecules, such as long non-coding RNAs (lncRNAs), may also play a similar role (Chen *et al.*, 2020).

A number of insect-derived candidate effectors have been identified, with diverse impacts on plant immune response (Table 1.1). The *A. pisum* effector protein C002 was identified to be necessary for the successful colonisation of fava bean, with knockdown mutants spending less time in contact with the phloem (Mutti *et al.*, 2008). C002 has been shown through immunohistochemistry to be synthesised in the aphid's salivary glands (Mutti *et al.*, 2008). In the work describing this protein, C002 was detected in fava bean plants following 24 hours of feeding by 500 adult *A. pisum*, but not in uninfested plants, suggesting it is translocated into plant hosts during feeding by pea aphids. Homologs of this effector have been found in a number of other species, including *Rhopalosiphum padi* (Escudero-Martinez *et al.*, 2020) and *M. persicae* (Bos *et al.*, 2010).

Through the use of transgenic plants, *M. persicae* has shown to be more reproductively successful on host plants producing the candidate effectors Mp1 and Mp2 (Pitino and Hogenhout, 2013). Mp1 has been shown to associate with the host plant's Vacuolar Protein Sorting Associated Protein 52 (VPS52), and aphids are able to supress production of this protein during infestation (Rodriguez *et al.*, 2017). Another protein, Armet, was upregulated during feeding by *A. pisum* in plant-feeding individuals when compared to diet-fed individuals (Wang *et al.*, 2015), suggesting an element of host-responsiveness for the synthesis and deployment of this effector. This protein was also shown to be secreted into the plant tissue as part of the watery saliva, and its knockdown led to detrimental changes to aphid feeding behaviour, further implicating the protein as a candidate effector. *In planta* expression of *Macrosiphum euphorbiae* genes encoding for *Me10* and *Me23* increased aphid fecundity on *Nicotiana benthamiana* (Atamian *et al.*, 2013).

Effector	Aphid species	Plant targets and mode of action	Host species	*In planta* localisation	Impact on aphid fitness	Reference
Ap4	*A. pisum*	*In planta* binding to PsBPL1/2, two proteins involved in plant immunity	*P. sativum*	Cytoplasmic puncta	Significant increase in fecundity of pea biotype on leaves expressing Ap4	Shih *et al.*, 2025
Ap25	*A. pisum*	Unknown	*P. sativum*	Unknown	Increase in fecundity of pea biotype on leaves expressing Ap25	Guy *et al.*, 2016
Armet	*A. pisum*	Unknown	*V. faba*	Unknown	Knockdown interferes with feeding behaviour	Wang *et al.*, 2015
ApC002	*A. pisum*	Unknown	*V. faba, A. thaliana, N. benthamiana*	Sieve elements	Increase in fecundity on plants over-expressing ApC002	Mutti *et al.*, 2008
CathB6	*M. persicae*	Recruitment of EDS1 to plant processing bodies	*A. thaliana*	Processing bodies	Increase in fecundity on plants over-expressing Cathepsin B6	Liu *et al.*, 2025
Mp1	*M. persicae*	Interacts with, and reduces levels of trafficking protein VPS52	*A. thaliana, N. benthamiana*	Prevacuolar compartments	Increases performance	Pitino and Hogenhout, 2013; Rodriguez *et al.*, 2017
Mp10	*M. persicae*	Inhibits DAMP triggered immunity and interacts with EDS1 mediated defence	*N. benthamiana*	Mesophyll cells adjacent to feeding tracks	Decrease in fecundity on plants over-expressing Mp10	Bos *et al.*, 2010, Gravino *et al.*, 2025
Mp42	*M. persicae*	Unknown	*N. benthamiana*	Unknown	Decrease in fecundity on plants over-expressing Mp42	Bos *et al.*, 2010
Mp56-58	*M. persicae*	Unknown	*N. benthamiana*	Unknown	Decrease in fecundity on plants over-expressing Mp56, Mp57, or Mp58	Elzinga *et al.*, 2014
Me10	*M. euphorbiae*	Interacts with TFT7, ultimate mechanism unknown	*S. lycopersicum*	Cytoplasm and nucleus	Increases performance	Atamian *et al.*, 2013
Me23	*M. euphorbiae*	Interacts with TFT7, ultimate mechanism unknown	*N. benthamiana*	Unknown	Decrease in fecundity on plants over-expressing Me23	Atamian *et al.*, 2013
Me47	*M. euphorbiae*	Encodes a glutathione S-transferase used to detoxify isothiocyanates	*S. lycopersicum, N. benthamiana*	Unknown	Increases fecundity on tomato, but reduces fecundity on *N. benthamiana*	Chaudhary *et al.*, 2015; Atamian *et al.*, 2013

Table 1.1: Species across the aphid phylogeny have evolved effectors to assist in host plant colonisation.
A majority of the described aphid effectors have been discovered in *M. persicae*, and another model species *A. pisum*. The effects of aphid effectors appear to be wide-ranging, from increasing aphid fecundity to extending the life span of the hemipteran pests. Though a number have been described, only a handful have been functionally characterised during aphid infestation. This table was adapted from Ray and Casteel, 2022.

There is also evidence that herbivores have co-opted proteins with diverse functions to act as effectors, most notably vitellogenin, a nutritious protein typically involved in reproduction, which has been implicated as an agent in the interaction between the small brown plant hopper *L. striatellus* and its host, rice. A C-terminal polypeptide from vitellogenin was shown to interact with the rice WRKY71 transcription factor, which is involved in the induction of plant resistance to small brown planthoppers. Silencing of vitellogenin in *L. striatellus* reduced feeding success and survival, with silenced nymphs eliciting higher hydrogen peroxide production in the host (Ji *et al.*, 2021).

It is also possible that herbivorous insects may benefit from effectors deployed by insect-vectored plant pathogens. Aster yellows witches' broom (AY-WB) phytoplasma, *Candidatus Phytoplasma asteris*, deploys a number of intracellular effectors, termed secreted AY-WB proteins (SAPs) to modulate immunity in the host plant (Bai *et al.*, 2009). This can provide a fitness benefit to the insect vector *Macrosteles quadrilineatus*, which show higher reproductive success on AY-WB infected *A. thaliana*, as well as plant lines expressing the SAP11 effector (Sugio *et al.*, 2011). SAP54 modulates floral development to give rise to leaf-like flowers in infected *A. thaliana* (MacLean *et al.*, 2011), upon which adult female *M. quadrilineatus* preferentially oviposit when compared to wild type flowers, suggesting improved vector colonisation potential conferred by the presence of AY-WB (MacLean *et al.*, 2014).

1.2.4: Plants respond to aphid effectors in a process termed Effector Triggered Immunity (ETI).

There are a number of aphid-responsive NLRs which are thought to detect aphid effectors. An NLR-encoding gene in tomato, *Mi*, so named due to the conferred resistance to the root knot nematode *Meloidogyne incognita*, was also shown to provide tomatoes with resistance to the potato aphid *Macrosiphum euphorbiae* (Rossi *et al.*, 1998). Additionally, the melon *Vat* (Virus aphid transmission) gene confers resistance to the melon/cotton aphid *Aphis gossypii* (Boissot *et al.*, 2016). Resistance genes to a number of aphid species have been identified in a wide breadth of plant species, including legumes, fruit trees, vegetables and cereals (Dogimont *et al.*, 2010).

The perception of aphid effectors by NLRs has led to a gene-for-gene immunity model, by which a single NLR recognises one effector, leading to an evolutionary arms race in specialist pathogens. Within this arms race, an effector can bypass recognition by a plant host's NLR through mutations. In the interaction between the brown rice planthopper, *N. lugens*, and its host plant, rice, a number of insect biotypes are able, or unable, to colonise host plants based on the presence of particular NLR or PRR genes within the host. At least 24 such resistance genes have been detected in rice (Cheng *et al.*, 2013).

However, the molecular basis of aphid resistance described above exclusively explores interactions between plants and specialist aphid species. How plants are able to detect and resist colonisation by generalist species, such as *M. persicae*, is less well known. It is unlikely that generalist aphid species have evolved ways to break down diverse plant metabolites, and as such, these aphids may have evolved a way to suppress production of specialist metabolites in plants, which may be conserved across plant phylogeny.

1.3: *M. persicae* deploys a repertoire of effectors to modulate PTI

Even within *M. persicae*, there is significant evidence for the deployment of a broad range of molecules with effector function, which may offer a route to the vast polyphagy observed in the species. These molecules, despite their shared ultimate function, vary greatly in their structure, composition, and evolutionary trajectory. There are a number of described *M. persicae* effectors which are anciently conserved singletons with homologs across hemiptera, and a number which belong to dynamically expanding gene families with varying degrees of specificity across the aphid phylogeny. Though the majority of these effectors are proteinaceous, a number of other classifications of molecules, most notably lncRNA transcripts, have been implicated in the modulation of plant host defence. These effectors have been identified as being translocated into host plants upon aphid feeding and elicit a spectrum of varying phenotypes across aphid and plant performance metrics.

Mp10 is an *M. persicae* chemosensory protein which has also been described as a candidate effector. Mp10 alters aphid fecundity (Bos *et al.*, 2010), and is delivered into plant mesophyll cells (Mugford *et al.*, 2016). It is also known to bind to the

acrostyle (Deshoux *et al.*, 2022), the structure at the tip of the aphid stylets which have also been implicated in binding aphid-transmitted plant viruses during feeding (Uzest *et al.*, 2010). Mp10 is thought to act by supressing PTI by interacting with enzymes which modulate pattern recognition kinase homeostasis (Gravino *et al.*, 2025). This protein appears to be evolutionarily ancient, with orthologues found across other sap-feeding hemipterans also supressing PTI by supressing ROS and Ca^{2+} bursts in response to the bacterial flagellin-derived elicitor flg22.

However, not all previously described effectors are so broadly conserved, nor do effector repertoires consist solely of singletons. For example, members of the Cathepsin B (CathB) cysteine protease family have also been explored due to a number of effector-like characteristics. Despite this family of proteins being found across animal phylogeny, primarily with digestive function, the evolutionary trajectory of the gene family within the lineage of *M. persicae* is interesting due to the presence of a recent expansion within the CathB family (Mathers *et al.*, 2017, Liu *et al.*, 2025). Firstly, transcriptomic analysis of *M. persicae* colonies adjusting to host-swap between phylogenetically distant host plant species reveals that CathB genes comprise part of a co-regulated cluster which show host-specific regulation, implicating them in the process of adaptation to novel hosts and the establishment of colonies following dispersal. Plant-mediated RNA interference of CathB genes, which relies on plants to express double stranded RNA molecules to be taken up by aphids upon feeding, which in turn reduces the expression of CathB in the aphid, has been shown to reduce aphid survival and fecundity (Mathers *et al.*, 2017). This suggests a fitness benefit for expression and deposition of CathB proteins in the life history of *M. persicae*. Follow-up proteomic analysis of aphid saliva has identified a number of CathB-associated peptides, pointing to direct transfer of peptides into a plant host through the saliva (Liu *et al.*, 2024). The peptides found in saliva are associated with CathB proteins belonging to the clade which has been recently expanded in *M. persicae*.

Additionally, transcripts from a number of long non-coding RNAs within the *M. persicae Ya* family, named for the Chinese word for aphids, were identified as being translocated into plants upon feeding, where three of the thirty members were identified as having migrated to distal leaves. Interestingly, this family

contains 30 members arranged as tandem repeats in the *M. persicae* genome (Chen *et al.*, 2020). The mobility of these transcripts might suggest a far-reaching effect in the modulation of plant immune response. One *Ya* transcript, *Ya1*, was identified as a virulence factor through the generation of stable transgenic *A. thaliana* plants (35S:Ya1), upon which *M. persicae* fecundity, which is an assumed proxy for fitness, was significantly increased when compared to wild-type (Col-0) and empty vector (35S:GFP) controls (Chen *et al.*, 2020).

1.3.1: CathB proteins have been identified as candidate effectors in M. persicae.

CathB proteins belong to a larger group of cysteine proteases, which are found across vertebrates and invertebrates, which are typically stored in the lysosomes (Turk *et al.*, 2000). In Lepidoptera, CathB proteins, amongst other Cathepsin proteins, are involved in metamorphosis and eclosion (Saikhedkar *et al.*, 2015). The soldier morphs of a social aphid species, *Tuberaphis styraci*, express a CathB protein in the midgut, which they use as a defensive venom to incapacitate predatory threats to the colony (Kutsukake *et al.*, 2004, Kutsukake *et al.*, 2008).

Notably, however, CathB proteins in *M. persicae* have been implicated in the manipulation of host plant immunity, pointing to their function as candidate effectors. This initial identification and functional analysis has relied on a breadth of molecular and -omics based approaches. For example, twelve members of the *M. persicae* CathB gene family were identified as being differentially expressed upon host swap (Mathers *et al.*, 2017), and one member of the family, CathB3, was detected in *Nicotiana tabacum* leaves infected with non-tobacco adapted *M. persicae* through western blotting (Guo *et al.*, 2020). The same study also showed that the same effector was present to a higher extent in the saliva of individuals from the non-tobacco-adapted *M. persicae* lineage, which might suggest a role for CathB3 in the specialisation of an otherwise generalist species (Guo *et al.*, 2020). Transient expression of CathB3 diminished fitness of tobacco adapted *M. persicae* by reducing survival and feeding efficiency, potentially as a result of interactions with the serine/threonine protein kinase Enhanced Disease Resistance 1-like (EDR1-like), which is capable of suppressing aphid feeding efficiency and activating phloem localised bursts of reactive oxygen species.

Mass spectrometry analysis of *M. persicae* saliva identified peptides corresponding to a number of CathB proteins (Liu *et al.*, 2024). Aphid CathB proteins consist of three domains: a signal peptide, a propeptide, and a mature domain. Based on the results of this proteomic analysis, it appears that only mature domains are present in the saliva, suggesting cleavage of the signal and propeptides prior to the oral secretion of CathBs (Liu *et al.*, 2025).

CathB6, one of the highly orally secreted CathB proteins, was shown to localise to dynamic punctate structures of varying sizes within the cytoplasm of *N. benthamiana* during transient expression (Figure 1.4A and 1.4B). These structures, which also showed strong localisation of the processing body (p-body) marker Varicose (VCS) (Figure 1.4C), were seen to migrate across cells and fuse with each other to form larger aggregates, suggesting that CathB6 is sequestered to processing bodies *in planta* (Liu *et al.*, 2025). The colocalization of CathB6 with VCS suggests that these punctate structures are p-bodies, which generally function to regulate the expression and degradation of mRNAs and proteins in the cell (Kearly *et al.*, 2024). Proximity-labelling mass spectrometry (PL-MS) was used to search for plant targets of CathB6, and identified EDS1 as an interactor of CathB6. EDS1 regulates several key aspects of plant immunity, including effector triggered immunity and systemic acquired resistance (Dongus and Parker, 2021). *M. persicae* reared on *A. thaliana* plants expressing CathB6 showed increased fecundity when compared to individuals reared on wildtype plants, suggesting that CathB6 has direct impacts on aphid performance (Liu *et al.*, 2025).

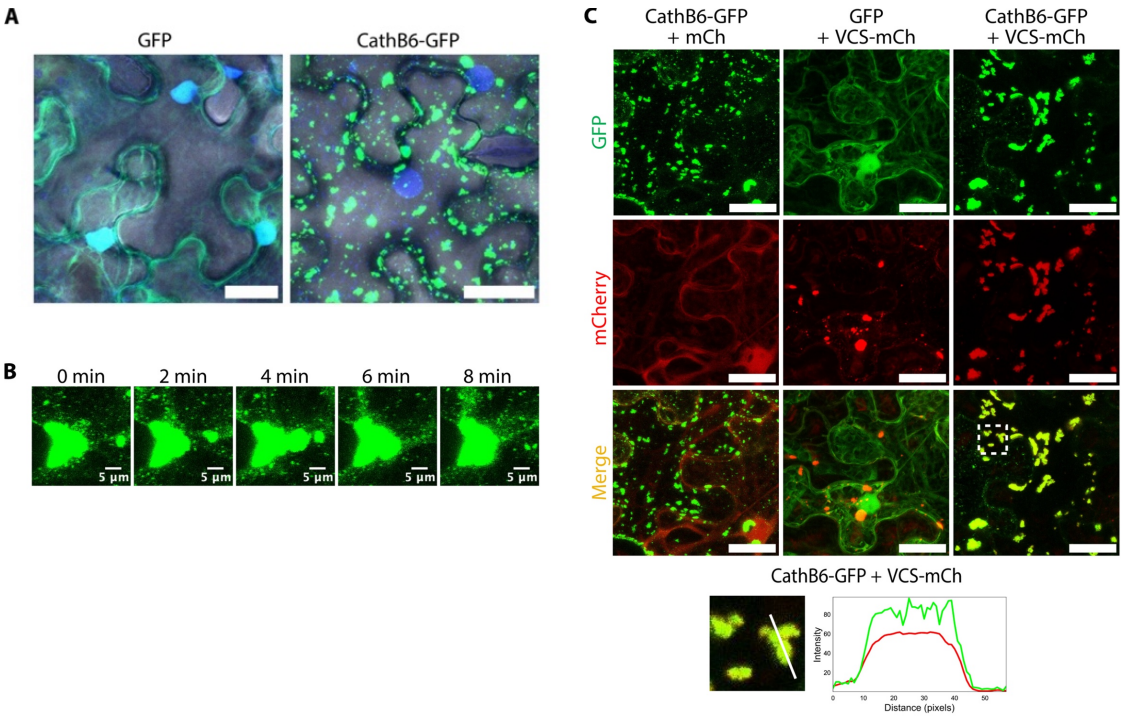


Figure 1.4: CathB6 forms dynamic puncta in *N. benthamiana* cells which correspond to processing bodies.
Confocal microscopy showing localisation of GFP-tagged CathB6 during Agrobacterium-mediated transient expression in *N. benthamiana* cells. **A:** CathB6 forms puncta in plant cells. Blue – DAPI, Green – GFP or CathB6-GFP. **B:** Puncta change shape and aggregate over time. Green – CathB6-GFP. **C:** CathB6-GFP puncta overlap with p-body marker Varicose, suggesting that puncta correspond to p-bodies. Figure reproduced from Liu *et al.*, 2025, under a Creative Commons License.

1.4: Heat shock protein biology

In order to respond to biotic and abiotic stresses, organisms including plants are able to modify the expression of a number of stress response genes. This includes the heat shock genes, which are upregulated in response to heat stress, and lead to the production of heat shock proteins (HSPs). These proteins assist in the maintenance of homeostasis in heat stressed cells, preventing damage and conferring increased organismal fitness (Swindell *et al.*, 2007). Heat shock proteins consist of several smaller protein families, termed the HSP100, HSP90, HSP70, HSP60, and small heat shock proteins (sHSPs) (Weston *et al.*, 2011). Broadly, HSPs act as molecular chaperones, and work to guide folding of their target proteins, as well as preventing the misaggregation of proteins, though various HSPs have also been implicated in the regulation of plant immunity to viral infection (Verchot, 2012).

Most sHSPs have molecular weights between 12 kD and 42 kD, earning the family the alternative identifier of HSP20 proteins (Bourgine and Guihur, 2021). sHSPs

are ubiquitous across the domains of life, being found within Archaea, Eukaryota and Bacteria (Roy *et al*., 2022). sHSPs are typically characterised by a relatively highly conserved alpha-crystallin domain (ACD), and as such appear to be evolutionarily related to the vertebrate lens protein alpha-crystallin (Ingolia and Craig, 1982). This domain usually takes the form of two opposing β-sheets forming a sandwich, separated by a hydrophobic domain which can be variable in size (Paul *et al.*, 2016). There are cases in which ACD-containing proteins may be multidomain proteins with other unrelated domains (Scharf *et al.*, 2001), but relatively little is known about the function or structure of such proteins.

Plants appear to have an expanded diversity of sHSPs compared to other eukaryotes (Waters, 2012), with sHSP counts between 19 in *A. thaliana* (Scharf *et al.*, 2001), and 109 in wheat, *Triticum aestivum*, (Muthusamy *et al.*, 2017). In angiosperms, there are 11 distinct subfamilies of sHSPs, including 5 subfamilies which localise to specific organelles, including the endoplasmic reticulum, peroxisomes, chloroplasts, and mitochondria (Waters, 2012).

There is evidence that in order to fulfil their role as chaperone proteins, sHSPs form oligomers formed from multiple individual subunits. For example, a crystal structure of the *T. aestivum* sHSP HSP16.9 shows the formation of a dodecamer made up of six dimeric subunits, creating a "hollow ball" which is used to bind unfolded substrate (van Montfort *et al.*, 2001). This remains, however, the only plant sHSP with a resolved oligomeric crystal structure. The pea (*Pisum sativum*) sHSP HSP18.1 has been shown through mass spectrometry to assemble almost exclusively into dodecamers at 22 °C, but this preferential arrangement appears to be diminished at higher temperatures, in which HSP18.1 dodecamers appear to be dissociated into monomers and dimers, or assembled into higher order oligomers with between 13 and 20 monomeric subunits (Stengel *et al.*, 2010).

There is also evidence that a number of ACD proteins are involved in silencing genes and transposable elements in *A. thaliana* (Boone *et al.*, 2023) by forming complexes with methyl-CpG binding domain (MBD) proteins MBD5 and MBD6, leading to aggregations of these proteins at methylation sites in a gene's promotor region, which prevents its transcription (Boone *et al.*, 2024).

1.4.1: Some sHSPs act as regulators of plant immunity.

Several sHSPs and other ACD-containing proteins in plants have been identified as being involved in the perception and suppression of invasion by plant pathogens and pests. In *N. benthamiana*, silencing of the sHSP NbHSP17 accelerated the growth of *Ralstonia solanacearum*, a soilborne pathogen which causes bacterial wilt. This silencing also reduced expression of defence related genes PR1 and PR4, marker genes for salicylic acid signalling and methyl-jasmonate signalling respectively (Maimbo *et al.*, 2007). This suggests a crucial role for NbHSP17 in disease resistance against this bacterium.

The *A. thaliana* ACD protein Restricted TEV Movement 2 (RTM2) is involved in the restriction of long-distance movement of *Tobacco Etch Virus* (TEV) (Whitham *et al.*, 2000) and other potyviruses (Decroocq *et al.*, 2006). This restriction appears to be highly specific to potyviruses, and completely diminishes viral titres in systemic tissues. rtm1 loss-of-function *A. thaliana* mutants do not show the same movement restriction for TEV (Whitham *et al.*, 2000).

Another ACD-containing protein from *A. thaliana*, sieve element-lining chaperone 1, or SLI1, is a 54.2 kD protein which has been characterised as a *M. persicae* resistance factor by genome-wide association mapping of aphid feeding behaviour on 350 *A. thaliana* accessions (Kloth *et al.*, 2017). Electrical penetration graphs showed that *M. persicae* displayed reduced feeding performance on sli1 loss-of-function Arabidopsis mutants when compared with the Col-0 wildtype control (Kloth *et al.*, 2017). The sli1 mutants lead to more sustained periods of phloem feeding, as well as a reduced proportion of time spent salivating during feeding. This led to a notable increase in the reproductive success of the aphids on the mutant plants, suggesting that SLI1 has a role in suppressing the fecundity of *M. persicae* on *A. thaliana*. However, an expanded screen of sli1 loss-of-function mutants suggests that SLI1-mediated resistance is not specific to *M. persicae* (Kloth *et al.*, 2021). An expanded screen of feeding behaviour of *M. persicae* and 4 other phloem-feeding species revealed one additional aphid species, *Brevicoryne brassicae*, and a whitefly species, *Aleyrodes proletella*, are also susceptible to fitness suppression by plants expressing SLI1. Reproductive success of *B. brassicae* and *A. proletella* was significantly reduced on Col-0 plants when compared to the sli1 mutants, with number of offspring or eggs being reduced by

at least 20%. *Myzus persicae nicotianae* and *B. brassicae* feeding on sli1 mutants outperformed those feeding on Col-0 plants. They spent less time salivating into the phloem prior to the establishment of feeding and displayed an increase in feeding duration on mutant plants. The aphid species which showed no reduction in fecundity, *Lipaphis ersyimi*, engaged in extended phloem feeding events on Col-0 than *M. persicae* and *B. brassicae*, and showed no significant divergence in behaviours during feeding between the wildtype and mutant hosts.

Another ACD-containing protein from *A. thaliana* has been implicated in immune response to the *M. persicae* CathB6 effector. The 28.9 kD ACD protein ACD28.9 was identified as an interactor of CathB6 through PL-MS (Liu *et al.*, 2024), and subsequent *in planta* co-expression analysis of ACD28.9 and CathB6 showed that ACD28.9 was responsible for the rescue of CathB6-EDS1 complexes from p-bodies, instead eliciting a cytoplasmic distribution of CathB6 (Liu *et al.*, 2025). *M. persicae* fecundity was increased on ACD28.9 knockout mutants of *A. thaliana* when compared to aphids reared on Col-0 wildtype plants, demonstrating that this ACD protein contributes to *A. thaliana* resistance to *M. persicae*.

1.4.2: Existing phylogenies of plant sHSPs focus on single-species protein families.

Genome-wide identification of HSPs has been performed for a number of plant species through a number of *in silico* methodologies (Table 2). Mining the genomes of plant species has proved powerful in the identification of a number of sHSPs, but larger phylogenies of plant sHSPs are yet to be performed, with previous studies focussing on a single species or group alongside model species.

Identification of HSP20 family members has been performed through the use of Hidden Markov Models (HMMs) of the conserved alpha-crystallin domains which are characteristic of small heat shock proteins. These domains are sufficiently highly conserved to be identified through sequence analysis in comparison to the reference HMM. This method was integral to the identification of small heat shock proteins in pumpkin (Hu *et al.*, 2021). HMMs have also been used for the identification and downstream analysis of other domains in a wide range of proteins. Notably, HMMs for protein kinase and leucine-rich repeat domains were utilised to identify and create a phylogeny of a *Phytopthora*-responsive immune receptor from wild potato called PERU (Torres Ascurra *et al.*, 2023).

Plant Species	Number of sHSPs	Method of identification	Reference
Arabidopsis (*Arabidopsis thaliana*)	19	Genome search	Scharf *et al.*, 2001
Rice (*Oryza sativa*)	23	Genome BLAST	Sarkar *et al.*, 2009
Cucumber (*Cucumis sativus*)	30	Domain analysis	Huang *et al.*, 2022
Pumpkin (*Cucurbita moschata*)	33	Genome BLAST	Hu *et al.*, 2021
Lettuce (*Lactuca sativa*)	36	Genome BLAST and HMM Search	Zhang *et al.*, 2024
Peach (*Prunus persica*)	42	HMM Search	Lian *et al.*, 2022
Sorghum (*Sorghum bicolor*)	47	NCBI search and genome BLAST	Nagaraju *et al.*, 2020
Potato (*Solanum tuberosum*)	48	HMM Search	Zhao *et al.*, 2018
Poplar (*Populus yunnanensis*)	53	Genome BLAST and domain analysis	Shi *et al.*, 2025
Wheat (*Triticum aestivum)*	109	Genome BLAST and HMM Search	Muthusamy *et al.*, 2017

Table 1.2: Several plant species have high numbers of sHSPs, which have been identified through various genomic analyses.
Selection of genome-wide sHSP identification studies for plant species, including the number of sHSPs identified as well as the methods used for identification.

1.5: Description of investigations

The aphid species *M. persicae* is an extremely polyphagous pest of plants, which is capable of modifying plant defensive processes using a broad repertoire of effectors, in order to facilitate the colonisation of a spectrum of plant hosts. This thesis aims to explore the transcriptional landscape of candidate *M. persicae* effectors, as well as determine potential modes of virulence *in planta*.

In Chapter 2, I describe the generation and analysis of a dataset consisting of organ-specific RNA-seq libraries from adult *M. persicae*. In tandem with previously established proteomic analysis of aphid oral secretions, I explore the contribution of various aphid organs to the effectorome.

Work described in Chapter 3 focuses on a family of candidate *M. persicae* effectors, the CathB proteins, which show high organ specificity and an interesting evolutionary trajectory. Using sequence analysis and molecular biology

approaches, I investigate the basis of binding affinity of proteins across this family to a characterised *A. thaliana* target, ACD protein ACD28.9.

Chapter 4 explores the evolutionary dynamics of the small heat shock / alpha-crystallin domain protein family in *A. thaliana* and across the plant phylogeny, which contains the CathB binding target ACD28.9. I also examine the potential regions responsible for ACD28.9 binding to CathB proteins, offering additional insight into the interaction between the two protein families.

In Chapter 5, I assess the significance of the work described in the course of the generation of this thesis, and consider how additional investigations can be performed to further propel the understanding of the evolutionary arms race between *M. persicae* and its array of host plants.

1.6: Contributions to this thesis

All experiments described herein were designed and performed by me. Proteomic analysis of *M. persicae* oral secretions which was used in tandem with transcriptomic analysis in Chapter 2 was performed by Dr Qun Liu (Hogenhout Lab alumni, John Innes Centre) and was published as a dataset on Zenodo (Liu *et al.*, 2024). The phylogenetic tree for the *M. persicae* CathB family used in figures across Chapters 2 and 3 was constructed and shared by Dr Mar Marzo (Hogenhout Lab, John Innes Centre), who is acknowledged in relevant figure legends. Dr Marzo also assisted in the identification of the genomic co-ordinates of the CathB gene clusters. The bioinformatic pipeline and plant proteomes used for the identification of alpha-crystallin domain proteins in Chapter 4 was kindly provided by AmirAli Toghani (Kamoun Lab, The Sainsbury Laboratory), based on methods described in Torres Ascurra *et al.*, 2023.

Chapter 2 - Generating and mining an organ-specific RNA-seq dataset for insights into effector expression.

2.1: Introduction

Aphids are highly adapted to feed on a diet of sap, which they access from the phloem of host plants using specialised piercing mouthparts termed stylets. This stylet is comprised of two exterior mandibular stylets, which encapsulate the inner maxillary stylets (Figure 1.2A). These four structures have been elongated to allow for piercing of plant tissues.

Aphid feeding has been characterised to consist of a number of distinct behavioural phases (Tjallingii, 2006), with aphid feeding beginning with a probing phase in order to assess host quality, and during which most non-persistent viruses are transmitted. During this phase, aphids will pierce cells to determine aspects such as pH, amino acid composition, and sucrose concentration (Hewer *et al.*, 2011). This is followed by a salivary phase, in which aphid saliva is injected into the plant to inhibit plant defence. Then, once a feeding site is established, the stylets access the phloem and feeding begins.

During feeding, aphids have been shown to produce two distinct types of saliva. The first, gelling saliva, is secreted during the movement of aphid stylets through the apoplast, and forms a salivary sheath around the stylets (Will and Vilcinskas, 2015). The latter, watery saliva, is secreted during the penetration and ingestion of sieve elements (Will *et al.*, 2007), and has, for instance, been shown to limit the levels of sieve element occlusion within the host plant (Will *et al.*, 2009). There is evidence that these two forms of saliva may have different origins within the aphid salivary gland, with watery saliva thought to originate mostly in the accessory salivary gland, and gel saliva in the principal salivary gland (Moreno *et al.*, 2011).

Also present in the saliva of aphids such as *Myzus persicae* are effectors, which improve aphid performance by modulating the basal immune response of host plants. A number of effectors have been identified in *M. persicae*, including the chemosensory protein Mp10, which was described as a candidate effector due to its capability to alter aphid fecundity on plants transiently expressing Mp10 (Bos *et al.*, 2010). Mp10 has been shown to be delivered by aphids directly into plant mesophyll cells (Mugford *et al.*, 2016), and is known to bind to a structure at the proximal tip of the stylets termed the acrostyle (Deshoux *et al.*, 2022).

Moreover, the *M. persicae* Cathepsin B (CathB) family has been identified as a family of host-responsive genes, which function to improve several aspects of aphid performance on a number of hosts (Mathers *et al.*, 2017). One member of this gene family, CathB3, has been shown to be upregulated in the heads of non-tobacco-adapted *M. persicae* lineages when compared with head samples from tobacco-adapted individuals (Guo *et al.*, 2020), while another, CathB6, has been shown to increase aphid fecundity on transgenic *Arabidopsis thaliana* (Liu *et al.*, 2025).

As such, the organ-specific RNA sequencing dataset generated and analysed in the following chapter seeks to address the hypothesis that the broad repertoire of *M. persicae* effectors are synthesised in the salivary glands, where they are secreted into the saliva to be translocated into the plant cell during feeding, in order to modulate the host's defence response and facilitate successful establishment and proliferation of aphid colonies. With particular focus on the transcriptomes of the salivary glands and stylets, based on the expression patterns of previously described effectors such as C002 in *Acyrthosiphon pisum* (Mutti *et al.*, 2008) and the binding of Mp10 to the acrostyle respectively, the synthesised dataset also included organs intended to offer comparative power when examining effector expression. Using this data in tandem with existing proteomics analysis from purified *M. persicae* saliva enabled a targeted search for novel effector candidates, as well as offered additional insight into previously described effectors.

2.2: Materials and methods

2.2.1: Aphid rearing, dissection and organ isolation

The *M. persicae* clone O colony has been sustained since 2010. This colony has been reared on Chinese cabbage *B. rapa* or *A. thaliana* Col-0 and maintained in a growth chamber (20°C, 14 h light / 10 h dark, 75% humidity).

In order to generate an age-matched population for dissection to remove age-based variation in gene expression, adults from the rearing colony were transferred from *B. rapa* stock plants and into clip cages containing 15 adults each on leaves of *B. rapa*. After 24 hours, all adults were removed from the experimental plant and clip cages were replaced to allow <1-day old nymphs to

mature. After 6 days under the same growth conditions as the rearing colony (20°C, 14 h light / 10 h dark, 75% humidity), remaining 7-day old aphids were removed from the clip cages and selected for dissection.

To generate organ-specific tissue samples, aphids were anaesthetised using dry ice, and organs were isolated in ice-cold PBS for three periods of thirty minutes. Following isolation, organs from each aphid were dropped into 1.5 mL LoBind Eppendorf tubes containing 25 µL PBS with 400 U/mL RNAseOut (Invitrogen) on ice to create a pooled subsample. After each 30 minutes of dissection, subsamples containing isolated organ samples were flash frozen in liquid nitrogen.

2.2.2: RNA extraction from organ tissues

Total RNA was extracted from tissue samples using a column-based extraction kit. Each 25 µL subsample was mechanically lysed while still frozen, using a polyethylene pestle and then solubilised in 100 uL TRIzol. For each total RNA sample, three tissue subsamples from the same round of dissection were pooled following lysis in TRIzol. 375 µL of 100% ethanol was added, before total RNA was extracted from all dissected organ samples from each replicate in tandem using the Direct-zol RNA microprep kit (Zymo, catalog no. R2061). To ensure highly concentrated RNA, isolated total RNA was eluted into 8 µL nuclease-free water, and RNA concentration was quantified using Qubit HS RNA assay. The quality of RNA in initial samples for each organ was determined using a Bioanalyser High Sensitivity RNA assay. Samples were flash frozen in liquid nitrogen and stored at -80 °c until being submitted for sequencing.

2.2.3: Sequencing by Novogene

Total RNA samples of sufficient yield were shipped on dry ice to Novogene, Cambridge, where the standard protocol was used to generate 150bp, strand-specific libraries. In summary, mRNA was purified from total RNA using magnetic poly-T oligo-beads. Purified mRNA was fragmented, and first strand cDNA was synthesized using random hexamer primers. Second strand cDNA was synthesized using dUTP. Directional libraries were prepared through end repair, A-tailing, adapter ligation, size selection, USER enzyme digestion, amplification, and purification. Qubit and real-time PCR was performed for library quantification and bioanalyzer assays were used for library size distribution detection. Quantified

libraries were pooled and sequenced on Illumina platforms, according to effective library concentration and data amount. Samples were sequenced using the Illumina NovaSeq platform and generated approximately 20 million 150 bp paired-end reads for each sample.

2.2.4: Expression analysis

To ensure high quality and robustness of generated RNA-seq libraries, sequencing output files containing read data were checked for quality using FastQC, and reads were trimmed for quality and length using TrimGalore (Martin, 2011). Reads were mapped to the *M.persicae* Clone O v2.1 reference genome (Mathers *et al.*, 2017) using HISAT2 (Kim *et al.*, 2019) and SAMtools (Li *et al.*, 2009). Read counts were generated using HTSeq-count. Ambiguously aligning reads were assigned to their aligning genes proportionately based on the number of unambiguous reads assigned to each gene, using the '--nonunique fraction' option for HTSeq-count tool. Code for the alignment and quantification of reads can be found in Supplementary Figure 1.

To perform general analysis of gene expression across tissues, transcript per million (TPM) values were calculated in R v.4.3.1 "Beagle Scouts" (RCoreTeam, 2023), using the raw read numbers generated by HTSeq-count and gene lengths extracted from the *M. persicae* Clone O v2.1 reference genome.

Differential expression analyses were performed in R using DESeq2 (Love *et al.*, 2014). Transcripts per million (TPM) values were generated in R, using gene lengths extracted from the *M. persicae* Clone O v2.1 reference genome. Plots were generated in R using ggplot2 (Wickham, 2016).

To determine organ-specific up- and down-regulated genes, DESeq2 analysis was performed on each tissue individually, with metadata labelling samples as either the organ of interest or 'other' (for example, when identifying genes expressed most highly in the salivary glands, samples from other tissues were assigned "Other" in the tissue metadata group). Genes with a $\log_2$ fold change of ≥1 or ≤-1, and an adjusted *p*-value of ≤0.05 were labelled as up- or down-regulated respectively.

2.2.5: Enrichment analysis

In order to assess the function or shared features of differentially expressed genes in each isolated organ, an Interpro scan output detailing the functional analysis of the *M. persicae* clone O v2.1 proteome was used to generate a gene matrix transposed (GMT) file displaying the gene ontologies (GO) assigned to each gene in the genome. A similar GMT file was created detailing PFAM domain predictions for each gene. The g:Profiler web tool (Kolberg *et al.*, 2023) was used to perform enrichment analysis by using the GMT file as a database and a list of upregulated genes in each organ as the query. An additional search was performed on the list of orally secreted proteins. The output of this tool, including enriched GOs or domains, gene number, and *p*-value, was saved and filtered. For gene lists with >20 enrichments, the top 20 enriched ontologies or domains based on *p*-value were filtered and dotplots of enrichment were created in RStudio with the ggplot2 package.

2.2.6: Mass spectrometry of oral secretions

Mass spectrometry was performed by Dr Qun Liu and methods were described alongside the data in a publication on Zenodo (Liu *et al.*, 2024). Lists of proteins contained in oral secretions based on this experiment, including the master list containing only representative proteins for larger protein families, were provided for analysis by Dr Qun Liu and Dr Sam Mugford.

2.2.7: Analysis of expression across the CathB family

For the comparison of expression of CathB family genes across aphid organs, *M. persicae* gene lengths were generated and used to calculate transcripts per million (TPM) values for all genes in the genome. This dataset was subsetted to only include CathB genes based on the *M. persicae* clone O v2.1 gene models. Visualisation of this data was performed with ggplot2, and a phylogenetic tree of CathB genes was created by Mar Marzo, with orally secreted CathB genes highlighted.

2.3: Results

2.3.1: Aphid organs were successfully and efficiently isolated for extraction of RNA.

For consistency in organ identification, initial rounds of dissection and organ isolation were used to generate images of each organ for reference in later rounds during data collection (Figure 2.1). An illustration was also generated to display the location of each organ, as a framework for the dissection process. Aphids were anaesthetised using dry ice and dissected in a petri dish, remaining on *Brassica rapa* leaves until being selected for dissection. Winged individuals were excluded from dissection experiments to avoid morph-specific transcriptomic variation. Leg, antenna, and stylet samples were removed prior to opening the abdomen, and salivary glands were removed through the space left by the removal of the stylet. Subsequently, the abdomen was opened, and the foregut, hindgut, and ovaries were separated and isolated.

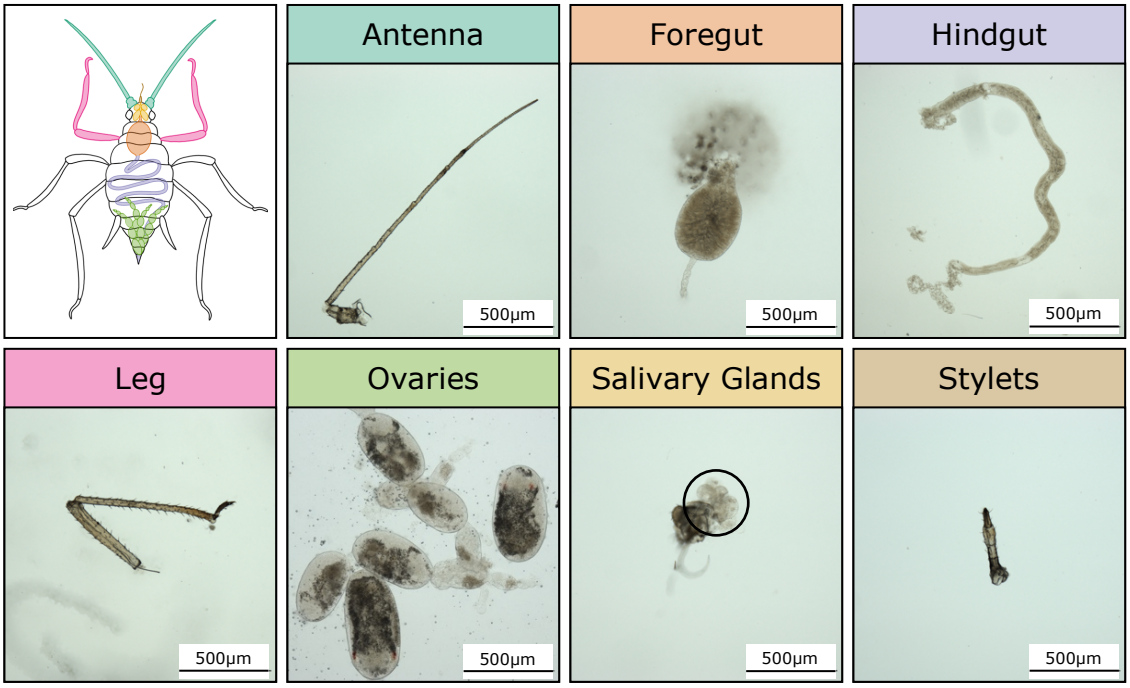


Figure 2.1: Reference images of isolated organs were generated prior to isolation of organs for RNA-seq.
Representative organs from dissected individuals were isolated in 1× phosphate buffered saline, then visualised and imaged on an Axio Zoom V16 II stereo microscope. Illustration showing the *in situ* location of organs is colourised based on organ label.

2.3.2: Organ-specific RNA-seq libraries are of high quality, and are correctly stranded.

Based on previous work undertaken in the lab to generate RNA-seq libraries from small amounts of tissue (Giolai, 2019), a SMART-Seq2 approach was initially trialled on single organs from single individuals. However, this complex, multistep approach led to extremely low yields of cDNA suitable for library preparation. A number of steps were taken to optimise yield from this protocol, but these were ultimately unsuccessful. As such, organs from a number of individuals were pooled to increase the volume of starting material, and the Direct-zol RNA microprep kit (Zymo, UK) was used as a more efficient means of RNA extraction. To ensure appropriate RNA quality, RNA extractions were analysed for RNA integrity (RIN) using the Bioanalyser High Sensitivity RNA assay. Once RNA extractions were consistently achieving RIN scores above 5, samples were collected for sequencing. Full metadata for RNA samples generated was recorded in case of variability arising between replicates. Dissection was performed and isolated organs were ultimately pooled into 36 samples, containing organs from between 12 and 21 adult aphids, across 7 organ types: antennae (n = 5), foregut (n = 7), hindgut (n = 4), legs (n = 5), ovaries (n = 4), salivary glands (n = 6), and stylets (n = 5) (Table 2.1).

Leg, antenna, and stylet samples typically produced lower total RNA yield, possibly due to reduced efficiency of mechanical lysis associated with the chitinaceous component of the samples. Novogene performed library preparation and sequencing steps according to their standard protocol to generate strand-specific, paired end libraries. Total RNA samples submitted to Novogene ranged in RNA quantity from 25 ng to 2.96 µg.

Sample Name	Pooled Organ Count	RNA Concentration (ng/µl)	Total RNA Quantity (ng)
Antenna_1	22	12.60	75.60
Antenna_11	36	16.8	168
Antenna_2	34	35.20	211.20
Antenna_6	42	40.00	240.00
Antenna_8	29	34.40	206.40
Foregut_1	12	126.00	756.00
Foregut_11	13	140.0	1400
Foregut_12	16	112.0	1120
Foregut_2	15	158.00	948.00
Foregut_6	21	170.00	1020.00
Foregut_8	14	128.00	768.00
Foregut_9	12	67.8	678
Hindgut_1	13	86.60	519.60
Hindgut_2	13	43.20	259.20
Hindgut_6	19	95.60	573.60
Hindgut_8	15	67.20	403.20
Leg_1	21	31.20	187.20
Leg_10	36	25.4	254
Leg_2	31	39.00	234.00
Leg_6	40	62.20	373.20
Leg_8	30	40.60	243.60
Ovary_10	16	200	2000
Ovary_11	17	200.0	2000
Ovary_12	14	200.0	2000
Ovary_9	13	150	1500
SGland_1	12	44.00	264.00
SGland_10	13	33.4	334
SGland_12	17	33.2	332
SGland_2	13	31.20	187.20
SGland_6	20	81.60	489.60
SGland_8	14	46.20	277.20
Stylet_1	18	42.60	255.60
Stylet_2	18	29.40	176.40
Stylet_6	21	40.20	241.20
Stylet_8	16	13.10	78.60
Stylet_9	13	14.4	144

Table 2.1: RNA extraction yielded RNA quantities sufficient for library preparation and sequencing. Organs were pooled during dissection and RNA was extracted then quantified. Organ counts for each pooled sample were noted during dissection and isolation steps. RNA concentration was quantified using a Qubit high sensitivity RNA assay.

RIN scores received from Novogene prior to sequencing ranged from 4.8 to 9.7, with a mean RIN of 7.88 (Table 2.2). 36 strand-specific, 150 bp libraries were generated using Illumina technology, with raw read counts from 33.3 million to 51.6 million (mean = 41.8 million raw reads). With a low error rate of 0.03% for all samples and over 90% of bases having a Phred quality score ≥30 (Q30), the generated libraries seem to be of a high standard.

Sample	RIN	Raw reads	Raw data	Effective (%)	Error (%)	Q20 (%)	Q30 (%)	GC (%)
Antenna_1	6.8	41542028	6231304200	98.13	0.03	97.45	92.96	41.68
Antenna_2	7.9	36513432	5477014800	98.57	0.03	97.55	93.13	39.64
Antenna_6	7.4	39660826	5949123900	97.88	0.03	97.39	92.9	37.64
Antenna_8	8	49073278	7360991700	98.67	0.03	97.57	93.19	40.08
Antenna_11	8	47312014	7096802100	98.27	0.03	96.78	91.49	39.12
Foregut_1	4.8	37625008	5643751200	98.3	0.03	97.41	92.81	37.9
Foregut_2	5.5	45774892	6866233800	98.76	0.03	97.64	93.24	37.55
Foregut_6	7.9	40872292	6130843800	98.88	0.03	97.51	93.04	38.54
Foregut_8	6.7	36562440	5484366000	98.93	0.03	97.29	92.49	38.06
Foregut_9	6	43771768	6565765200	98.54	0.03	96.85	91.52	37.61
Foregut_11	8.4	44206448	6630967200	99.17	0.03	96.83	91.54	41.74
Foregut_12	6.8	42655150	6398272500	98.88	0.03	97.75	93.43	38.35
Hindgut_1	8.1	42875388	6431308200	98.42	0.03	97.63	93.25	38.64
Hindgut_2	7.6	45496650	6824497500	98.75	0.03	97.69	93.33	37.08
Hindgut_6	8.2	37347944	5602191600	98.89	0.03	97.54	93.03	37.46
Hindgut_8	7.7	44563616	6684542400	99.07	0.03	97.65	93.27	37.55
Leg_1	8.9	40122122	6018318300	98.81	0.03	97.54	93.2	43.87
Leg_2	8.5	41062434	6159365100	98.35	0.03	97.42	92.9	41.63
Leg_6	7.8	47083928	7062589200	98.68	0.03	97.48	93.05	40.65
Leg_8	8.2	35192666	5278899900	98.73	0.03	97.57	93.18	39.82
Leg_10	5	51579984	7736997600	98.84	0.03	97.15	92.24	41.29
Ovary_9	9.4	42750686	6412602900	99.16	0.03	97.56	92.99	38.45
Ovary_10	9.6	45342552	6801382800	98.96	0.03	97.88	93.7	38.63
Ovary_11	9.3	44250796	6637619400	98.66	0.03	97.19	92.25	40.19
Ovary_12	9.4	34348332	5152249800	99.04	0.03	97.08	91.98	38.26
SGland_1	9	48300288	7245043200	98.81	0.03	97.6	93.22	37.8
SGland_2	7.9	41447626	6217143900	98.53	0.03	97.4	92.8	37.4
SGland_6	8.2	44581714	6687257100	98.69	0.03	97.41	92.79	37.18
SGland_8	8.5	33800178	5070026700	98.56	0.03	97.28	92.65	37.86
SGland_10	8.5	43061558	6459233700	98.55	0.03	96.64	91.23	37.58
SGland_12	9	33331848	4999777200	98.63	0.03	97.28	92.58	37.1
Stylet_1	8.1	39681500	5952225000	98.15	0.03	97.53	93.05	38.74
Stylet_2	8.1	41233140	6184971000	98.49	0.03	97.43	92.81	37.65
Stylet_6	6.7	38989154	5848373100	98.22	0.03	97.6	93.23	37.91
Stylet_8	9.7	36638982	5495847300	98.14	0.03	97.42	92.97	39.77
Stylet_9	7.9	45035032	6755254800	98.6	0.03	97.46	92.74	36.77

Table 2.2: Samples submitted to Novogene provided high level sequencing results with around 42 million raw reads on average.
RNA samples provided to Novogene for library construction were high quality and did not show evidence of substantial degradation. RNA integrity numbers (RINs) ranged from 4.8 to 9.7, with a mean RIN of 7.88.

Due to the strand-specific, paired nature of the generated libraries, it was necessary to determine the direction of each of the paired reads. As such, the pipeline used to generate read counts from raw reads was performed for each library using different strandness flags (forward, reverse or none) during the HISAT2 and HTSeq-count steps. Using the reverse strandness option for these two steps vastly increased the proportion of used read counts for each library when compared to using the forward strandness option (Figure 2.2). Using reverse strandness also eliminated the ambiguous alignment of reads observed when using the no strandness options. As such, the libraries were determined to be reverse stranded, and read counts using reverse strandness options for the HISAT2 and HTSeq-count steps were utilised during downstream transcriptomic analysis.

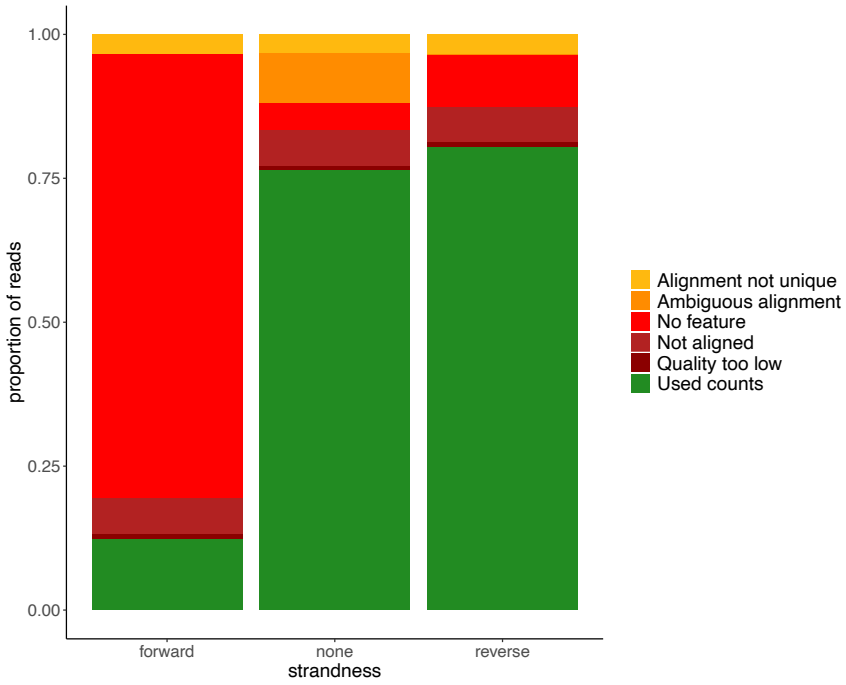


Figure 2.2: Generated organ-specific RNA-seq libraries were determined to be reverse stranded. To determine the strandness of generated libraries, the quantification pipelines were performed using three strandness options in the HISAT2 and HTSeq-count tools. Using reverse strandness options for the respective tools yielded the highest percentage of used counts, and eliminated ambiguous alignments. As such, all further RNA-seq count analyses used negative strandness options for HISAT2 and HTSeq-count steps.

2.3.3: Tissues do not express all genes in the aphid genome, and many genes are not expressed at all.

Genes were determined to be expressed in a given tissue if they had a mean transcripts per million (TPM) value ≥ 0.1. The ovaries expressed the highest number of genes, expressing 16323 of the total 37720 genes in the *M. persicae* genome, while the foregut expressed the lowest number, with only 14313 genes being expressed. Otherwise, the antennae expressed 15254 genes, the hindgut expressed 14995 genes, the legs expressed 14951 genes, the salivary gland expressed 15900 genes, and the stylet expressed 15828 genes. A total of 17402 genes did not have a mean TPM value ≥ 0.1 in any of the seven organs sampled, while 11817 genes were found to be expressed in all seven sampled organs.

2.3.4: Internal M. persicae organs show a high degree of transcriptomic specialisation.

Samples from the foregut, hindgut, and ovaries show a high degree of pairwise correlation between samples within the same organs, suggesting high levels of

transcriptomic specialisation within those tissues (Figure 2.3). Samples from the antennae, legs, and stylet, do not show such a high intra-organ transcriptomic correlation, potentially due to their similar functions or to the presence of shared structures such as musculature and exoskeleton.

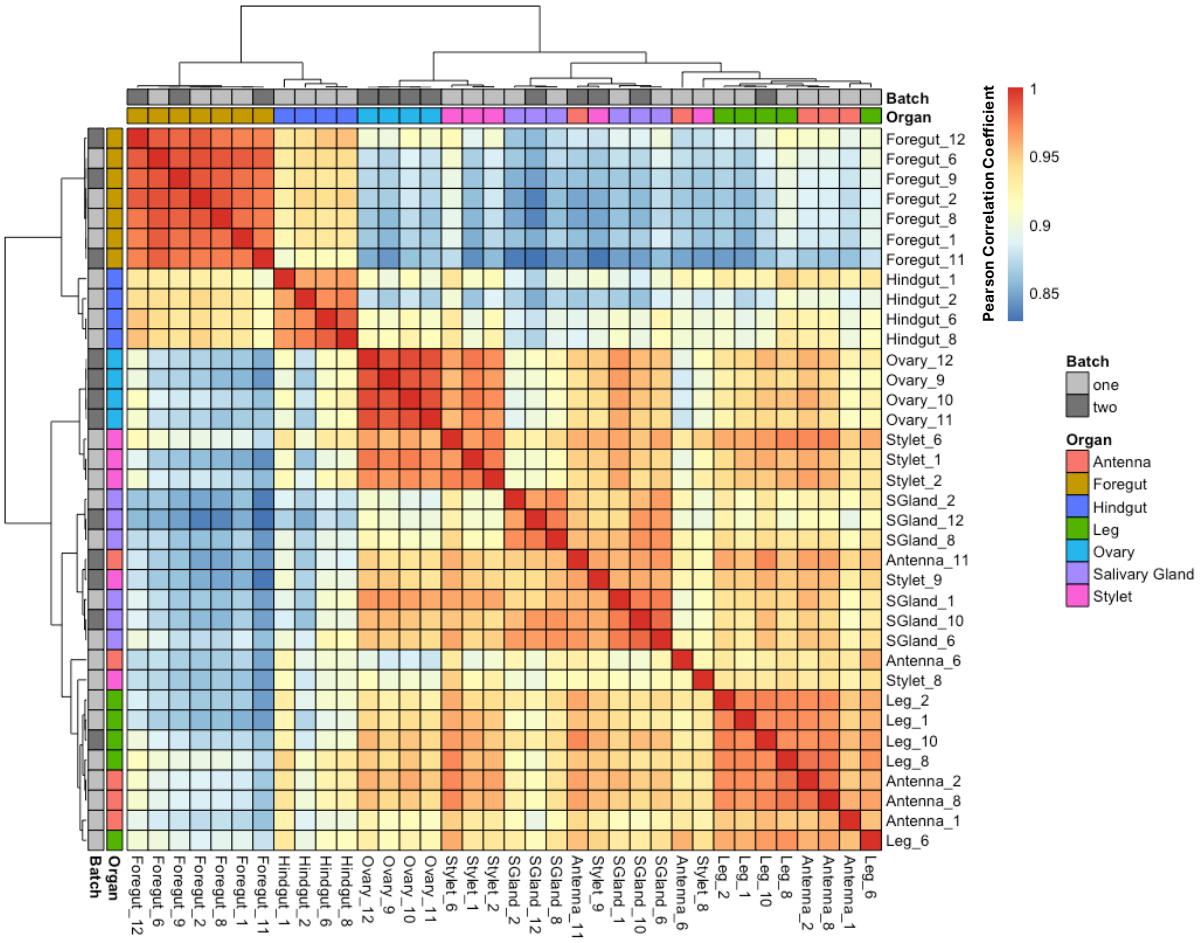


Figure 2.3: Gene expression in foregut and hindgut samples correlate strongly within and between the two organs.
Salivary gland samples also correlate quite closely with each other, but appear to fall into two distinct subclusters. This heatmap also points to a lack of sequencing batch effect, as samples from batch one and two do not appear to segregate. Heatmap showing the pairwise Pearson's correlation coefficient of normalised read counts (TPM) in each RNA-seq library. Heatmap generated using pheatmap R package (Kolde, 2025).

Given that sequencing was performed in two batches (Supplementary Table 1), it was important to rule out technical variability between sequencing batches, and so the sequencing batch of each sample was included as metadata during differential expression analysis. Based on pairwise correlation analysis, there

appears to be no effect of sequencing batch on expression data, with organ samples from both batches clustering well together (Figure 2.3).

Principal component analysis of gene expression for the top 500 most variable genes in the generated libraries showed that principal component 1 (PC1) and principal component 2 (PC2) contributed to 47% and 17% of variation respectively (Figure 2.4). Hindgut and foregut samples cluster together on the PC1 axis separate from the other organs, suggesting that they are the two most transcriptomically distinct organs. 5 of the 6 salivary glands separate from the remaining four organs on the PC2 axis, suggesting a notable, but less pronounced specialisation than observed in the foregut and hindgut. The other salivary gland sample, SGland_1, does not seem to show the same separation from the remaining 4 organs' samples, but exploratory analysis of gene expression across the salivary gland samples indicated that this has not arisen as a result of human error such as misidentification of the tissue or mishandling of the sample, but may be a result of an as yet unknown heterogeneity in expression between salivary gland samples. Similar to the results of the pairwise correlation analysis, samples from across sequencing batched cluster together. Therefore, I am confident that no biases arose as a result of sequencing batch.

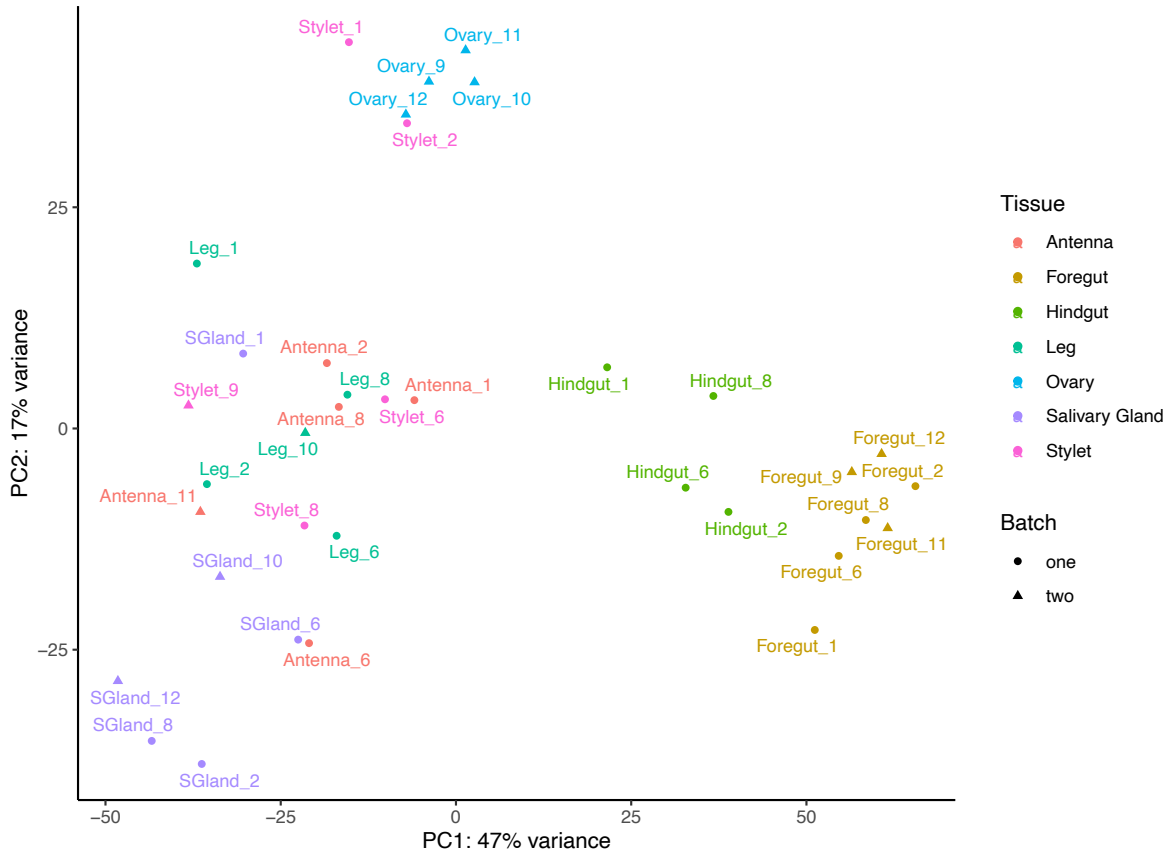


Figure 2.4: Organs within the alimentary tract show highly specialised transcriptomes when compared with other organs.
Principal component analysis reveals that organs such as the salivary glands, foregut, and hindgut, all cluster strongly based on the expression of the 500 genes most strongly driving inter-organ variation.

2.3.5: The aphid foregut contains the highest number of organ-enriched genes.

Reflecting the results of the clustering analysis, the foregut transcriptome shows the highest number of both up-regulated (log_2 fold change ≥ 1, adjusted *p*-value ≤ 0.05) and down-regulated (log_2 fold change ≤ -1, adjusted *p*-value ≤ 0.05) genes when compared to all other organs (n=977 and n=4726 respectively) (Figure 2.5). Hindgut samples contained 643 up-regulated and 1064 down-regulated genes, and salivary gland samples showed up- and down-regulation of 646 and 641 genes respectively. The ovaries also appear to have extensive transcriptomic specialisation, containing 105 up-regulated and 1208 down-regulated genes. The remaining organs show reduced numbers of differentially regulated genes, potentially pointing to their overlapping physiology and function.

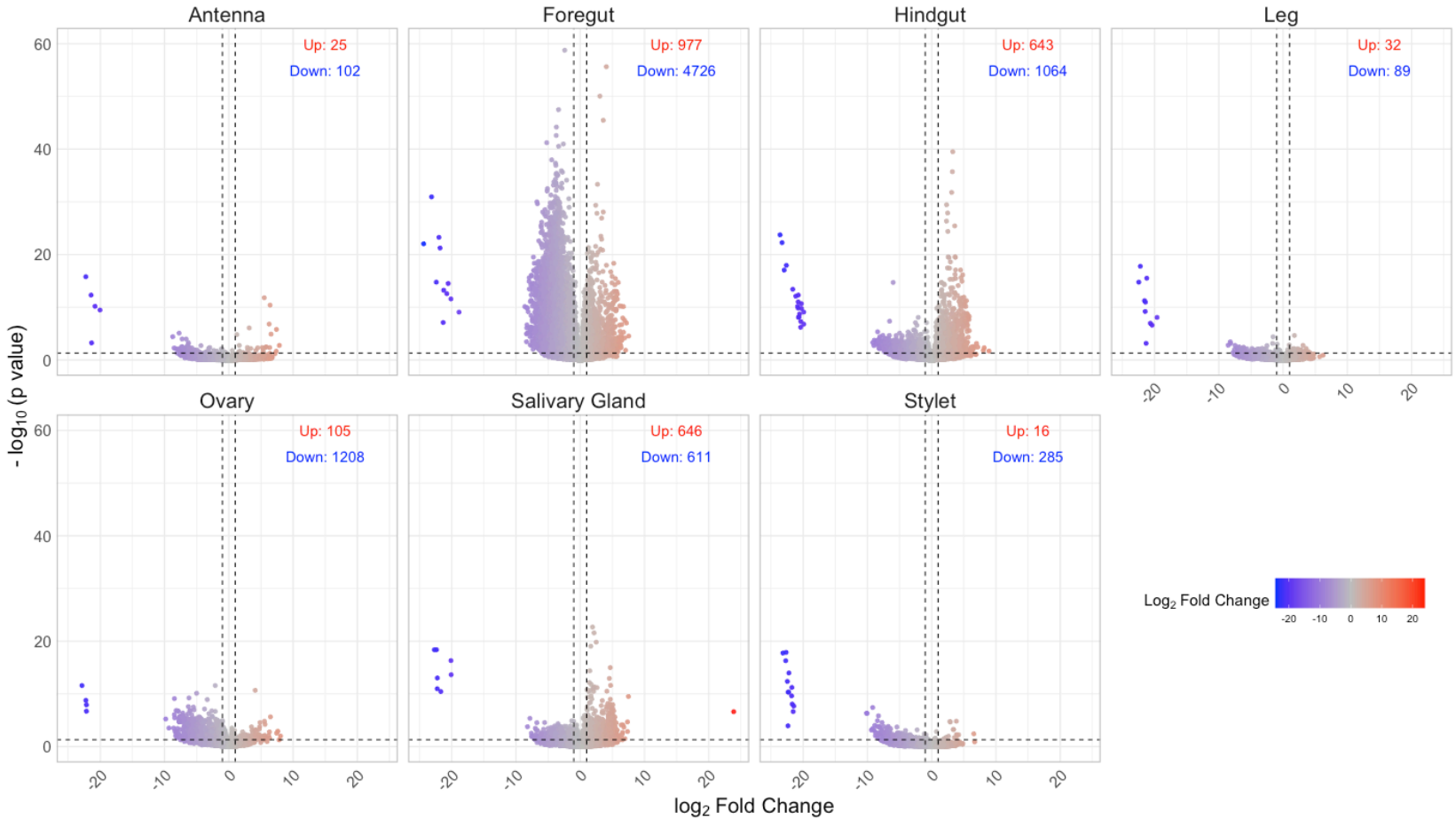


Figure 2.5: The foregut contains the highest number of organ-specific genes.
Volcano plots displaying genes which show up-regulation (log_2 fold change ≥ 2, adjusted *p*-value ≤ 0.05) and down-regulation (log_2 fold change ≤ -2, adjusted *p*-value ≤ 0.05) in each of the seven isolated tissues. The foregut has the highest number of up-regulated genes (n=977), though the salivary gland and hindgut also show a high number of up-regulated genes.

2.3.6: Organ-enriched genes are enriched for function-associated gene ontologies.

Gene ontology (GO) enrichment analysis of the genes identified as being up-regulated in an organ-specific manner revealed a number of enriched GO terms which are reflective of physiological and functional specialisation, with the more transcriptionally distinct organs being enriched for a higher number of GO terms (Figure 2.6). Antenna samples were only enriched for up-regulated genes involved in odorant binding (Figure 2.6A), which points to their function as olfactory organs, while samples from the legs were exclusively enriched for chitin binding (Figure 2.6D). The stylet showed up-regulation of genes which did not appear to be specifically enriched for any GO terms.

The foregut showed specific up-regulation of genes which were enriched for ontologies including regulation of catalytic activity, phospholipid transport, and cysteine-type (endo)peptidase activity (Figure 2.6B). All of these ontologies are in keeping with the digestive function of the foregut. Interestingly, several of the genes with cysteine peptidase activity ontologies belong to the established candidate effector family of *M. persicae* CathB proteins. The salivary glands are enriched for membrane-associated gene ontologies, as well as signal peptidase complex activities, suggesting enrichment of proteins involved in the cleavage of signal peptides from secreted proteins (Figure 2.6F)

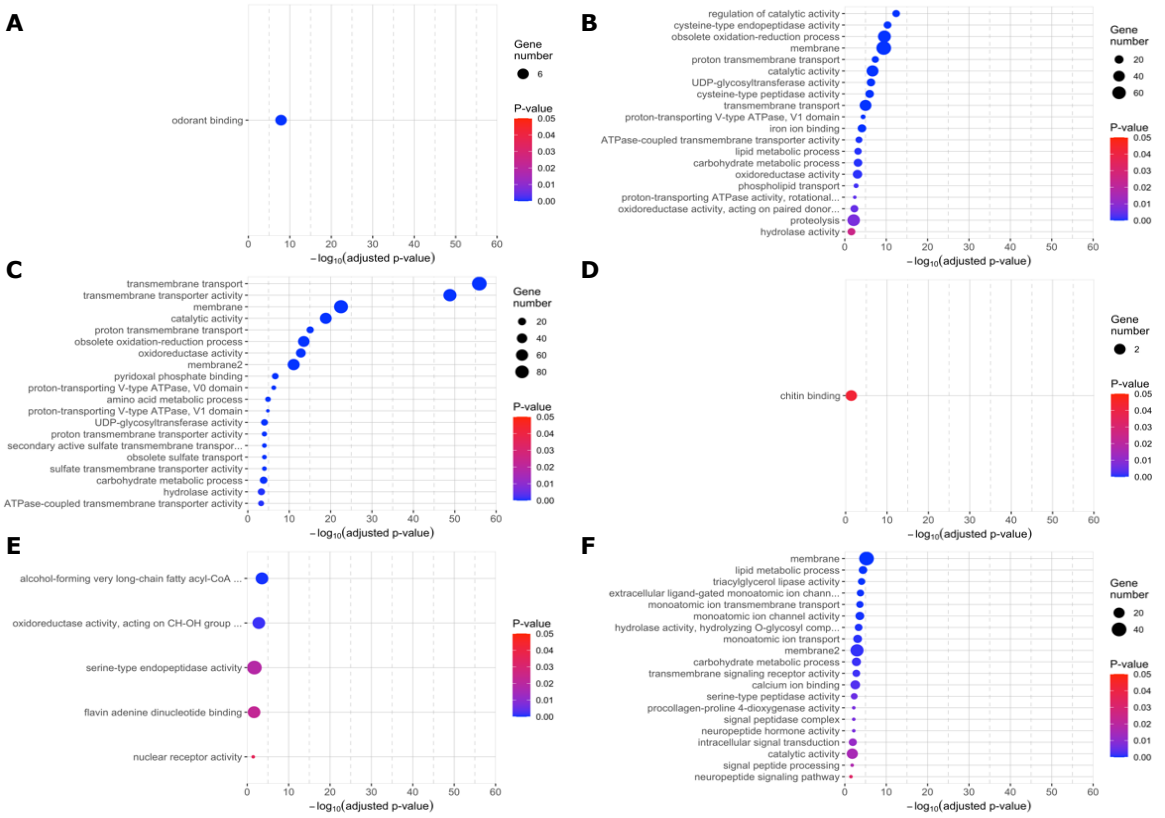


Figure 2.6: Foregut-, hindgut-, and salivary-gland-specific genes were enriched for a high number of gene ontologies.
Notably, the foregut contains organ-specifically up-regulated genes enriched for cysteine type (endo)peptidase activity. **A:** Antenna **B:** Foregut **C:** Hindgut **D:** Legs **E:** Ovaries **F:** Salivary Glands. The stylet-specific genes were not enriched for any gene ontology terms, so were omitted from this figure.

2.3.7: Genes encoding orally secreted proteins are enriched for signal peptides and transmembrane domains.

Peptides derived from transcripts associated with 238 genes were identified in *M. persicae* saliva. Genes encoding 60.08% (143 of 238) of peptides identified in the saliva were predicted to have a signal peptide by SignalP v5. This suggests that genes encoding peptides in saliva are enriched for signal peptides when compared to the whole genome, in which only 12.89% (3831 of 29730) of genes are predicted to code for signal peptides (Chi squared test, χ^2 = 472.3, d.f. = 1, p = 2.2×10^{-16}). Additionally, salivary transcripts showed an enrichment from the whole genome in terms of the abundance of transmembrane domains, with 19.33% (46 of 238) of salivary transcripts being predicted to contain transmembrane domains, compared to 13.51% (4016 of 29730) of transcripts globally (Chi squared test, $\chi2$ = 6.89, d.f. = 1, p = 0.0086).

2.3.8: Genes encoding most, but not all, orally secreted proteins are enriched in the aphid salivary glands.

Of the 238 unique genes encoding peptides identified in the mass spectrometry analysis of *M. persicae* oral secretions (Liu *et al.*, 2024), 118 were identified as being most highly expressed in the salivary glands when compared to all other tissue samples (log_2 fold change ≥ 1, adjusted p ≤ 0.05) (Figure 2.7). Orally secreted genes most highly expressed in the salivary glands were enriched for oxidoreductase (n=4, adjusted *p*-value < 0.0001) and trehalase (n=2, adjusted *p*-value = 0.014) gene ontologies, as well as predicted vitellogenin/lipase (n=3, adjusted *p*-value = 0.0005) and a number of enzymatic domains. 32 genes were most highly expressed in the foregut, 11 in the hindgut, and 2 in the stylets. The remaining 77 genes showed no significant organ-specificity but varying levels of general expression (Figure 2.8).

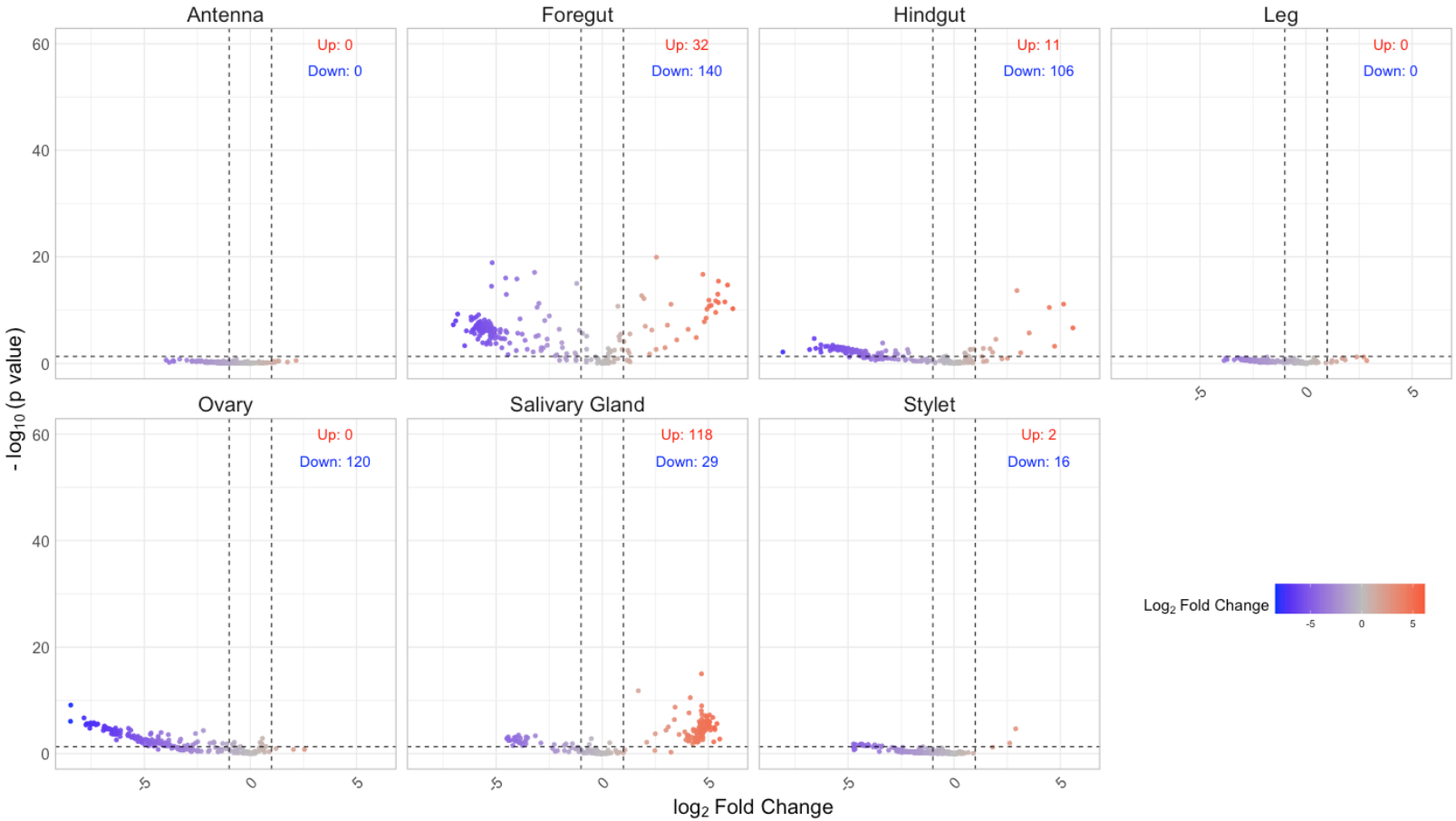


Figure 2.7: Most, but not all, orally secreted proteins are enriched in the salivary glands compared to other tissues.
The salivary gland most highly expresses the highest number of orally secreted proteins (n=118), though there also appears to be proteins encoded most extensively in the foregut (n=32), hindgut (n=11), and stylets (n=2). Volcano plots displaying genes from the OS mass spectrometry analysis which show organ-specific up-regulation (log_2 fold change ≥ 1, adjusted *p*-value ≤ 0.05) and down-regulation (log_2 fold change ≤ -1, adjusted *p*-value ≤ 0.05) in each of the seven isolated tissue

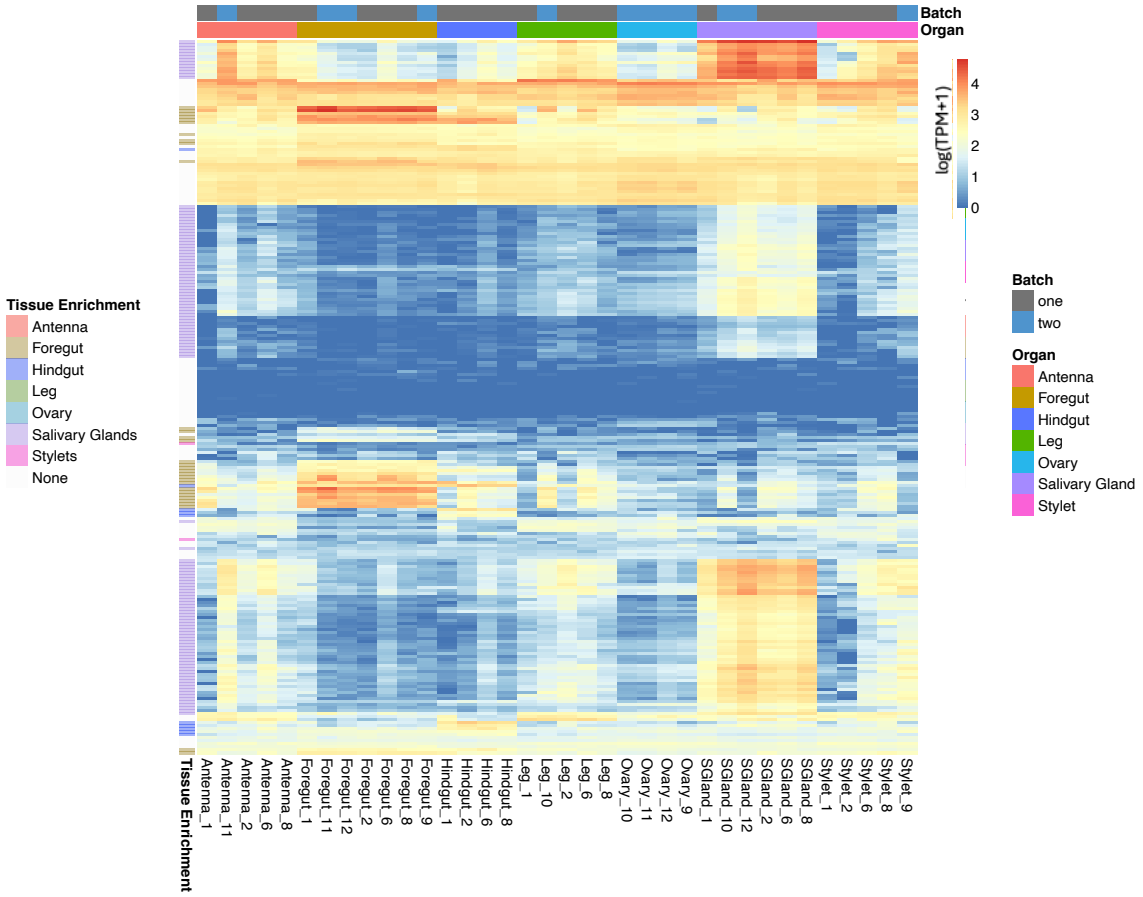


Figure 2.8: Orally secreted proteins are encoded by genes showing a variety of organ-specific enrichments, including the foregut.

Though the majority of genes encoding the 238 peptides identified in the mass spectrometry of aphid saliva are most highly expressed in the salivary glands, there are also a number most highly expressed in the foregut, hindgut, and stylets. There are also several genes which show non-organ-specific expression. Heatmap showing normalised expression of genes associated with OS proteins across the seven organs in the organ-specific RNA-seq dataset. Annotation on the left side of the heatmap shows organ specificity, with genes showing highest expression in one specific organ based on Figure 2.7.

2.3.9: Genes encoding orally secreted proteins are enriched for cysteine proteases including CathB proteins.

Functional enrichment analysis of the entire list of peptides from the oral secretion mass spectrometry revealed enrichment of a number of gene ontology terms, most significant of which were regulation of catalytic activity and cysteine-type endopeptidase activity (Figure 2.9). This list of genes also revealed enrichment of predicted PFAM domains including Histone H2A and Peptidase C1A. Both of these PFAM domains are characteristic of gene families with high degrees of conservation between members, which may have skewed the initial analysis. As such, the functional enrichment analysis was performed on a filtered list of genes which contained only genes encoding "master proteins", single representatives of

otherwise ambiguous peptides, which may be encoded by multiple members of a highly conserved gene family. Using this refined list, the enrichment of Histone H2A domains observed in the original list was eliminated, but the C1A Peptidase domains were still significantly enriched in the "master protein" list, suggesting that this enrichment is legitimate. The C1A peptidase domain is present in CathB proteins, suggesting an enrichment of CathB genes in the mass spectrometry analysis of saliva. Of the 238 salivary-peptide-associated genes, 14 were identified as members of the CathB gene family. In tandem with the enrichment of CathB genes in the foregut, a decision was made to examine the expression of all CathB genes across the *M. persicae* genome.

2.3.10: Orally secreted CathB genes are most highly expressed in the foregut of M. persicae.

To examine the organ-specific expression of CathB genes, I generated standardised counts for CathB genes in each of the samples, in the form of transcripts per million (TPM). Of the 26 CathB genes identified in the *M. persicae* clone O v2.1 genome, 19 showed highest expression in the foregut (log_2 fold change ≥ 1, adjusted $p \leq 0.05$) (Figures 2.7 and 2.8). This includes all 9 of the orally secreted CathB proteins from the recently expanded, *Myzus*-specific clade. 8 of these 9 genes also show an elevated level of foregut expression when compared to other CathBs in the *M. persicae* genome (Figure 2.10). CathB6 exhibits the highest expression of all CathB genes, which mirrors its high abundance in the mass spectrometry of *M. persicae* saliva (Liu *et al*., 2024). Inversely, CathB15, which is not found in the oral secretions, shows no significant expression across any organs.

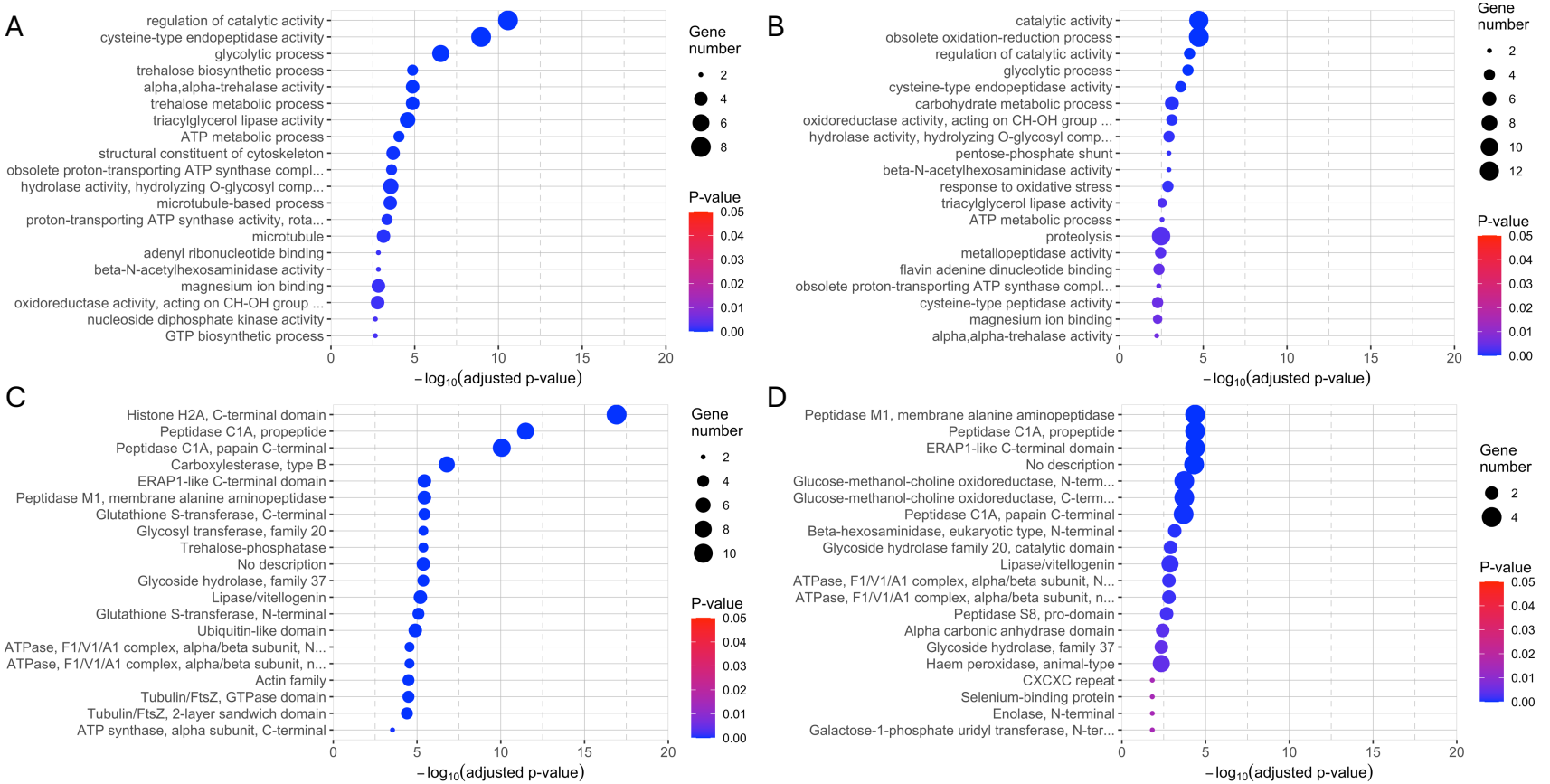


Figure 2.9: The oral secretions of *M. persicae* are enriched for cysteine proteases, including members of the CathB family.
Dot plots showing the top 20 most significantly enriched ontologies (GO) and domains (PFAM) in the oral secretions. In the unfiltered list, the oral secretions appear to be enriched for histone H2A C-terminal domains, but when highly conserved transcripts are removed, as in the master protein list, this enrichment is lost, suggesting that the enrichment of this domain is an artefact of BLAST hits to multiple genes due to high conservation. However, even when using the filtered lists, genes involved in cysteine (endo)peptidase activity are significantly enriched in the oral secretions, as are C1A peptidase domains, which are characteristic of the CathB cysteine protease genes. This matches the pattern of CathB enrichment in the foregut samples of *M. persicae*, suggesting a foregut-derived component of oral secretions. **A:** Enriched gene ontologies of all genes encoding peptides identified in the salivary proteomic analysis. **B:** Enriched gene ontologies of "master proteins" used to account for ambiguous attribution of peptides encoded by highly conserved genes. **C:** Domain enrichment of all genes encoding salivary peptides. **D:** Domain enrichment of "master protein" peptides.

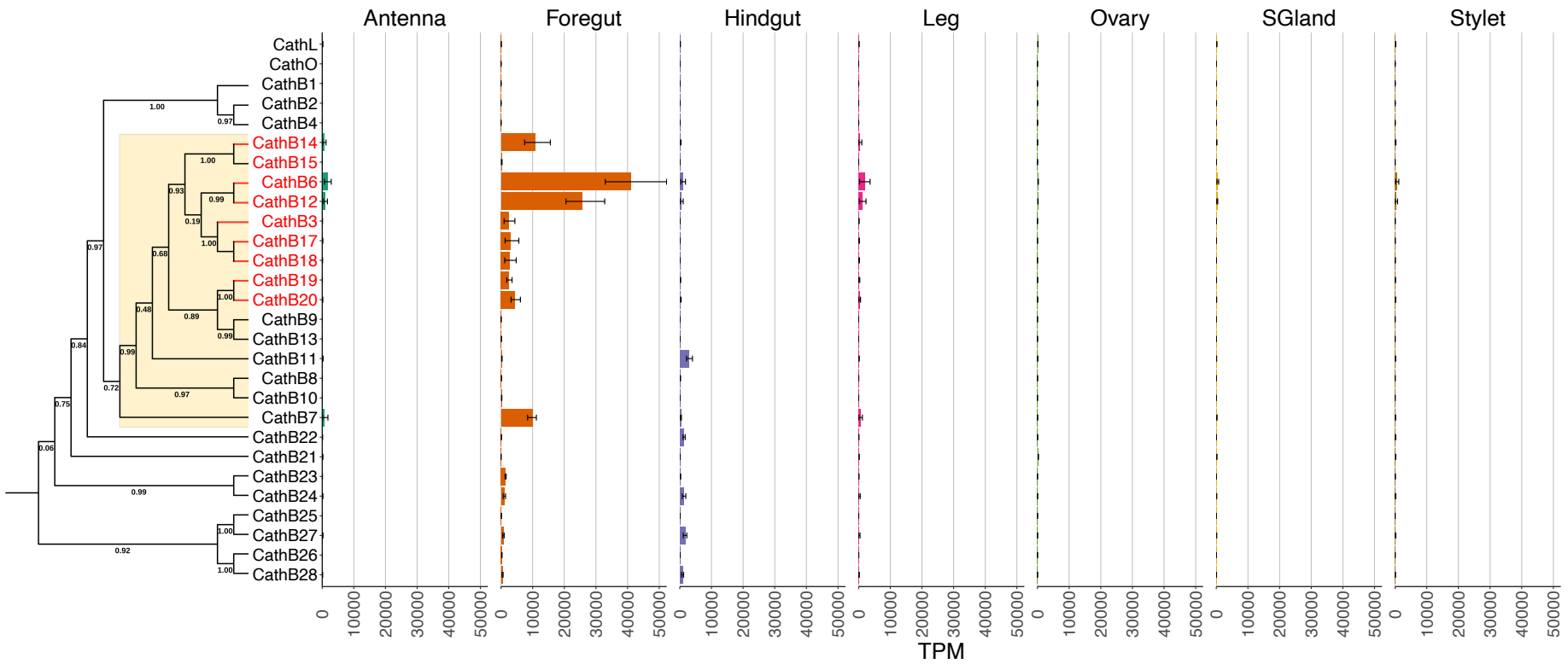


Figure 2.10: CathB genes in oral secretions of *M. persicae* show high levels of foregut specificity.

Bars show mean transcripts per million (TPM) value for each CathB gene across each organ. Orally secreted CathB proteins (Liu *et al.*, 2024) are labelled in red, and an aphid-specific expanded clade is shaded in yellow. Phylogenetic tree of CathB proteins generated by Dr Mar Marzo. Eight of the nine genes encoding peptides identified in the oral secretions showed significantly upregulated in the foregut. CathB7, though not identified in the oral secretions, also showed notably high expression in the foregut. Other members of the CathB family were expressed lowly across all organs, with a small number being expressed in the foregut.

2.4: Discussion

2.4.1: Organ-specific RNA-seq in tandem with OS mass spec reveals foregut-derived proteins in oral secretions.

In keeping with the pre-existing central dogma of aphid effector biology, the majority of proteinaceous components identified in purified aphid saliva correspond to genes most highly expressed in the salivary glands. However, a proportion of orally secreted proteins show significant up-regulation within the foregut, suggesting an alternative mechanism by which aphid effectors may be translocated into plants. This new model would also call into question whether material deposited into plants by aphids during feeding is salivary in nature, and as such may demand a re-evaluation of the language used when discussing these secretions. Accordingly, aphid-derived material secreted into host plants during the course of feeding will be referred to henceforth as "oral secretions", in order to recognise the presence of non-salivary-gland-derived components.

While the exact mechanisms of the translocation of foregut-derived orally secreted components into plants are not fully elucidated, they may follow a similar route of transmission to a number of semi-persistent, foregut-binding plant viruses for which *M. persicae* is a well described vector. Binding to the aphid foregut during the uptake of phloem sap from already infected plants, viruses including *cauliflower mosaic virus* (CaMV), and closteroviruses such as *leaf infectious yellows virus* (LIYV), are deposited into non-infected plants during aphid dispersal and colony establishment, likely during the probing phase of feeding (Figure 2.11).

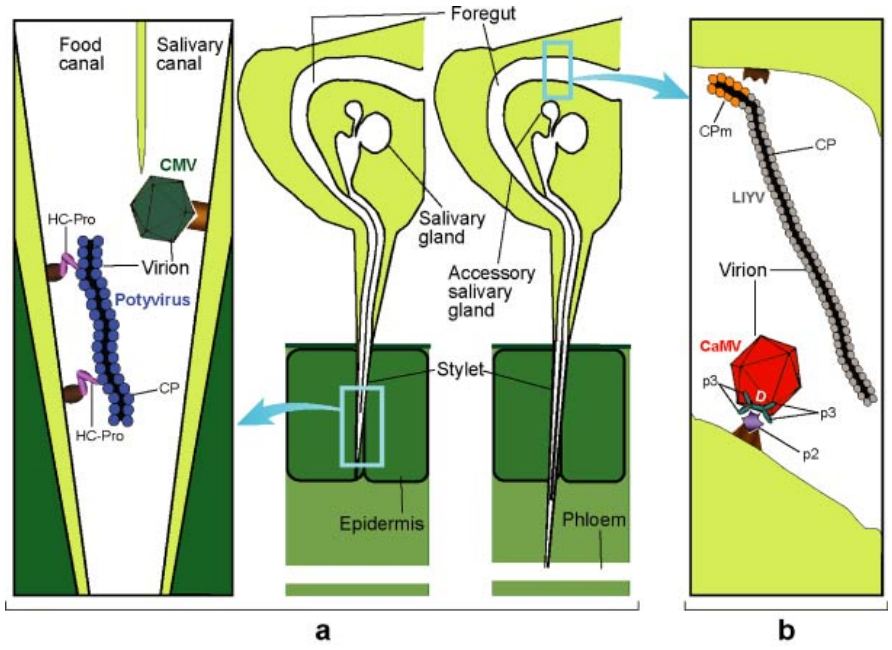


Figure 2.11: CathB proteins may follow a similar route into plants as several foregut-bound viruses. Illustration highlighting a number of aphid-vectored viruses which are translocated into plants during feeding. **A:** CMV and Potyviruses adhere to the stylet. **B:** Cauliflower mosaic virus (CaMV) and other closteroviruses bind to the foregut upon acquisition, and are deposited during feeding. Figure reused with permission from Ng and Falk, 2006.

This suggests deposition of viral particles into plants by aphids through the food channel of the stylet, which may travel in the same matrix as foregut-derived effectors. Further evidence for the deposition of foregut-derived effectors is found in social aphid species *Tuberaphis styraci*, in which a CathB protein is regurgitated through the stylet and injected into enemies, where it functions as a paralytic agent (Kutsukake *et al.*, 2008, Kutsukake *et al.*, 2004).

This injection of foregut contents may be analogous to the regurgitation which can be observed in the larvae of a multitude of lepidopteran species in which oral secretions can be comprised of true saliva, glandular secretions, as well as regurgitated material, all of which are capable of eliciting or modulating the immune response of the host plant (Snoeck *et al.*, 2022). However, feeding mechanisms are divergent between hemipteran and lepidopteran herbivores, with the chewing mouthparts of lepidopteran larvae eliciting much more mechanical damage than hemipterans' piercing mouthparts, and as such the delivery method of host-manipulating oral secretions varies, with lepidopteran oral secretions deposited on the surface of the leaf as opposed to the direct injection style of hemipteran oral secretions. This targeted delivery of effectors into the feeding site

may offer an additional layer of fitness for piercing-sucking insects in the arms race between host and pest.

2.4.2: Internal organs show high levels of transcriptomic specialisation, and are enriched for function-associated genes.

The four organs isolated from inside dissected aphids show a high level of transcriptomic specialisation, reflective of their highly specialised molecular and physiological functions. The extent of this specialisation is not extended to the external organs, which, although likely expressing a number of genes not expressed in the other organs, likely lack significant up-regulation due to their shared physiological structure and overlapping functions.

Additionally, the libraries created from the internal organs of *M. persicae* were enriched for up-regulated genes annotated with a number of gene ontologies which point towards correct identification of organs within the body cavity during dissection. Coupled with the reference images of dissected organs, these libraries can be used in future with confidence that organs were identified and isolated correctly. The reference images (Figure 2.1) will also allow for a direct comparison of libraries in the instance that this dataset is expanded to encompass more of the physiological heterogeneity observed within the *M. persicae* life history strategy. For example, it is possible that the organ-specific transcriptomic regulation may be different between wingless and winged morphs, sexual and asexual female morphs, and females and males, reflective of their varying biological function. With the data provided above, these morphs could be explored in a similarly high resolution to more comprehensively understand the gene regulatory trade-offs which give rise to this phenotypic plasticity.

2.4.3: Advances and future potential.

The generation of this dataset will provide an incredibly useful tool for researchers with interests across *M. persicae* biology. Though generated to cover the spatial transcriptomics of candidate effectors, with specific focus on the foregut, hindgut, and salivary glands, the presence of samples from other organs such as the legs, antennae, and ovaries, broadens the scope of this dataset significantly. Transcriptomes of these tissues will allow for future analysis regarding

reproduction and development, in the case of the ovaries, as well as chemosensory perception and mobility, in the case of the antennae and legs respectively.

While this dataset offers a novel insight into the organ-specific transcription of genes in *M. persicae*, the libraries were only derived from adult, apterous, asexual females. As such, it does not currently represent the sheer complexity of the morphological and reproductive polyphenisms present in *M. persicae*. Therefore, additional datasets replicating the methods described above on alate (winged) females, including libraries derived from the wings, as well as sexually reproducing females and males, would amplify the power of this dataset, and would allow for comparative transcriptomic analysis of the various adult morphs, potentially allowing researchers to untangle the mechanisms by which these specialised phenotypes originate. One specific aspect of interest in such a comprehensive dataset would be the comparative transcriptomic analysis of sexual and asexual ovaries, which differ not only in their morphology, but also in the way which they provide nutrition to the housed progeny. Understanding the variation in gene expression between these two functionally distinct organs would be invaluable in elucidating the polyphenism, and may also uncover the pathways involved in the reproductive mode switch in *M. persicae*.

Chapter 3 - *Myzus persicae* CathB effectors have diversified in their ability to bind *A. thaliana* ACD28.9

3.1: Introduction

Plants, as sessile organisms, are particularly vulnerable to infection and infestation by pathogens and pests. Since these invaders cannot be avoided or overcome behaviourally, plants must rely heavily on their immune system to minimise the damage caused by such challenges. To achieve resistance to pathogens and pests, plants deploy a complex immune response, launched across two phases, and involving a number of intricate metabolic pathways, the purpose of which is to limit mechanical and biological damage caused by pathogens.

The first line of immunity is termed pattern triggered immunity (PTI) and relies on the perception of pathogen-associated molecular patterns (PAMPs), sometimes termed herbivore associated molecular patterns (HAMPs) in the case of infestation by herbivorous insects, which are molecules secreted by pathogens and pests upon invasion. However, well-adapted pests and pathogens can supress the resistive effects of PTI, using specialised effector molecules.

Suppression of pattern triggered immunity by effectors deployed by pathogens and pests presents an additional layer of susceptibility to plants being infected. As such, more hardily resistant plants have evolved a mechanism by which to identify pathogenic effectors, and mount an additional wave of immune response, termed effector triggered immunity (ETI).

In *Myzus persicae*, a number of Cathepsin B (CathB) proteins have been determined to be candidate effectors through a series of genomic and molecular approaches, which have allowed for initial identification and downstream functional analyses. CathB proteins are a family of cysteine proteases, with a structure comprised of three distinct domains (Fig. 3.1). Members of the *M. persicae* CathB gene family were identified as being differentially expressed upon host swap (Mathers *et al.*, 2017), as well as contributing to the establishment of *M. persicae* colonies on *Arabidopsis thaliana*.

Figure 3.1: CathB proteins are comprised of 3 distinct domains.

Proteins in the *M. persicae* CathB family are made up of an N-terminal signal peptide, which allows for the secretion of the protein into aphid oral secretions. The active catalytic form of CathB proteins is contained within the mature domain, which is released from the pro-domain through autocatalytic cleavage under acidic conditions. The identified putative ACD28.9-binding region of CathB6 was identified as lying between amino acids 263 and 333 (Liu *et al.*, 2025).

Additionally, a number of peptides corresponding to the mature domains of several members of the CathB family were identified in a mass spectrometry analysis of *M. persicae* oral secretions (Liu *et al.*, 2024), providing evidence that CathB proteins are efficiently translocated into plants during aphid feeding. By examining the subcellular localisation of CathB6, one of the most abundantly orally secreted CathB proteins, following infiltration into *Nicotiana benthamiana* leaves, it was observed that CathB6 localises to punctate structures ranging from <5 μm^2 to >120 μm^2 in size within the cytoplasm (Liu *et al.*, 2025). These puncta showed dynamic localisation, migrating across the cell and fusing with larger aggregates (Liu *et al.*, 2025). These punctate structures were shown to overlap with expression of processing body (p-body) scaffold Varicose (VCS), suggesting that CathB6 is localising in the p-bodies (Liu *et al.*, 2025).

Given this evidence that CathB6 displays this atypical localisation, a proximity-labelling mass spectrometry (PL-MS) approach was employed to identify potential plant proteins which interact with the deposited CathB protein. With CathB6, one of the most abundant CathB proteins identified in the oral secretions, as the bait, a number of *A. thaliana* proteins were identified as interactors, including Enhanced Disease Susceptibility 1 (EDS1) (Liu *et al.*, 2025), which regulates a number of key processes in plant responses to pathogens, including effector triggered immunity (ETI) and systemic acquired resistance (SAR) (Dongus and Parker, 2021).

In the same study, ACD28.9 was also identified as a plant target of *M. persicae* CathB6 through PL-MS. ACD28.9 is an alpha-crystallin domain (ACD) containing,

28.9kD from *A. thaliana* (Scharf *et al.*, 2001). Confocal microscopy has revealed that the presence of ACD28.9 in leaves co-infiltrated with CathB6 and EDS1 eliminates the formation of puncta by CathB6, instead causing a cytoplasmic distribution of CathB6. This suggests that ACD28.9 is involved in removing the EDS1-bound CathB6 from processing bodies, potentially allowing the previously captured EDS1 to continue its key functions in the regulation of plant immunity. Interestingly, another ACD-containing protein from *A. thaliana*, sieve element-lining chaperone 1, or SLI1, is a 54.2 kD protein which has been characterised as a *M. persicae* resistance factor by genome-wide association mapping of aphid feeding behaviour on 350 *A. thaliana* accessions (Kloth *et al.*, 2017).

Myzus persicae CathB effectors found in oral secretions belong to a recently expanded clade within the CathB protein family (Chen *et al.*, 2020, Liu *et al.*, 2025). Additionally, analyses described in chapter 2 revealed evidence that these proteins correspond to genes most highly expressed in the *M. persicae* foregut. As such, CathB proteins are likely delivered into aphid oral secretions via regurgitation. Given recent evidence that CathB members of the expanded clade have clear virulence functions, and interact with at least two distinct plant proteins (Liu *et al.*, 2025), it is possible that members of the CathB family within the expanded clade have diversified in their interactions with plant proteins, either to increase the specificity of targeting, or to avoid binding to plant proteins involved in defence. It was found that ACD28.9, a plant protein belonging to the small heat shock protein family, a family characterized by the presence of one or more alpha-crystalline domains (ACDs), suppresses the activity of *M. persicae* CathB6 to sequester plant immune proteins (Liu *et al.*, 2025). Therefore, ACD28.9 is effectively increasing plant immunity to aphids via inhibiting CathB6-mediated suppression of host immunity. However, it is not yet clear if other CathB proteins bind ACD28.9, and which amino acids within CathB proteins are involved in binding to ACD28.9.

Work in this chapter addresses the hypothesis of whether the members of the expanded clade of the CathB family, particularly those present in *M. persicae* oral secretions, harbour specific amino acid variations that prevent binding to ACD28.9. It was found that additional *M. persicae* CathB family members bind ACD28.9, but that ACD28.9-binding was not ubiquitous across the family, with

several CathB proteins not binding to ACD28.9. Interestingly, a close paralog of CathB6, CathB12, did not bind to ACD28.9 in yeast, revealing two specific amino acids within the two proteins which appear to be important for binding ACD28.9. This difference has downstream impacts on the *in planta* subcellular dynamics and maintenance of virulence of CathB proteins in response to ACD28.9.

3.2: Materials and methods

3.2.1: Sequence analysis of CathB genes

Full coding sequences of all CathB genes were aligned using MAFFT (Katoh *et al.*, 2002), and alignments were visualised using CLC Genomics Workbench v.24.0.1 (QIAGEN), in which the pairwise comparison tool was used to generate heatmaps of sequence identity. This same analysis was performed on the regions of each CathB gene aligning to the putative ACD28.9-binding region of CathB6, which spans residues 263-333.

Two residue differences between the ACD28.9 binding region of CathB6 and its analogous region on CathB12 were identified, at residues 307 and 314. Mutants were designed containing single residue swaps between CathB6 and CathB12 at these two residues, as well as a double mutant containing both swapped residues. Gene fragments for each of these mutants were synthesised by Integrated DNA Technologies (IDT) and shipped as lyophilised gene fragments. To resuspend these fragments, nuclease-free water was added to each fragment for a final concentration of 100 ng/µl, and vortexed. Resuspended fragments were stored at -20 °C until required for use in the BP reaction with pDONR207.

3.2.2: Gateway cloning of SLI1

The coding sequence of ACD-containing hemipteran resistance protein SLI1 was cloned into yeast-2-hybrid vector plasmids using Gateway cloning technology, in order to perform a yeast-2-hybrid experiment exploring the potential of SLI1 to bind aphid CathB effectors. This was performed using a two-step approach, involving a BP reaction into the donor plasmid pDONR207, and then an LR reaction into destination plasmids pDESTGADT7 and pDESTGBKT7.

The methods below were also used to generate yeast-2-hybrid vectors containing single- and double-residue swap mutants of CathB6 and CathB12 at residues 307 and 314.

3.2.2.1: BP Reaction

A gene fragment for the full coding sequence of SLI1 containing AttB adapters was synthesised by Integrated DNA Technologies. A liquid culture of LB media with 100 µg/mL gentamicin containing empty pDONR207 in *Escherichia coli* DB3.2 taken from lab glycerol stocks was incubated overnight at 37 °C with shaking at 220 rpm, then plasmids were purified from the culture using QIAprep spin miniprep kit. A BP cloning reaction was performed with a mix containing 1 µL (100 ng) gene fragment and 4 µL pDONR207 in TE buffer to a final volume of 8 µL. 2 µL BP clonase was added and the mix was incubated at room temperature overnight. Following the overnight incubation, 1 µL Proteinase K was added, and the mixture was incubated at 37 °C for 10 minutes.

3.2.2.2: Transformation of E. coli

25 µL heat-shock competent DH5α *E. coli* cells were thawed on ice, then 5 µL recombinant plasmid mix was added. Competent cells were incubated on ice for 30 minutes, then heat shocked at 42 °C for 30 seconds. Following heat shock, cells were returned to ice for 2 minutes, then 500 µL S.O.C. medium was added. Cells were incubated at 37 °C for 1 hour, then plated onto LB plates with 100 µg/mL gentamicin. Selective plates were incubated overnight at 37 °C, and resulting colonies were picked and inoculated into 10 mL gentamicin selective liquid LB cultures. Liquid cultures were incubated overnight at 37 °C, then plasmids were purified using QIAprep spin miniprep kit. Cloning efficiency was verified using colony PCR, restriction digest, and whole plasmid Nanopore sequencing (Plasmidsaurus, UK).

3.2.2.3: LR Reaction

1 µL recombinant pDONR207 was added to an LR cloning reaction containing 1 µL empty pDESTGADT7 or 1 µL empty pDESTGBKT7, both from lab glycerol stocks, 6 µL TE buffer, and 2 µL LR Clonase II. This mixture was incubated at room temperature overnight, then the reaction was stopped by adding 1 µL proteinase

K and 10 minutes incubation at 37 °C. Recombinant plasmids were transformed into DH5α competent cells as described above, and cloning efficiency was verified with colony PCR, restriction digest, and whole plasmid Nanopore sequencing (Plasmidsaurus, UK).

3.2.3: Yeast-2-hybrid experiments

3.2.3.1: Preparation of competent cells

A stock of the yeast growth media 2× yeast extract peptone adenine dextrose (YPAD) was generated by the Innes Centre Laboratory Support Media Preparation Team, consisting of 2% w/v yeast extract (Merck, Catalogue no. 1-03753-0500), 4% w/v peptone (Formedium, Catalogue no. PEP02), 4% w/v glucose (Fisher, Catalogue no. G/0500/53), and 0.004% w/v adenine (Sigma, Catalogue no. A2786).

A glycerol stock of *Saccharomyces cerevisiae* strain AH109 competent cells was streaked on to 2× YPAD with 1% agarose, and incubated for 48-72 hours at 28 °C. Following this incubation, 10 mL liquid 2× YPAD was inoculated with colonies from the incubated media, and was incubated overnight at 28 °C with shaking at 220 rpm. 200 mL liquid 2× YPAD was inoculated with a number of volumes of overnight liquid culture, 10-50 µL, and incubated for 14-16 hours at 28 °C with shaking at 220 rpm. The optical density at 600 nM (OD600) of the liquid cultures were measured at 1:10 dilution using a spectrophotometer, with uninoculated 2× YPAD media from the same batch used as a blank. Liquid cultures with OD600 of 0.6-1.0 were used in transformation.

Cultures were centrifuged at 3900 ×G for 10 minutes at room temperature in 50 mL falcon tubes. Supernatant was discarded and the cell pellets were washed by resuspension in 10 mL sterile water, then centrifuged at 3900 ×G for 10 minutes. The resulting supernatant was discarded, and a second wash by resuspension in 1 mL sterile water was performed, before another 10-minute centrifugation. A final wash was performed by resuspending the pelleted yeast in 1 mL 1× TE/LiOAc (10 mM Tris-HCl, 1 mM EDTA, 100 mM LiOAc, pH 7.5), followed by a final

centrifugation step. Prepared cells were resuspended in 250 µL 1× TE / 100 mM LiOAc per falcon tube for a final volume of 1 mL competent cells in solution.

3.2.3.2: Transformation of competent cells

100 µL single-stranded salmon sperm DNA was boiled for 10 minutes, returned to ice, and added to 1 mL competent cells. 20 µL competent cell solution was added to each well of a round-bottomed 96-well plate containing 4 µL (400 ng) purified recombinant pDESTGADT7 and 4 µL (400 ng) purified recombinant pDESTGBKT7 and mixed with repeated pipetting. 100 µL PEG solution (40% PEG4000 / 1× TE / 100 mM LiOAc) was added per well, and the plate was sealed with a PCR sealing foil.

The plate was incubated at 28 °C with shaking at 220 rpm for 60 minutes. Following incubation, cells were heat shocked in the plate by submersion in a water bath at 42 °C for 15 minutes, taking care to avoid flux of water into the wells. The plate was centrifuged at 1800 rpm for 5 minutes, then 100 µL supernatant was removed. 200 µL 1× TE was added to each well without disrupting the cell pellet, then 150 µL was removed.

3.2.3.3: Selecting for integrations and interactions

Synthetic defined medium with glucose (SD+) was produced by John Innes Centre Laboratory Support Media Preparation Team. To generate 1 L of SD+ media, 6.9 g yeast nitrogen base and 20 g glucose were added to 800 mL distilled water, then resulting media was mixed thoroughly and volume made up to 1000 mL. The relevant amino acid dropout mixtures were added to this media according to packaging concentration instructions.

Transformed cells were resuspended in the remaining 78 µL supernatant from the transformation step above, then 4 µL of each transformation was dropped out onto plates containing: 1% agarose SD+ media lacking leucine and tryptophan (SD+ -LT); 1% agarose SD+ media lacking leucine, tryptophan, and histidine (SD+ -LTH); 1% agarose SD+ media lacking leucine, tryptophan, histidine and adenine (SD+ -LTHA). Plates were incubated at 28 °C for 72 hours, then successfully transformed colonies were resuspended in 150 µL sterile water. Each suspension was dropped out onto new plates with the media described above, as

well as two additional SD+ -LTH plates containing 5 mM 3-Amino-1,2,4-triazole (3AT) and 10 mM 3AT respectively. Selection plates were incubated at 28 °C for 72 hours, then imaged.

3.2.4: Gibson cloning

The coding sequence of CathB19, as well as those of the CathB6 and CathB12 mutants, were amplified through PCR using gene specific primers containing adapters with overlaps for the Agrobacterium tumefaciens expression plasmid pB7WFG2.0, which contains a C-terminal green fluorescent protein (GFP) sequence. The template for each PCR was 1 µL recombinant pDESTGBKT7 containing the fragment of interest, which was added to a reaction mix containing 1× Phusion High Fidelity Buffer, 200 µM dNTPs, 0.5 µM each of the respective forward and reverse primers, and 1 U Phusion DNA polymerase, in a final reaction volume of 25 µL. Reaction mixes were added to a PCR Max thermocycler programmed to run the following cycles: polymerase activation at 98 °C for 2 minutes, 35 cycles of (98 °C for 10 seconds, 55 °C for 15 seconds, 72 °C for 1 minute 30 seconds), a final extension step at 72 °C for 10 minutes, followed by a final store step indefinitely at 4 °C. pB7WFG2.0 purified from a lab glycerol stock was linearised through restriction digest using restriction enzymes EcoRI and XHoI. Subsequently, each fragment was assembled with the linearised pB7WFG2.0 backbone using the NEBuilder® HiFi DNA Assembly master mix and protocol.

3.2.5: Transformation of Agrobacterium tumefaciens

A 50 µL aliquot of competent *A. tumefaciens* GV3101 cells was thawed on ice, and 5 µL of purified plasmid, approximately 500 ng, was added, then mixed by repeated pipetting. Each tube was placed in liquid nitrogen for five minutes, then thawed at room temperature. 200 µL of LB was added to the thawed cells, which were then incubated at 28 °C for 3 to 4 hours with shaking at 220 rpm. Cells were plated on LB with appropriate selective antibiotics—10 µg/mL gentamicin, 50 µg/mL rifampicin and 50 µg/mL spectinomycin —and incubated at 28 °C for 4 to 5 days. Cells from successfully grown colonies were used to inoculate 10 mL of LB with the same selective antibiotics. Glycerol stocks were prepared by adding 300 µL overnight culture to 600 µL 40% glycerol, and stored at -80 °C for future use.

Agrobacterium strains transformed with empty pB7WFG2.0-ATG (free GFP), pB7WG2.0 (free mCherry), pK7-VCS-BFP (VCS-BFP) and pB7WG2.0-ACD28.9-mCherry (ACD28.9-mCherry) were synthesised and provided by Dr Qun Liu (Liu *et al.*, 2025, Table S5) for use in confocal microscopy of transient expression analyses.

3.2.6: Agroinfiltration

3.2.6.1: Nicotiana benthamiana *growth conditions*

Wild-type *N. benthamiana* plants were grown in 9cm plastic pots, in Levington starter medium at 22 °C with 80% humidity, under a long-day photoperiod of 16 hours light and 8 hours darkness.

3.2.6.2: Preparation of A. tumefaciens and infiltration

A. tumefaciens GV3101 cells containing the relevant recombinant plasmids were scraped from frozen glycerol stocks, and plated onto LB agar plates with appropriate selective antibiotics—10 µg/mL gentamicin, 50 µg/mL rifampicin and 50 µg/mL spectinomycin for *A. tumefaciens* expressing fluorophore-tagged constructs, and 10 µg/mL gentamicin, 50 µg/mL rifampicin and 50 µg/mL kanamycin for *A.* tumefaciens expressing silencing suppressor gene P19—and incubated at 28 °C for 48 hours. Colonies were picked from the incubated plates and inoculated into 10 mL of liquid LB with appropriate antibiotics, then incubated overnight at 28 °C with shaking at 220 rpm.

The 10 mL overnight cultures were transferred to 15 mL Falcon tubes and centrifuged at 3900 ×G for 10 minutes. The supernatant was discarded and the remaining pellet was resuspended by repeated pipetting in 1 mL MMA infiltration buffer (10 mM MES pH 5.6, 10 mM $MgCl_2$, 0.1 mM acetosyringone). The suspension was centrifuged again at 3900 ×G for 5 minutes, the supernatant was discarded, and the pellet was resuspended in 2 mL MMA infiltration buffer.

Agrobacterium content was quantified by measuring the OD600 of 1 mL of 1:10 dilution of each suspension in water, with a 1:10 dilution of MMA infiltration buffer in water used as a blank, using a spectrophotometer. Infiltration mixes were generated by adding the necessary volumes of each Agrobacterium suspension to

result in a 0.3 OD of *A. tumefaciens* expressing each construct. This included *A. tumefaciens* expressing silencing suppressor gene P19 at an OD of 0.3. Infiltration mixtures were made up to a final volume of 2 mL using sterile water. These mixtures were incubated at room temperature for 1 to 2 hours. Infiltration solutions were used to infiltrate the underside of the leaves of 4-week old *N. benthamiana* plants using a 1 mL needle-less syringe. Infiltrated plants were returned to their standard growth conditions (see 3.2.6.1).

3.2.7: Confocal microscopy

48 hours following infiltration with *A. tumefaciens*, a Leica TCS SP8X upright confocal laser scanning microscope was used to track *in planta* expression of fluorophore-tagged proteins. Raised stages were created on microscope slides using high-vacuum grease to create a square in which distilled water could be added, in order to allow samples to be submerged without compromising the structural integrity of the tissue. Square tissue samples were excised from regions infiltrated with agrobacteria, and added to the stage on the microscope slide. A 22 mm x 26mm glass cover slip (0.5 mm thickness), was placed over the stage, and prepared samples were visualised and imaged using a 20×/0.75 mixed immersion objective. Sequential, unidirectional scans were carried out using the following laser lines: 405 nm (405 nm diode laser), 488 nm (65 mW Argon ion laser), and 580 nm (pulsed SuperK EXTREME supercontinuum white light laser, 470-670 nm, 1.5 mW per line) to excite BFP, GFP, and mCherry respectively. Fluorescence emissions were collected at 420-450 nm, 490-544 nm, and 598-618 nm. Images were acquired on hybrid detectors at laser power < 5%, with line averaging 2 and a pinhole of 1 Airy unit. Images were generated with varying gains and z-steps. The pixel size was set 0.28 µm × 0.28 µm and pixel dwell time 600 ns. Z stacks were produced and scale bars were added in ImageJ (Schindelin *et al.*, 2012).

3.3: Results

3.3.1: CathB6 belongs to a clade of CathB proteins which are recently expanded and show high sequence identities.

As a first step to assess the extent of sequence differences among *M. persicae* CathB family proteins, and particularly of those within the expanded clade, I

performed a pairwise sequence identity analysis among the CathB family members. As expected, the CathB members within the expanded clade showed the highest levels of sequence identities (top-right, Figure 3.2) and the lowest number of amino acid differences (bottom-left, Figure 3.2). Strikingly, CathB6, which is present in *M. persicae* oral secretions along with other CathB proteins (Figure 3.2) and has been investigated extensively as an aphid effector (Liu *et al.*, 2025), has 98% identity across the whole amino acid sequence compared to CathB12. CathB6 and CathB12 differ only by 6 amino acids, corresponding to positions 17, 34, 82, 100, 307 and 314 of CathB6. Four of these amino acids lie in the virulent mature domain of CathB6 (residues 61-338), and of these, two are positioned in the putative ACD28.9-binding region (Figure 3.1).

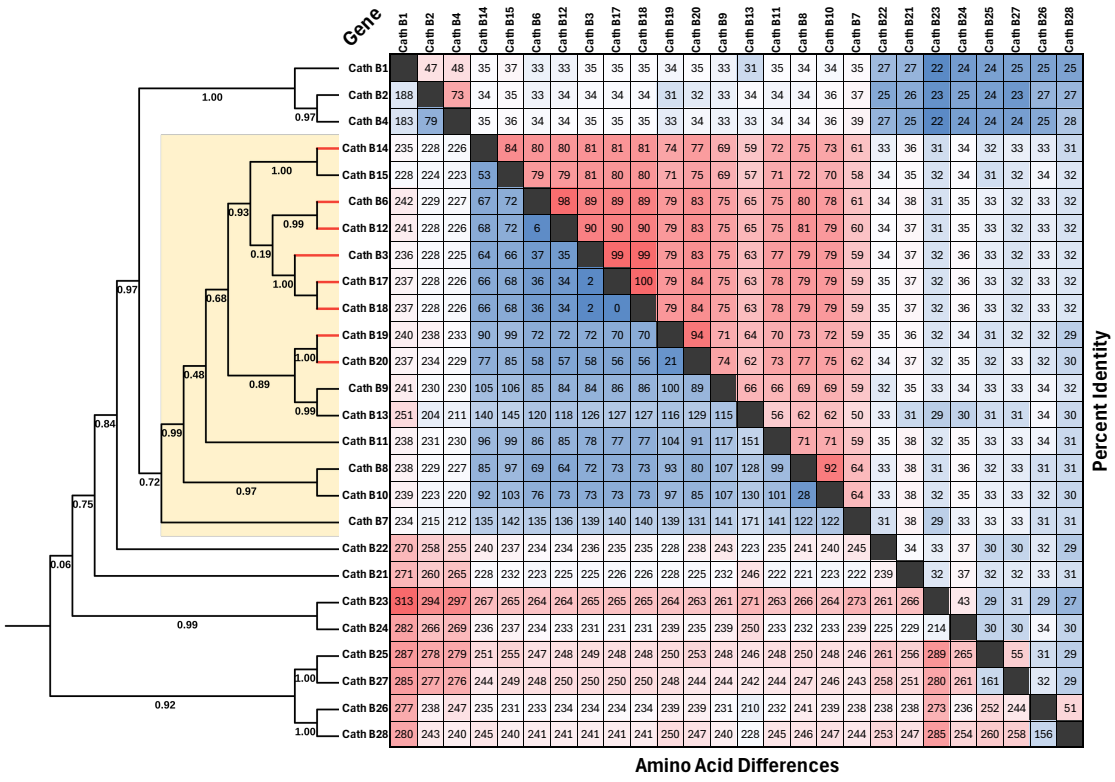


Figure 3.2: Orally secreted CathB proteins belong to a clade with elevated levels of pairwise percentage identity.
Top-right: high sequence identities are shaded red and low identities are shaded blue; Bottom-left: fewer amino acid differences are shaded blue and more differences are shaded red. CathB proteins present in M. persicae oral secretions (Liu *et al.*, 2025) are highlighted by red branches on the phylogenetic tree constructed by Dr Mar Marzo (left) and belong to the recently expanded clade which is labelled in yellow (Chen *et al.*, 2020).

In coding regions aligning to the 70 amino acid region of CathB6 involved in ACD28.9 binding, amino acids 263-333, Cathepsins B6 and B12 only differ in 2 positions (Figure 3.3B). The remaining 4 divergent amino acids between these two

protein sequences lie outside of the putative ACD28.9-binding region of CathB6 (Figure 3.3A). The homologous region of CathB3, B17, and B18 are identical in the putative binding regions.

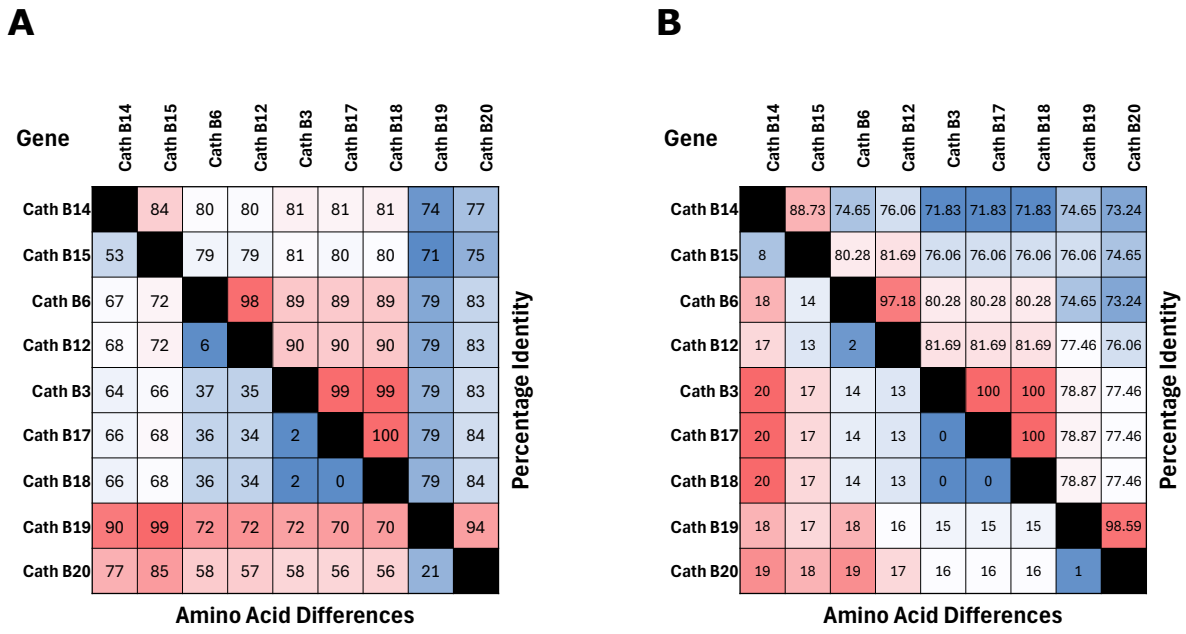
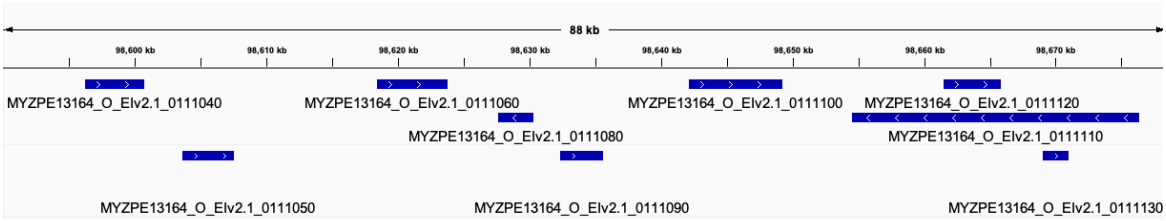


Figure 3.3: Pairwise comparisons of orally secreted members of the Myzus persicae CathB protein family.
Top-right of each panel: high sequence identities are shaded red and low identities are shaded blue; Bottom-left of each panel: fewer amino acid differences are shaded blue and more differences are shaded red. **A:** Pairwise comparisons showing amino acid differences across the full coding sequence across each gene **B:** Pairwise comparisons showing amino acid differences across regions aligning to amino acids 263-333 of CathB6, the region of CathB6 implicated in the binding of CathB6 and ACD28.9.

Intriguingly, the 9 genes corresponding to the CathB found in oral secretions are arranged in close proximity over a span of approximately 80 kilobases the *M. persicae* genome assembly scaffold 1, which corresponds to the X chromosome (Figure 3.4A). A separate cluster of CathB genes that do not belong to the expanded clade lie on a different location corresponding to *M. persicae* genome assembly 4, which is one of the 4 autosomes (Figure 3.4B).

A

scaffold_1:98,589,929-98,678,344



B

scaffold_4:47,748,876-47,837,291

C

Gene Name	Gene model ID	Scaffold	Start	End	DNA strand	Figure
CathB8	MYZPE13164_O_Elv2.1_0111040	scaffold_1	98596245	98600707	leading	A
CathB10	MYZPE13164_O_Elv2.1_0111050	scaffold_1	98603662	98607588	leading	A
CathB11	MYZPE13164_O_Elv2.1_0111060	scaffold_1	98618461	98623795	leading	A
CathB12	MYZPE13164_O_Elv2.1_0111080	scaffold_1	98627677	98630301	lagging	A
CathB6	MYZPE13164_O_Elv2.1_0111090	scaffold_1	98632373	98635613	leading	A
CathB13	MYZPE13164_O_Elv2.1_0111100	scaffold_1	98642170	98649292	leading	A
CathB9	MYZPE13164_O_Elv2.1_0111110	scaffold_1	98654523	98676417	lagging	A
CathB14	MYZPE13164_O_Elv2.1_0111120	scaffold_1	98661548	98665846	leading	A
CathB15	MYZPE13164_O_Elv2.1_0111130	scaffold_1	98669090	98671039	leading	A
CathB25	MYZPE13164_O_Elv2.1_0302340	scaffold_4	47767108	47772877	lagging	B
CathB26	MYZPE13164_O_Elv2.1_0302350	scaffold_4	47774218	47779877	lagging	B
CathB27	MYZPE13164_O_Elv2.1_0302390	scaffold_4	47809921	47813048	lagging	B
CathB28	MYZPE13164_O_Elv2.1_0302400	scaffold_4	47816062	47819501	lagging	B

Figure 3.4: *Myzus persicae* CathB genes are organized as tandem repeats within at least two locations in the genome assembly.
A: Genes of CathB proteins in *M. persicae* oral secretions lie within a region of ±80 kb of the *M. persicae* genome assembly scaffold 1, which correspond to the X chromosome. Schematic representation shown was generated with Integrative Genomics Viewer (IGV) (Robinson *et al.*, 2011). **B:** Genes encoding several non-secreted CathB proteins are arranged across a 55 kb region of scaffold. **C:** Additional information about CathB genes displayed in panels A and B. Figure created with the assistance of Dr Mar Marzo, Hogenhout Lab.

3.3.2: Proteins across the CathB family are capable of binding to ACD28.9 in yeast.

To determine whether the interaction between ACD28.9 and CathB6 was specific, I performed a yeast-2-hybrid binding assay of ACD28.9 with additional members of the CathB family. Interactions between CathB family proteins and broad-spectrum herbivore resistance factor SLI1 (Kloth *et al.*, 2017, Kloth and Kormelink, 2020, Kloth *et al.*, 2021) was also tested, to assess whether the

conferred resistance arises as a result of the neutralisation of CathB effectors by SLI1.

Successful double transformation was confirmed for all transformants by growth of yeast on the double drop-off media of leucine and tryptophane (-LT)(Figure 3.5, left panel).

All of the tested CathB proteins were capable of binding to the positive control, CathB6, on the quadruple drop-off media of leucine, tryptophane, histidine and adenine (-LTHA), giving evidence for successful production and folding of CathB proteins within the yeast. At this stringency, 18 of the 19 tested CathB proteins showed evidence of binding to ACD28.9 (Figure 3.5, middle panel).

Using the highest stringency media, SD+ media lacking leucine, tryptophane, and histidine (-LTH), and supplemented with 3-amino-1,2,4-triazole (3AT) at a concentration of 5 mM, 10 of the 19 successfully transformed CathB proteins were able to restore yeast growth when co-transformed with ACD28.9. Notably, CathB3 and CathB6 bind ACD28.9, whereas CathB9 does not (Figure 3.5, right panel). These findings are consistent with published data (Liu *et al.*, 2025). ACD28.9-binding CathB proteins were CathB3, CathB6, CathB7, CathB11, CathB14, CathB17, CathB18, CathB19, CathB20, and CathB21.

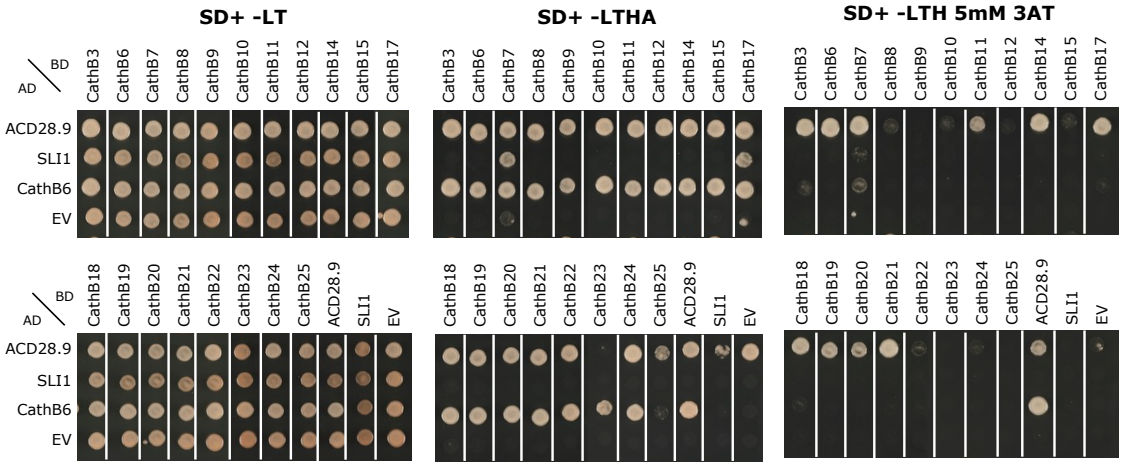


Figure 3.5: Proteins across the CathB family show differences in their ability to bind ACD28.9.
SD+ -LT plates showed successful integration of both activation and binding domain-containing plasmids. SD+ -LTHA plates showed dimerisation of CathB proteins with CathB6, acting as a positive control for the production of CathB proteins in the yeast. Highest stringency media (SD+ -LTH 5mM 3AT), showed differential ACD28.9 binding capacity within the CathB family.

Of the CathB proteins that were detected in the *M. persicae* oral secretions, and may therefore be subject to perception by to ACD28.9, most, but not all, interacted with ACD28.9 in the Y2H assay (Figure 3.5). CathB17 and CathB18, two of the proteins closest to CathB3 and B6, also interacted with ACD28.9, whereas CathB12 did not. Similarly, two further closely related proteins show a difference in ACD28.9 binding, with CathB14 binding ACD28.9 in yeast, and CathB15 not binding. As such, yeast-2-hybrid results show that ACD28.9 binding is not ubiquitous across the CathB protein family (Figure 3.5).

On -LTHA media, yeast co-transformed with SLI1 and two CathB proteins, CathB7 and CathB17, showed a positive result for interaction. However, both of these proteins also show a degree of autoactivation in the empty vector control, and so the interaction between SLI1 and these proteins cannot be confirmed (Figure 3.5). Additionally, this growth is not observed on the highest stringency media (-LTH 5mM 3AT), suggesting that the growth on the lower stringency media is likely a result of autoactivation.

3.3.3: An evolutionarily distant CathB, CathB19, shows similar subcellular localisation dynamics to CathB6 in the presence of ACD28.9.

Based on the interesting subcellular dynamics of CathB6 *in planta* during transient expression, a distantly related, similarly ACD28.9-binding CathB was selected to determine whether this interaction between CathB6 and ACD28.9 was specific. CathB19 has been shown to bind ACD28.9 in yeast (Figure 3.5), so I performed confocal microscopy analysis of *Nicotiana benthamiana* leaves transiently expressing CathB19.

CathB19 was cloned into green fluorescent protein (GFP) labelled *Agrobacterium tumefaciens* plasmid pB7WG2F using Gibson cloning technology. Transformed *A. tumefaciens* expressing the C-terminal GFP-tagged CathB19 was infiltrated into leaves of *N. benthamiana* alongside BFP-tagged p-body marker VCS to determine whether potential puncta corresponded to p-bodies. CathB9-GFP was also co-expressed with either mCherry-tagged ACD28.9 or free mCherry to explore the subcellular localisation of the CathB proteins in the presence and absence of ACD28.9.

CathB19-GFP forms punctate structures of varying sizes in *N. benthamiana* cells. These puncta also show expression of VCS-BFP, suggesting that they correspond to p-bodies (Figure 3.6A). This is the same localiation as is displayed by CathB6 *in planta*. However, similar to CathB6, when GFP-tagged CathB19 was co-expressed with mCherry-tagged ACD28.9 in *N. benthamiana* leaves, this recruitment to the p-bodies appeared to be diminished, with CathB19 being at least partially rescued from p-bodies and showing a more cytoplasmic distribution (Figure 3.6B). As such, it seems that the phenotype of CathB removal from p-bodies by ACD28.9 is not specific to CathB6 alone, but may extend to any CathB proteins which can be strongly bound to ACD28.9.

3.3.4: CathB6 and CathB12 interact differently with ACD28.9 in yeast and in planta.

Based on the high sequence similarity between CathB6 and CathB12, a yeast-2-hybrid assay was performed to examine their potential for interaction with ACD28.9, as well as with SLI1, the previously described ACD-containing protein from *A. thaliana* previously identified as a resistance factor in response to hemipteran infestation. While CathB6 and CathB12 both appear to be capable of binding to ACD28.9, the maintenance of growth of yeast with CathB6 and ACD28.9 integrations at lower yeast concentrations and higher media stringencies suggests that the interaction between CathB6 and ACD28.9 is markedly stronger than that between CathB12 and ACD28.9 (Figure 3.7). SLI1, however, does not appear to bind to CathB6 or CathB12 at any concentration or stringency, suggesting that SLI1 does not bind CathB6 or CathB12 at all.

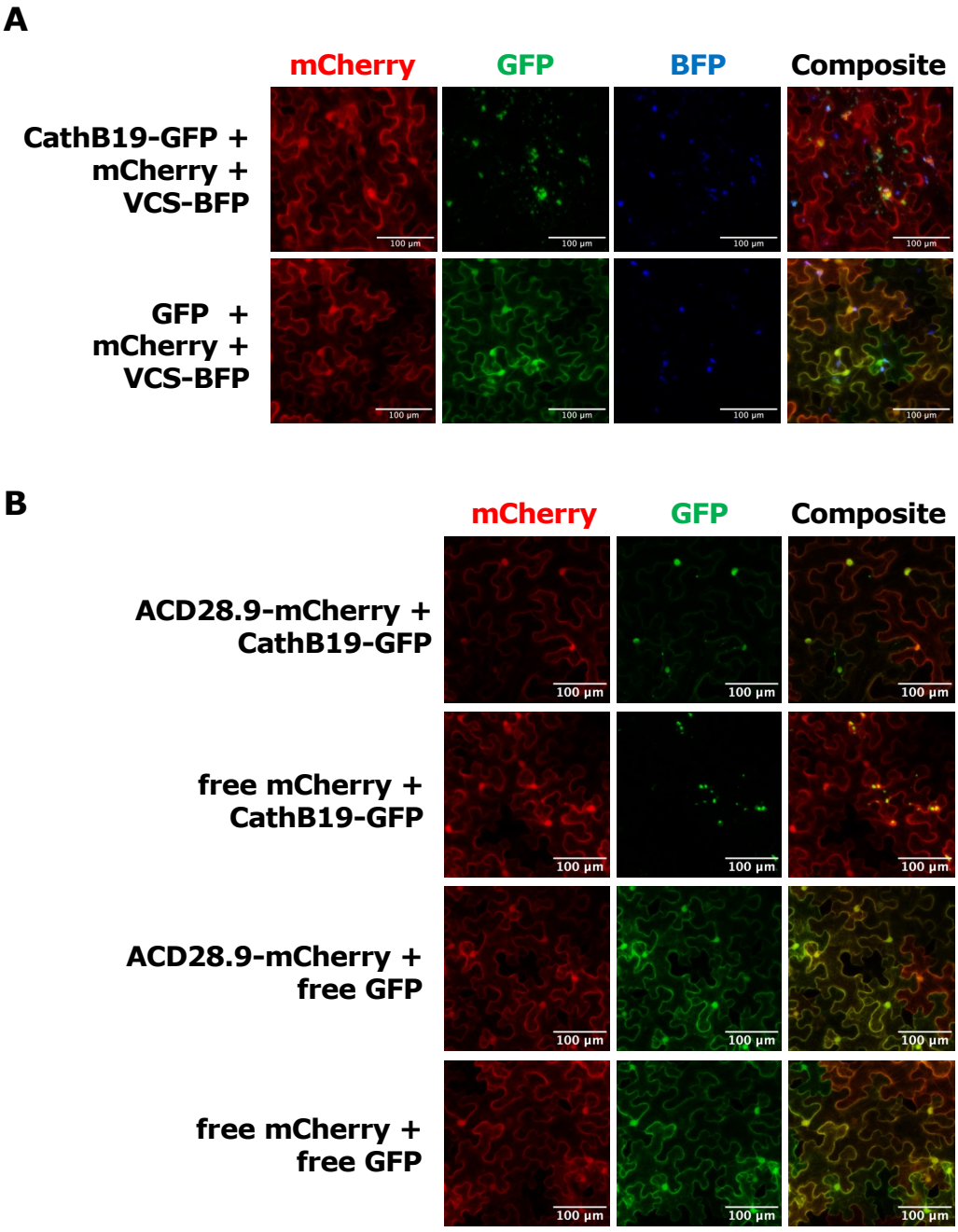


Figure 3.6: CathB19 is removed from p-bodies by ACD28.9 to a degree similar to CathB6.
Confocal microscopy of *N. benthamiana* leaves during transient expression of CathB19-GFP, ACD28.9-mCherry and VCS-BFP. **A:** CathB19 displays p-body localisation characteristic of all currently tested CathB proteins *in planta,* with CathB19-GFP colocalising with varicose-BFP. **B:** When co-expressed with ACD28.9-mCherry, this p-body localisation is diminished, with CathB19 also being observable in the cytoplasm and nucleus of cells. This is similar to the localisation dynamics of CathB6.

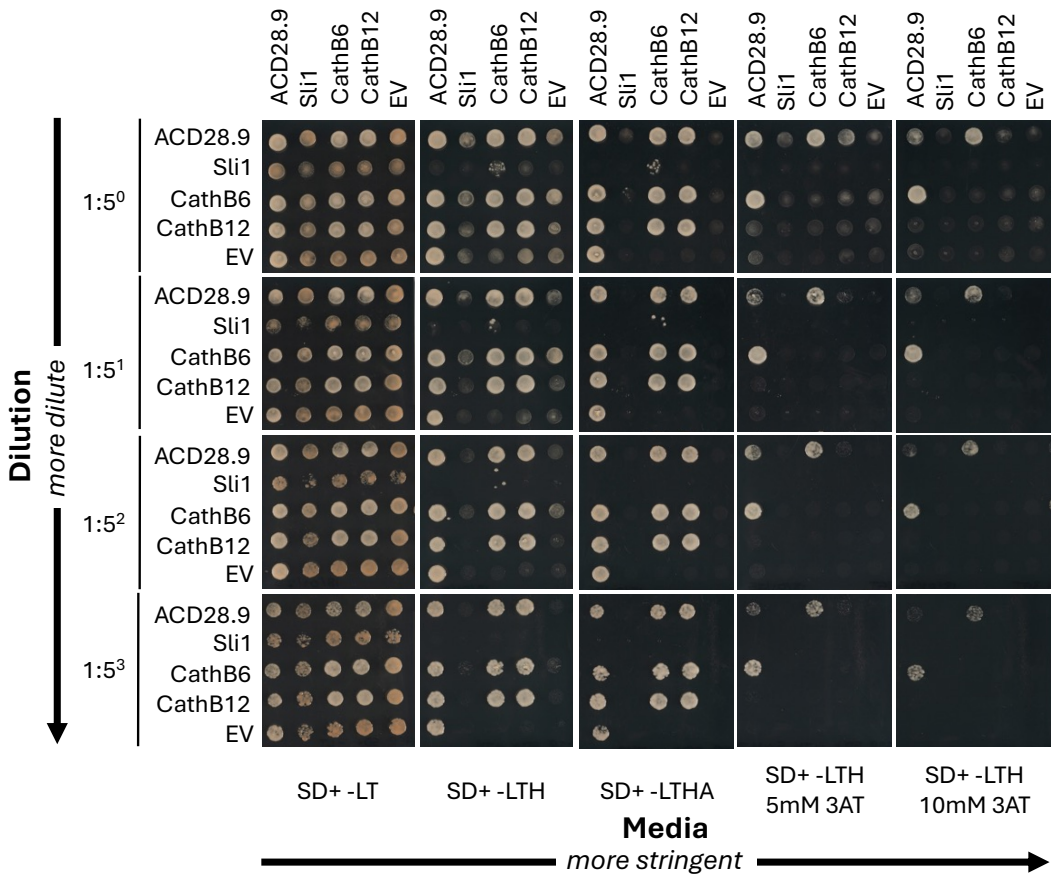


Figure 3.7: CathB6 and CathB12 show differences in binding affinity to ACD28.9.
Yeast-2-hybrid results show that while both CathB6 and CathB12 are both capable of binding ACD28.9, only the interaction between CathB6 and ACD28.9 is maintained at lower yeast concentrations and higher levels of stringency, suggesting a higher binding efficiency between CathB6 and ACD28.9. These results also imply that neither CathB6 nor CathB12 bind the ACD protein SLI1.

In planta, CathB12 forms puncta in a similar way to CathB6 and CathB19. These puncta also appear to represent p-bodies, due to the colocalization of CathB12 and p-body marker varicose (Figure 3.8A). However, during co-expression of CathB12 and ACD28.9, the recruitment of CathB12 to p-bodies is maintained (Figure 3.8B). This marks a stark difference from the dynamics of CathB6, which loses p-body localisation when co-expressed with ACD28.9 (Figure 3.8B).

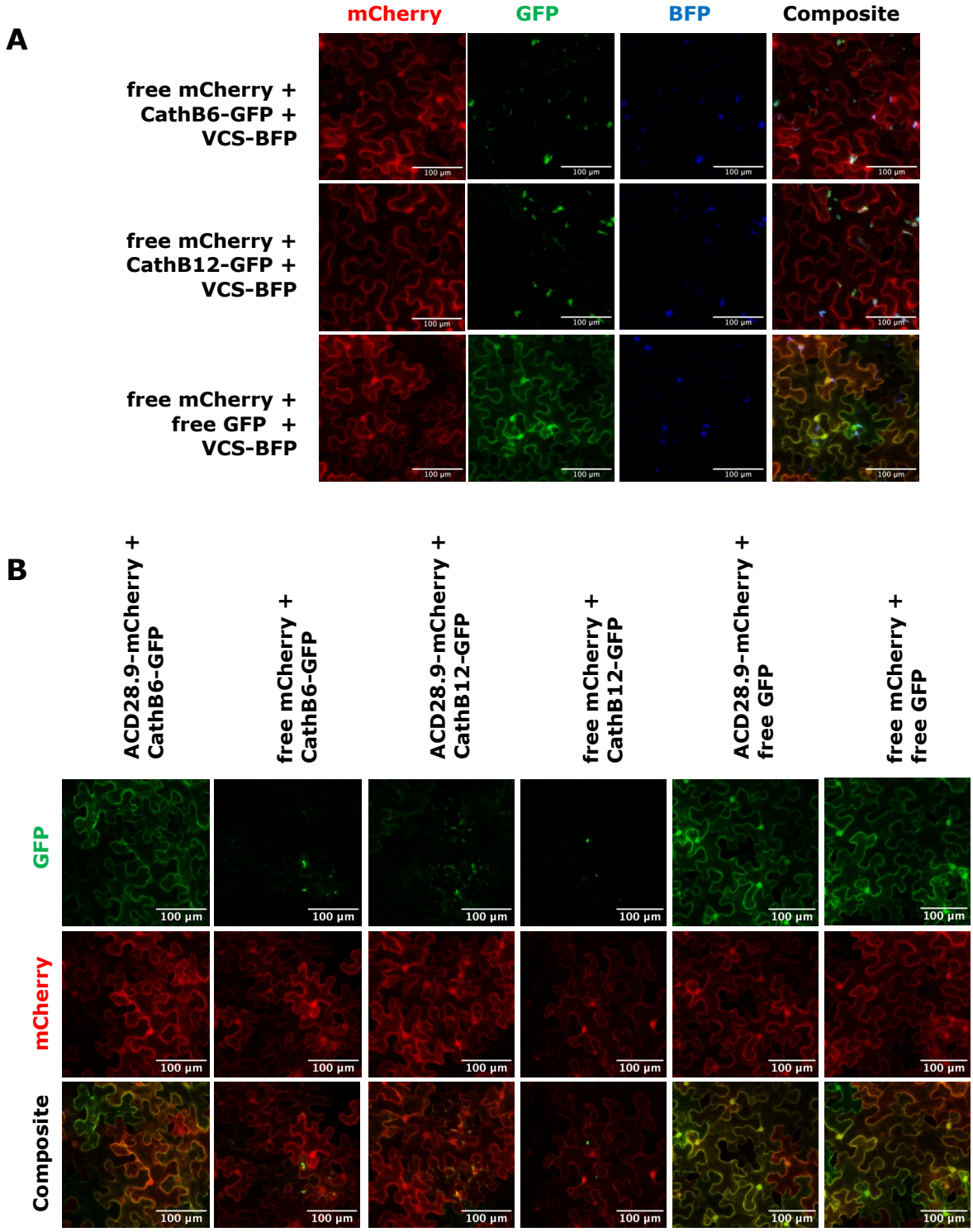


Figure 3.8: CathB12 forms puncta corresponding to p-bodies when expressed *in planta*, even when co-expressed with ACD28.9.
Co-infiltration of GFP-tagged CathB12 with mCherry-tagged ACD28.9 shows that ACD28.9 is incapable of removing CathB12 from p-bodies. This localisation during co-infiltration with ACD28.9 differs from the localisation of CathB6, which displays a cytoplasmic distribution when co-infiltrated with ACD28.9. **A:** Co-expression of CathB6 and CathB12 with free mCherry and Varicose-BFP shows that both CathB proteins are localised to p-bodies *in planta*. **B:** Co-expression of CathB6 and CathB12 with ACD28.9 highlights differing subcellular dynamics of the two CathB proteins in response to ACD28.9.

3.3.5: Two amino acids confer differential ACD28.9-binding capacity in CathB6 and CathB12.

MAFFT amino acid sequence alignment of the putative ACD28.9-binding regions of B6 and B12 (amino acids 263-333) shows two divergent residues between the two proteins. The first is a substitution of glutamate at position 307 in CathB6 to alanine in B12 (Figure 3.9), and the second is a phenylalanine to leucine substitution at position 314. Based on this minute difference in sequence, I generated constructs containing full coding sequences for the mature domain for CathB6 and B12, with single and double residue mutations to make the CathB6 more B12-like, and vice versa. Constructs encoding the full mature domain of CathB6 with single point mutations of glutamate to alanine at point 307 (CathB6_E307A), and phenylalanine to leucine at 314 (CathB6_F314L) were produced, as well as a double point mutation containing both of these changes (CathB6_E307A_F314L). The opposite changes were made in CathB12 to generate three more mutant constructs (CathB12_A307E, CathB12_L314F, CathB12_A307E_L314F).

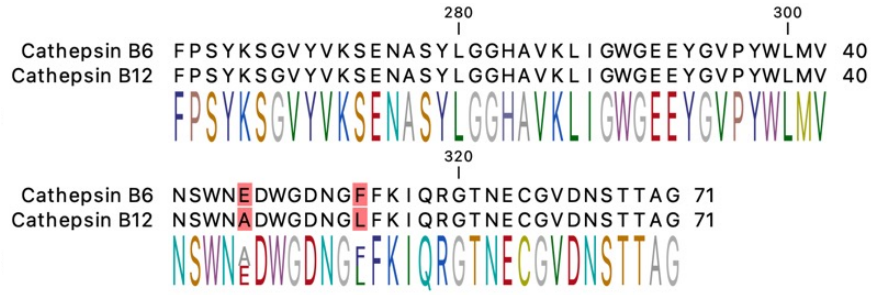


Figure 3.9: The putative ACD28.9-binding region of CathB6 has a high identity with the homologous region of the CathB12.
Amino acids 263-333 of CathB6 have been shown to be the region responsible for binding to ACD28.9. In CathB12, this region contains only two divergent residues at positions 307 and 314.

Mutated constructs were cloned into yeast-2-hybrid vector pDEST-GADT7 to assess binding potential of each construct with ACD28.9. Mutants were incapable of restoring growth on SD+ -LTHA media when co-transformed with ACD28.9, suggesting no interactions (Figure 3.10). There was also no evidence for dimerisation between mutants and CathB6 on this media, suggesting that there may be a loss of dimerisation between mutated CathB proteins. On the most stringent media SD+ -LTH 10mM 3AT, only CathB6 is capable of binding to

ACD28.9, once again pointing to a strong degree of binding between CathB6 and ACD28.9.

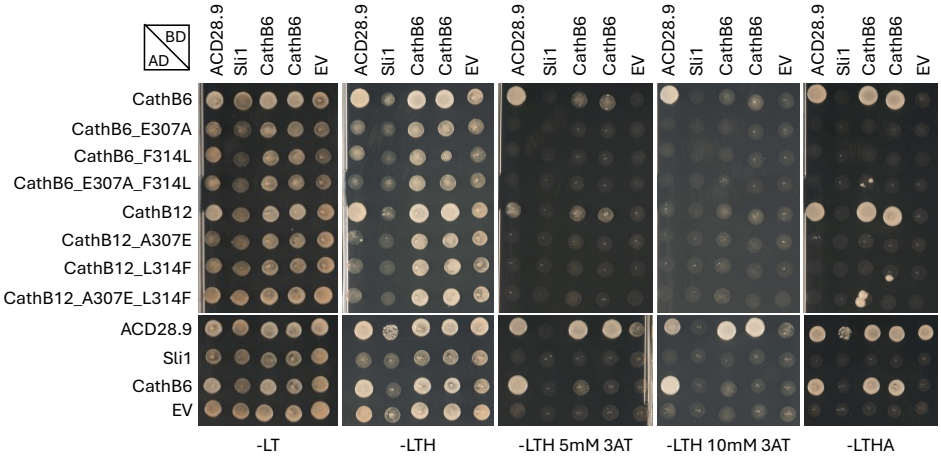
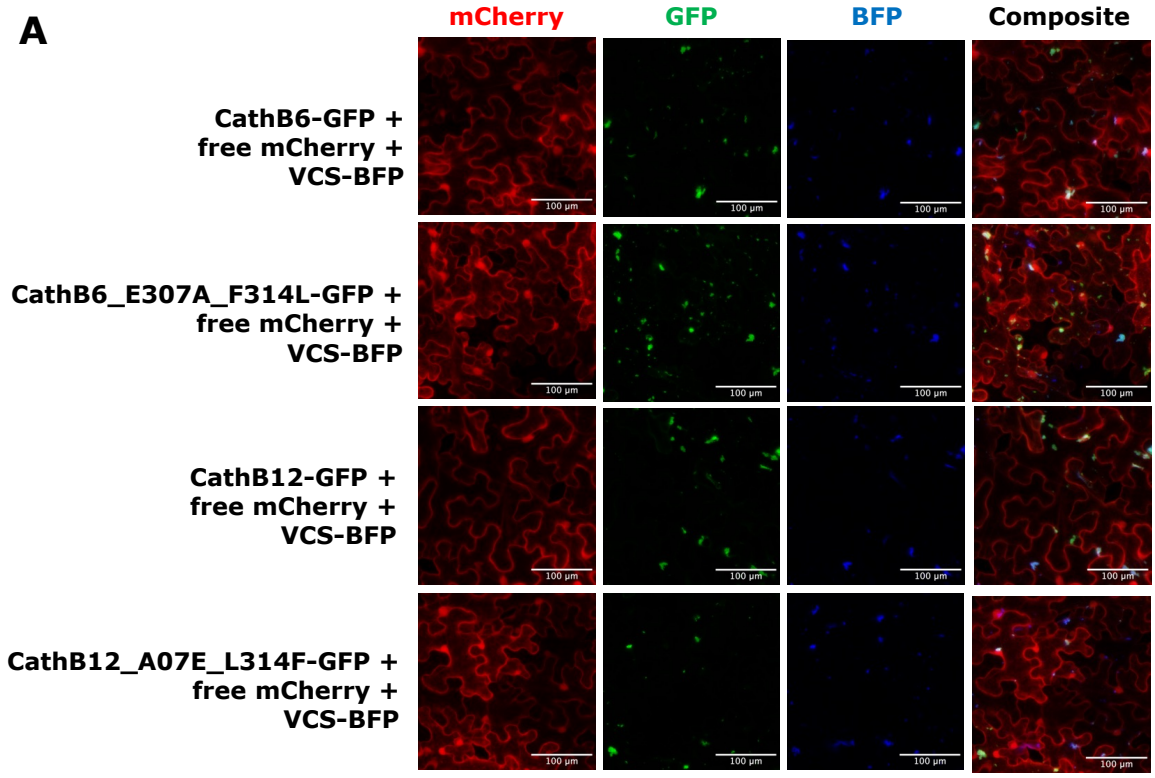
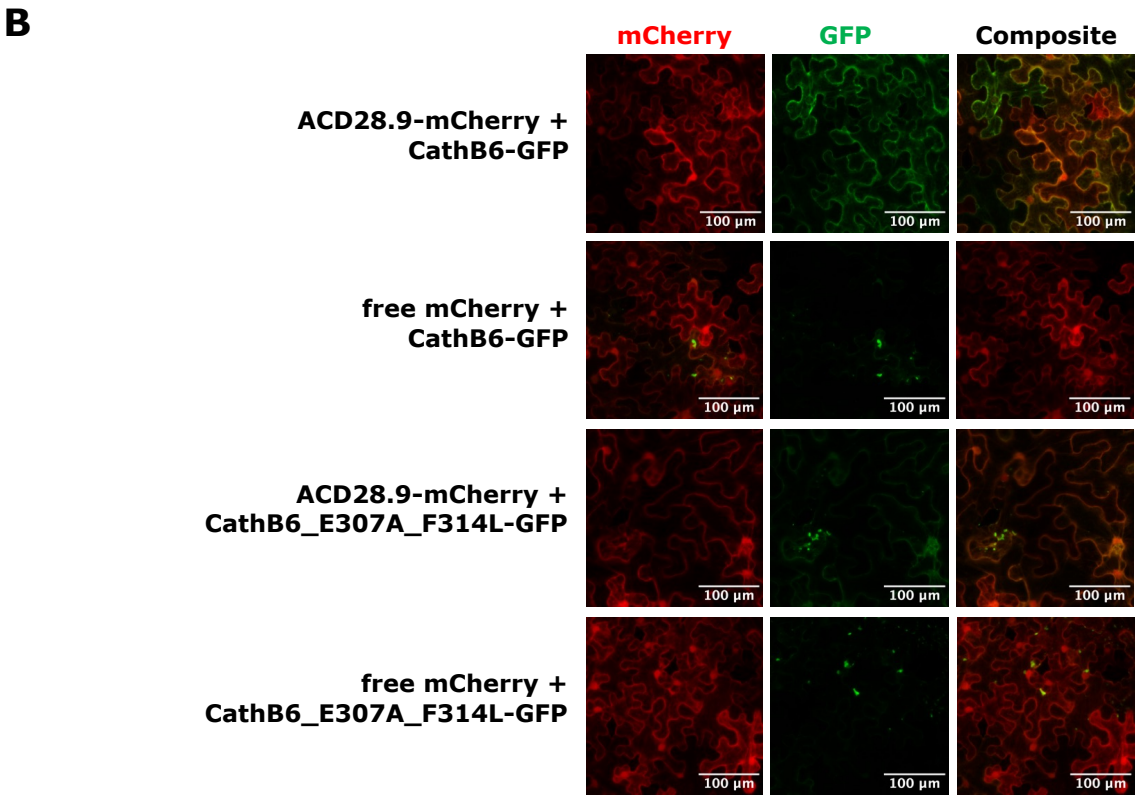


Figure 3.10: CathB6 and CathB12 mutants show no evidence of binding to ACD28.9 in yeast. Mutagenised CathB6 and CathB12 sequences containing variable residues at the two divergent sites were generated and tested for interactions with ACD28.9 and SLI1 in yeast-2-hybrid assays. Mutants showed no interactions, including homodimerization.

CathB6 and CathB12 double mutants (CathB6_E307A_F314L and CathB12_A307E_L314F respectively) were also cloned into pB7WG2F for use in agroinfiltration and confocal microscopy analysis. GFP-tagged CathB6_E307A_F314L and CathB12_A307E_L314F both colocalise with BFP-varicose, suggesting that both are recruited to p-bodies when expressed *in planta*, suggesting that the mutations do not alter the localisation of CathB6 or CathB12 in the absence of ACD28.9 (Figure 3.11A). The recruitment of CathB6_E307A_F314L to p-bodies appears to be maintained in the presence of ACD28.9, suggesting a loss of binding affinity compared to the wildtype CathB6, which is removed from p-bodies by ACD28.9 (Figure 3.11B). This suggests a lack of ACD28.9-binding affinity as a result of the two mutations in CathB6_E307A_F314L.

However, when CathB12_A307E_L314F is co-expressed with ACD28.9, there appears to be an elevated degree of removal from p-bodies by ACD28.9 when compared to CathB12, with CathB12_A307E_L314F-GFP being present in the cytoplasm of *N. benthamiana* cells when co-expressed with ACD28.9-mCherry (Figure 3.11C), suggesting that these mutations make CathB12 more available for ACD28.9 binding, leading to enhanced ACD28.9-mediated rescue from p-bodies.


A
mCherry
GFP
BFP
Composite
CathB6-GFP +
free mCherry +
VCS-BFP
CathB6_E307A_F314L-GFP +
free mCherry +
VCS-BFP
CathB12-GFP +
free mCherry +
VCS-BFP
CathB12_A07E_L314F-GFP +
free mCherry +
VCS-BFP
100 µm


B
mCherry
GFP
Composite
ACD28.9-mCherry +
CathB6-GFP
free mCherry +
CathB6-GFP
ACD28.9-mCherry +
CathB6_E307A_F314L-GFP
free mCherry +
CathB6_E307A_F314L-GFP
100 µm

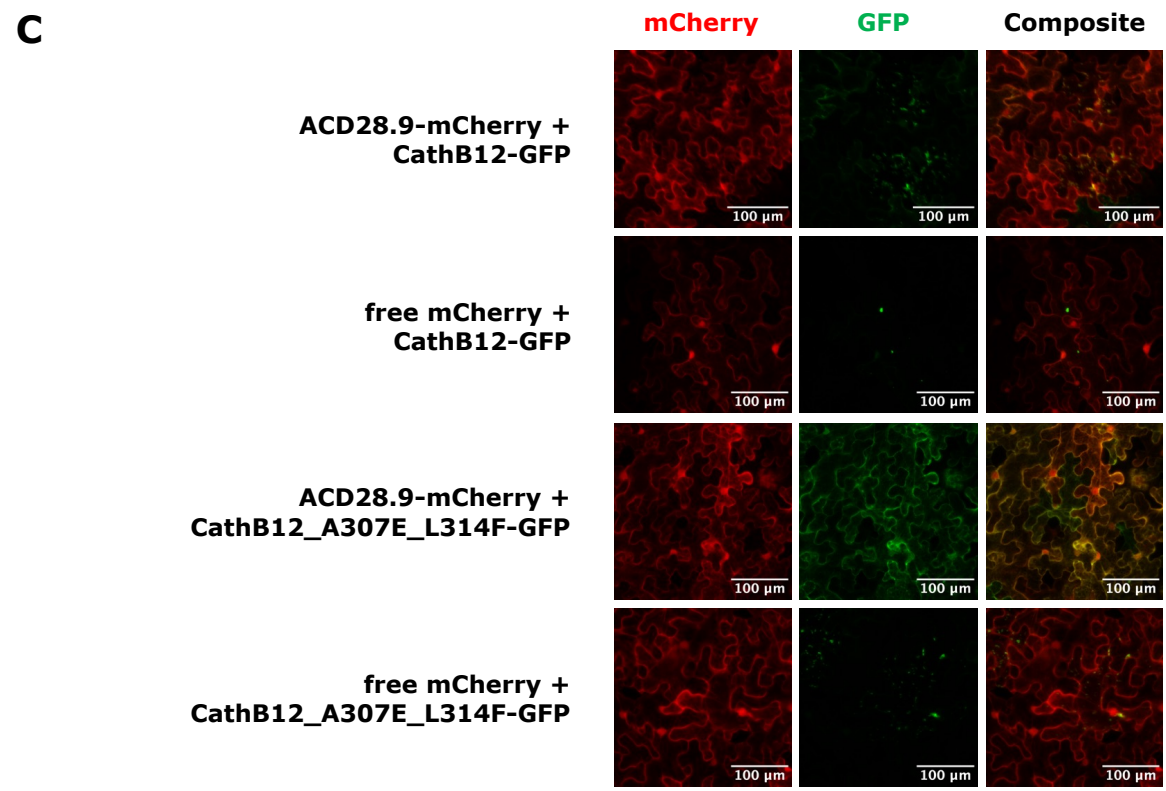


Figure 3.11: CathB6 and CathB12 double mutants with residue swaps at positions 307 and 314 lead to swapped ACD28.9-mediated p-body removal of CathB proteins.
Mutagenised CathB6 displays a more CathB12-like punctate localisation in the presence of ACD28.9, whereas the CathB12 double mutant is rescued from puncta by ACD28.9, similar to wildtype CathB6. **A:** Colocalisation of double mutants with varicose-BFP suggests that both CathB6_E307A_F314L-GFP and CathB12_E307A_L314F are recruited to p-bodies. **B:** CathB6_E307A_F314L-GFP maintains p-body localisation even when co-expressed with ACD28.9, similar to wildtype CathB12. **C:** CathB12_A307E_L314F-GFP is rescued from p-bodies by ACD28.9-mCherry, similar to CathB6.

3.4: Discussion

3.4.1: CathB proteins in a recently expanded clade bind differentially to ACD28.9.

Based on the results of yeast-2-hybrid assays to assess binding between members of the CathB protein family and the small heat shock protein ACD28.9, it appears that despite their high degree of sequence similarity, CathB proteins within the *Myzus*-specific expanded clade show differential binding affinity to ACD28.9. Most striking of these differences in binding affinity is between CathB6 and CathB12, in which the ACD-binding interface contains only two diverging nucleotides. In spite of this minimal divergence, the binding between CathB6 and ACD28.9 displayed in the above yeast-2-hybrid experiments is markedly stronger than that between CathB12 and ACD28.9. If such a small mutation can confer such a stark difference

in binding affinity, then it is possible that CathB12 diverged from CathB6 in order to avoid detection and relocalisation by ACD28.9.

3.4.2: ACD28.9 does not remove CathB12 from p-bodies during co-expression in N. benthamiana.

Based on the high degree of sequence similarity between CathB6 and CathB12, and the difference in potential for ACD28.9 binding, I tested whether the *in planta* localisation of the two proteins differs in the absence and presence of ACD28.9. Similar to CathB6, CathB12 is recruited to puncta when expressed in *N. benthamiana* cells, which appear to correspond to processing bodies based on the colocalization of p-body marker VCS. However, unlike CathB6, this recruitment to p-bodies is not diminished by the presence of ACD28.9. This difference in *in planta* dynamics may be a result of the small number of amino acid polymorphisms between the two proteins, particularly those in the region homologous to the ACD28.9 binding region of CathB6 (residues 263-333).

Creation of single and double mutants of CathB6 and CathB12 at these divergent sites (residues 307 and 314) appears to eliminate dimerisation in the two proteins in yeast-2-hybrid experiments, and eliminate the ability for both proteins to bind to ACD28.9. However, transient expression and subsequent confocal microscopy analysis of the two double mutants *in planta* reveals that in the absence of ACD28.9, both mutants are still recruited to p-bodies in a manner that is characteristic of the two wildtype proteins, suggesting that there may still be a degree of pathogenicity. Thus, it appears that there may be further troubleshooting required for the analysis of binding between ACD28.9 and CathB mutants in the yeast-2-hybrid system.

Interestingly, in the presence of ACD28.9, the CathB6 double mutant shows a subcellular distribution more similar to CathB12, with p-body localisation maintained in spite of ACD28.9. The opposite is true for the CathB12 double mutant, which is more readily rescued from p-bodies and returned to the cytoplasm by ACD28.9, akin to what is seen in co-expression of ACD28.9 and wildtype CathB6. This switch in localisation dynamics heavily implicates residues 307 and 314 as being integral to the binding of ACD28.9, and suggests that the

divergence of CathB12 was a response to the pressure from ACD28.9 recognition and the associated neutralisation.

While it is known that CathB6 and CathB12 localise to p-bodies based on their colocalization with varicose, the mechanism by which CathB proteins are held in p-bodies remains unknown. Processing bodies are multi-subunit aggregates composed of mRNA and proteins (Maldonado-Bonilla, 2014) and are responsible for the storage, translation and degradation of mRNAs (Kearly *et al.*, 2024). As a result, the composition of p-bodies is constantly in flux, with components being transported in and out based on a number of biological and molecular factors (Huang *et al.*, 2024). As such, there may be additional proteins which anchor the sequestered CathB to the p-bodies through direct interaction. It is possible that the capacity of ACD28.9 to remove CathB proteins from p-bodies may be dependent on the strength of CathB binding by both ACD28.9 and the p-body 'anchor' protein. I would hypothesise that if ACD28.9 binds the CathB protein more strongly than the binding between CathB and the 'anchor' protein, then p-body localisation is diminished, but if the opposite is true, then CathB proteins maintain their p-body specific distribution. To test this hypothesis, identification of such an 'anchor' protein would be required, and binding affinity of the various CathB proteins would need to be quantified.

3.4.3: Gene duplication may have allowed CathB proteins to avoid detection by ACD28.9.

Results of the yeast-2-hybrid assay point to differential ACD28.9-binding capacity across the CathB phylogeny. 90% of the CathB proteins which show evidence of ACD28.9 binding belong to the *Myzus*-specific expanded clade, and 70% are represented in the oral secretions of *M. persicae*. Interestingly, there are two instances of differential binding capacity between highly related pairs of CathB proteins. CathB6 binds ACD28.9, while CathB12 does not appear to bind as readily. Likewise, CathB14 binds ACD28.9, while CathB15 does not. These binding differences are in spite of a low number of residue differences in the regions homologous to the ACD28.9-binding region of CathB6, two in the case of CathB6 and CathB12, and eight in the case of CathB14 and CathB15.

Notably, CathB6 and CathB14 are encoded by genes present in a genomic cluster of CathB proteins on the *M. persicae* X chromosome (Figure 3.3A), and are both located directly adjacent to their closest related homolog, CathB12 and CathB15 respectively, which are incapable of binding ACD28.9. This points towards repeated duplication of CathB genes, and subsequent divergence of gene duplicates to avoid detection by ACD28.9. This supports the hypothesis that the expansion of this clade was driven by the selective pressure to avoid detection and neutralisation by ACD28.9, and points to multiple instances of ACD28.9-driven divergence within the expanded clade as a result of repeated gene duplication.

3.4.4: CathB proteins do not seem to be the target of SLI1-mediated resistance.

According to yeast-2-hybrid experiments, there is no evidence for SLI1 binding to any *M. persicae* CathB proteins. This suggests that while ACD28.9 likely targets CathB6 and other members of the CathB family to confer resistance, this mode of action may not be ubiquitous across other small heat shock proteins. It also suggests an alternative mode of action for SLI1-mediated resistance, though this line of investigation is yet to be explored.

3.4.5: Advances and future potential.

The research discussed in this chapter have enhanced the understanding of the evolutionary pressures which have driven a lineage-specific expansion of CathB effectors in *M. persicae*. ACD28.9 is capable of binding to, and as such potentially diminishing the virulence of, multiple CathB proteins, which has driven multiple instances of divergence within the CathB family, giving rise to proteins which are capable of being imperceptible to ACD28.9.

However, the exploration of the interactions between members of the CathB family and ACD28.9 can be progressed in a number of ways which would improve both the breadth and depth of our understanding. Currently, only 3 CathB proteins have been assessed for localisation *in planta*, only 2 of which are capable of binding ACD28.9. While p-body localisation of CathB proteins in the absence of ACD28.9 appears to be consistent across the tested proteins, it is unclear whether this localisation is conserved across all CathB proteins present in *M. persicae* oral secretions. Additionally, the sequences of differentially-binding pairs of CathB

proteins could be mined to identify signs of positive selection, indicating additional residues of interest in the binding between some CathB proteins and ACD28.9.

Chapter 4 – *A. thaliana* proteins from a Brassicales-expanded clade of ACD proteins contain multiple ACDs and are capable of binding to *M. persicae* CathB

4.1: Introduction

All organisms, including plants, are capable of adjusting gene expression levels in response to stresses. A group of genes initially identified based on their differential regulation in response to temperature are those encoding heat shock proteins, which have since been shown to be responsive to a number of biotic and abiotic stress factors. Heat shock protein gene expression in stressed cells contributes to cellular homeostasis, preventing cell death and conferring increased organismal fitness.

These heat shock proteins have been broadly categorised into several families, including, but not restricted to, HSP100, HSP90, HSP70, HSP60, and HSP20 protein families, with multiple members of each family found in most organisms. Members of the HSP20 family are also termed small heat shock proteins (sHSPs), due to their relatively low molecular weight. Members of this family also feature one or more alpha-crystallin domains (ACDs) (Caspers *et al.*, 1995). While the sHSPs include majoritively small proteins with molecular weights ranging from 15 to 30 kD, ACD-containing proteins have been identified with a much larger range of molecular weights, some over 100 kD (Scharf *et al.*, 2001).

Scharf *et al.* 2001 described the HSP20 family of *Arabidopsis thaliana*, and additionally described a distinct family of proteins containing ACDs. Recently, two of these ACD-containing proteins have been implicated in plant defence to aphid herbivory. Sieve element-lining chaperone 1 (SLI1)—alternatively identified as ACD54.2 (Scharf *et al.,* 2001)—is a 54.2 kD protein which has been characterised as a broad-spectrum herbivore resistance factor by genome-wide association mapping of aphid feeding behaviour on 350 *A. thaliana* accessions (Kloth *et al.*, 2017). Notably, SLI1 reduces fitness of *Myzus persicae* as well as another aphid species, and one whitefly species. Another small heat shock protein, ACD28.9, has been found to interact with *M. persicae* Cathepsin B (CathB) effector proteins within plant cell cytoplasm, and suppresses the activities of the CathB effectors to modulate aspects of effector triggered immunity. More specifically, ACD28.9 reduces the capacity for CathB6 to recruit immune regulator Enhanced Disease Susceptibility 1 (EDS1), and its partners Phytoalexin Deficient 4 (PAD4) and Activated Disease Resistance 1 (ADR1), to processing bodies, instead restoring the cytoplasmic distribution of CathB and EDS1 (Liu *et al.*, 2025).

Existing phylogenies of heat shock proteins in plants have typically focused on small subsets of the broader phylogeny of plants, focusing on a single species or group alongside model species. For example, genome wide analyses of peach (Lian *et al.*, 2022), tomato (Yu *et al.*, 2016), wheat (Muthusamy *et al.*, 2017, Wang *et al.*, 2017), and Chinese cabbage (Tao *et al.*, 2015), have provided identification and expression analysis of small heat shock proteins in response to a wide array of biotic and abiotic stresses.

HSP20 family members have previously been identified using Hidden Markov Models (HMMs) that capture conserved amino acid sequence features of ACDs. For example, this method was integral to the identification of small heat shock proteins in pumpkin (*Cucurbita moschata*) (Hu *et al.*, 2021). HMMs are widely used for the identification and downstream analysis of a broad range of protein domains. Notably, HMMs were utilised to identify and create a phylogeny of a *Phytopthora*-responsive immune receptor from wild potato called PERU (Torres Ascurra *et al.*, 2023).

In previous studies described in this thesis, it was uncovered that the evolutionary trajectory of the CathB protein family may have been shaped by an arms race between *M. persicae* and plants. Given that CathB effectors belong to a recently expanded clade within the *M. persicae* CathB phylogeny and ACDs are part of a larger family, it is possible that proteins related to ACD28.9 may interact with one or more aphid CathB proteins and affect their effector activities. As a first step to investigate this further, I wished to better understand the diversity of ACDs in *A. thaliana* and other Brassicaceae and how these relate to those in other plant species. As such, work in this chapter describes a domain discovery pipeline to identify proteins containing ACDs across 180 plant species contained in the NCBI RefSeq database (Pruitt *et al.*, 2007, Toghani and Kamoun, 2024) and a phylogenetic analysis of HSP20/ACD protein family members discovered in these plants. This revealed that ACD28.9 belongs to a clade of HSP20s which appears to be rosid-specific and hugely expanded in Brassicales. ACD28.9 and the other *A. thaliana* members of this clade contain at least two ACDs, and appear to interact with *M. persicae* CathB proteins. Moreover, the N-terminal alpha-crystallin domain, but not the C-terminal one, of ACD28.9 is responsible for CathB binding.

4.2: Materials and methods

4.2.1: ACD phylogeny

The raw hidden Markov model for the alpha-crystallin/HSP20 domain was downloaded from EMBL-EBI Interpro (PFAM database: PF00011). A compiled proteome of 180 plant species was generated and shared by AmirAli Toghani (Kamoun Lab, The Sainsbury Laboratory), using genomes downloaded from the National Center for Biotechnology Information (NCBI) genome database (Supplementary Table 2). All genomes for green plants (Viridiplantae) which had been annotated by NCBI RefSeq (Pruitt *et al.*, 2007) were downloaded and used to create a concatenated proteome. This dataset was mined for instances of the alpha-crystallin/HSP20 domain using the HMMSearch command from the HMMER tool package.

Alpha-crystallin/HSP20 domain-containing-proteins from *Mus musculus*, *Homo sapiens*, *Marchantia polymorpha*, and *Pseudomonas syringae* were downloaded from the Interpro sequence database, and concatenated to be used as potential outgroups for downstream phylogenetic analysis. A list of previously identified HSP20 proteins from *Arabidopsis thaliana* was used as a reference dataset for phylogenetic analysis.

The tabular output from HMMSearch, as well as the outgroup set, the reference set, and the taxonomic metadata for the 180-species combined plant proteome, were read into RStudio. Sequences with a whole sequence e-value <0.01 were deemed significantly likely to contain a true alpha-crystallin domain, and filtered from the HMMSearch tabular output. IDs of proteins identified as containing an alpha-crystallin/HSP20 domain by HMMSearch was extracted from the tabular output, and the resulting list was used to extract protein sequences from the proteome dataset using the seqkit grep function. The resulting fasta file containing protein sequences for alpha-crystallin/HSP20 domain -containing proteins was read into R for gene length analysis and filtering. Following gene length filtering, protein fasta files from the outgroup, reference, and resultant datasets were concatenated and exported for multiple sequence alignment using MAFFT (Katoh *et al.*, 2002). Phylogenetic trees were constructed using FastTree (Price *et al.*, 2009).

4.2.2: Identification and alignment of ACDs in Arabidopsis thaliana HSP20s

HMMSearch was used against the protein sequences of previously identified *Arabidopsis thaliana* ACD family members (Scharf *et al.*, 2001) using the raw HMM file downloaded from EMBL-EBI Interpro. The HMMSearch –domtblout flag was used to generate a tab separated values (.tsv) file with gene names and sequence coordinates of each determined alpha-crystallin domain. The SeqKit (Shen *et al.*, 2024) subseq function was used to extract these regions from the original protein fasta file, and alpha-crystallin domains from proteins with a whole sequence e-value < 0.01 were aligned using MAFFT, and phylogenetic trees were constructed using FastTree (Price *et al.*, 2009).

4.2.3: AlphaFold modelling and visualisation

Models were created by entering protein sequences of the relevant proteins into the ColabFold web server (Mirdita *et al.*, 2022, Kim *et al.*, 2025). Modelling was performed on the site with "recycle_early_stop_tolerance" = 0.0, "num_recycles" = 6, and all other parameters as default. The highest ranked model prediction was identified upon completion of the run, and the predicted template modelling (pTM) score was noted. Models were visualised and customised in Chimera X (Meng *et al.*, 2023).

4.2.4: Gateway cloning

Fragments corresponding to M1-K162 and G163-E264 respectively of ACD28.9 (AT5G47690) were amplified and flanked with AttB adapters through PCR using Phusion polymerase. 1 µL pDESTGADT7-ACD28.9 was added to a PCR mix containing 1× Phusion High Fidelity Buffer, 200 µM dNTPs, 0.5 µM each of the respective forward and reverse primers, and 1 U Phusion DNA polymerase, in a final reaction volume of 25 µL. Reaction mixes were added to a PCR Max thermocycler programmed to run the following cycles: polymerase activation at 98 °c for 2 minutes, 35 cycles of (98 °C for 10 seconds, 55 °C for 15 seconds, 72 °C for 1 minute 30 seconds), a final extension step at 72 °C for 10 minutes, followed by a final store step indefinitely at 4 °C. AttB adapted fragments were cloned into pDONR207 through a BP Clonase reaction (see 3.2.2.1), then into pDESTGADT7 and pDESTGBKT7 through an LR Clonase reaction (see 3.2.2.3).

Yeast-2-hybrid destination plasmids were also created for the full coding sequences of ACD26.0, ACD25.4, and ACD81.4. To achieve this, 1 µL plasmid containing each full coding sequence was used as a template for PCR with Phusion polymerase, using gene-specific primers with flanking AttB sequences. PCR set up was identical to that used to amplify ACD28.9 fragments (see above). AttB adapted fragments were cloned into pDONR207 through a BP Clonase reaction (see 3.2.2.1), then into pDESTGADT7 and pDESTGBKT7 through an LR Clonase reaction (see 3.2.2.3).

Site-directed mutagenesis was used to generate yeast-2-hybrid destination plasmids ACD16.9, ACD22.1, ACD21.4, and ACD28.7. Overlapping primers containing the desired mutation were used to amplify pTWIST-ENTR plasmids containing a stop codon immediately following the AttB site. Reaction mixes were added to a PCR Max thermocycler programmed to run the following cycles: polymerase activation at 95 °C for 5 minutes, 40 cycles of (95 °C for 30 seconds, 50 °C for 30 seconds, 72 °C for 6 minutes), a final extension step at 72 °C for 5 minutes, followed by a final store step indefinitely at 4 °C. PCR products were visualised through gel electrophoresis, and appropriately sized bands were excised. QIAquick gel extraction kit (Qiagen) was used to purify DNA from excised bands. Extracted DNA was quantified using the Qubit dsDNA broad range assay (Invitrogen). 10 µL NEBuilder HiFi DNA Assembly Master Mix was added to 0.1 pmol purified dsDNA, and deionised water was added to make a final reaction volume of 20 µL. Samples were incubated at 50 °C for 15 minutes, then transformed into NEB5alpha (New England Biosciences) competent *E. coli* cells by heat shock (see 3.2.2.2). Transformants were screened for successfully mutagenized plasmids using Sanger sequencing.

4.3: Results

4.3.1: ACD28.9 and SLI1 (ACD54.2) belong to distinct clades in the A. thaliana ACD protein family.

Previous work in this thesis revealed that *M. persicae* CathB effectors do not interact with SLI1 in yeast, in contrast to ACD28.9 (Figure 3.9). To determine the evolutionary relationship between ACD28.9 and SLI1 (ACD54.2), which are both implicated in resistance to herbivorous hemipterans including *M. persicae*, protein

sequences of the 23 *A. thaliana* ACD-containing proteins detailed in Scharf, 2001 were aligned to generate a phylogeny. Multiple sequence alignment of the two proteins revealed a low sequence similarity (14.75%). Moreover, phylogenetic tree construction of the *A. thaliana* ACD proteins established that ACD28.9 and SLI1 belong to distinct clades (Figure 4.1). Thus, whereas ACD28.9 and SLI1 both contain ACDs, their protein sequences are relatively distinct, hereby providing an explanation for the finding that *M. persicae* CathB proteins bind ACD28.9, but not SLI1.

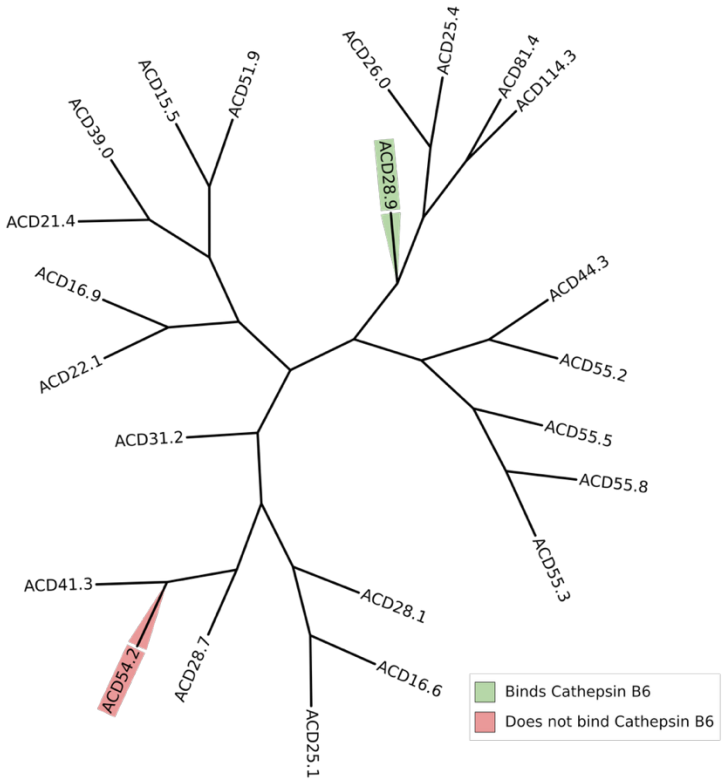


Figure 4.1: ACD28.9 and SLI1 (ACD54.2) belong to different clades within the *A. thaliana* ACD protein phylogeny.
Phylogenetic tree of *A. thaliana* ACD-containing proteins described in Scharf *et al.*, 2001, based on whole sequence MAFFT alignment. SLI1 labelled in red, ACD28.9 labelled in green.

4.3.2: HMMSearch is capable of identifying HSP20 family members using the ACD HMM.

To expand the understanding of phylogenetic relationships among HSP20 and ACD-containing protein families beyond *A. thaliana*, I performed a bioinformatics pipeline modified from Torres Ascurra *et al.* 2023, to identify additional ACD-containing proteins within 180 plant proteomes from the NCBI Refseq database (Pruitt *et al.*, 2007) . To do this, the raw HMM for the HSP20/alpha-crystallin family

(PF000111) was downloaded from the PFAM database (Figure 4.2), and used as a template in an HMM search using the HMMSearch tool from the HMMER package (Eddy, 2011).

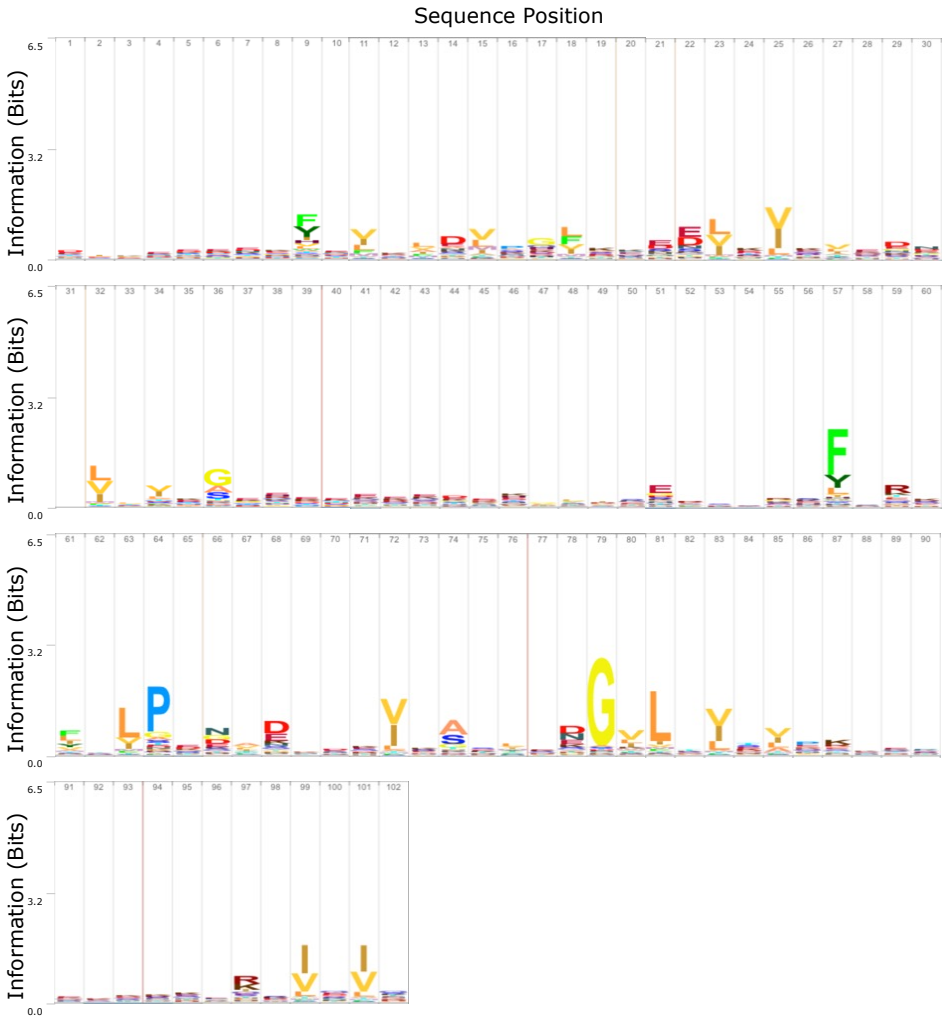


Figure 4.2: Sequence logo of the Hidden Markov Model (HMM) for HSP20/ACD protein family used in HMM identification pipeline.
Logo showing likelihood of a given amino acid residue at each sequence position based on the probability matrix in the HMM, adapted from https://www.ebi.ac.uk/interpro/entry/pfam/PF00011/logo/.

The HMM search with this profile captured 11521 sequences from the concatenated proteomes of the 180 Refseq plant species provided by AmirAli Toghani (Kamoun Lab, The Sainsbury Laboratory) (Supplementary Table 2). Of these, 9832 proteins had e-values < 0.01 for HMM matches against the whole sequences and were thus retained. The 9832 proteins had a mean length of 232 amino acids (standard deviation, ±202 amino acids), with a minimum length of 55 amino acids, and a maximum length of 4897 (Figure 4.3, top panel). To improve

the quality of the alignments used to generate a phylogenetic tree, 52 sequences with lengths over 1000 amino acids were removed, leaving 9780 sequences that had a mean length of 220 amino acids (standard deviation, ±113) (Figure 4.3, middle panel). To further optimize the alignment and phylogeny, an additional length filtering step was implemented to only include proteins <600 amino acids in lengths. The final filtered list contained 9638 proteins with a mean length of 213 amino acids (standard deviation, ±94) (Figure 4.3, bottom panel).

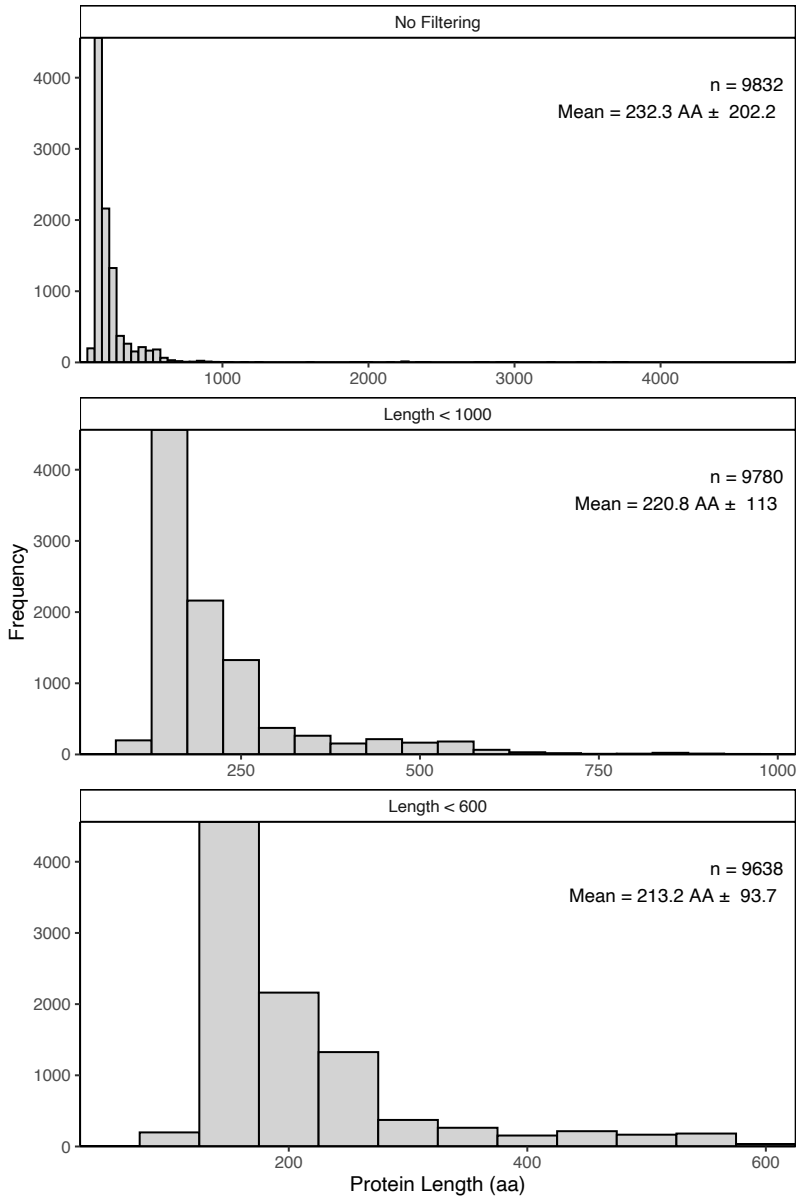


Figure 4.3: Length filtering was used on identified ACD proteins to remove outliers and improve alignments.
Protein length distributions of Hsp20/ACD family proteins identified in 180 plant species upon various filtering steps. **Top panel:** 9832 proteins with e-values < 0.01 against the ACD HMM. **Middle panel:** as top panel, but upon removal of 52 sequences with lengths over 1000 amino acids. **Bottom panel:** as middle panel, but upon removal of sequences with lengths over 600 amino acids.

The HMM Search against the Ref-seq database identified 17 additional *A. thaliana* proteins containing ACDs, which correspond to those described as HSP20 proteins in Scharf *et al.* 2001. This provides evidence that this HMM profile is suitable for the identification of HSP20 proteins. Alignment and phylogenetic analysis of HSP20 proteins with the ACD proteins named in Scharf *et al.* 2001, show that HSP20 proteins belong to a distinct clade from the ACD proteins containing ACD28.9 and SLI1 (Figure 4.4), and as such that the HMM is capable of comprehensive coverage of ACD-containing proteins in *Arabidopsis* including HSP20 proteins, and likely beyond. As such, hereafter all proteins containing ACDs—including members of the HSP20 family—will be referred to as ACD proteins.

Figure 4.4: The ACD HMM search captured 17 *A. thaliana* HSP20 proteins
Phylogenetic tree showing the *A. thaliana* ACD proteins identified by the HMM Search pipeline. These proteins include those identified as both HSP20 and ACD proteins in Scharf *et al.*, 2001. HSP20 proteins (pink clade) are evolutionarily distinct from the ACD family proteins (green clade) which were used as references in the HMM search. ACD28.9 and SLI1 (ACD54.2) are highlighted in the tree.

4.3.3: ACD28.9 and 4 other A. thaliana ACD proteins belong to a Brassica-expanded subclade of the plant ACD protein family.

ACD protein sequences from a non-vascular plant species (*Marchantia polymorpha*), mouse (*Mus musculus*), human (*Homo sapiens*), and a bacterium (*Pseudomonas syringae*) were downloaded from the PFAM database, and added to the sequence output of the HMMsearch. However, none of the sequences of the above species clustered into a separate clade that may constitute an outgroup, and so the phylogenetic tree of plant ACD proteins remained unrooted. This is consistent with the notion that the ACD protein family is ancient, likely predating the divergence of bacteria and eukaryotes.

The 9638 ACD proteins from the 180 analysed plant proteomes grouped into multiple clades containing subclusters of proteins from orders of plants such as Brassicales, Solanales and Fabales, pointing to significant lineage-specific expansion of early diverging ACD proteins (Figure 4.5).

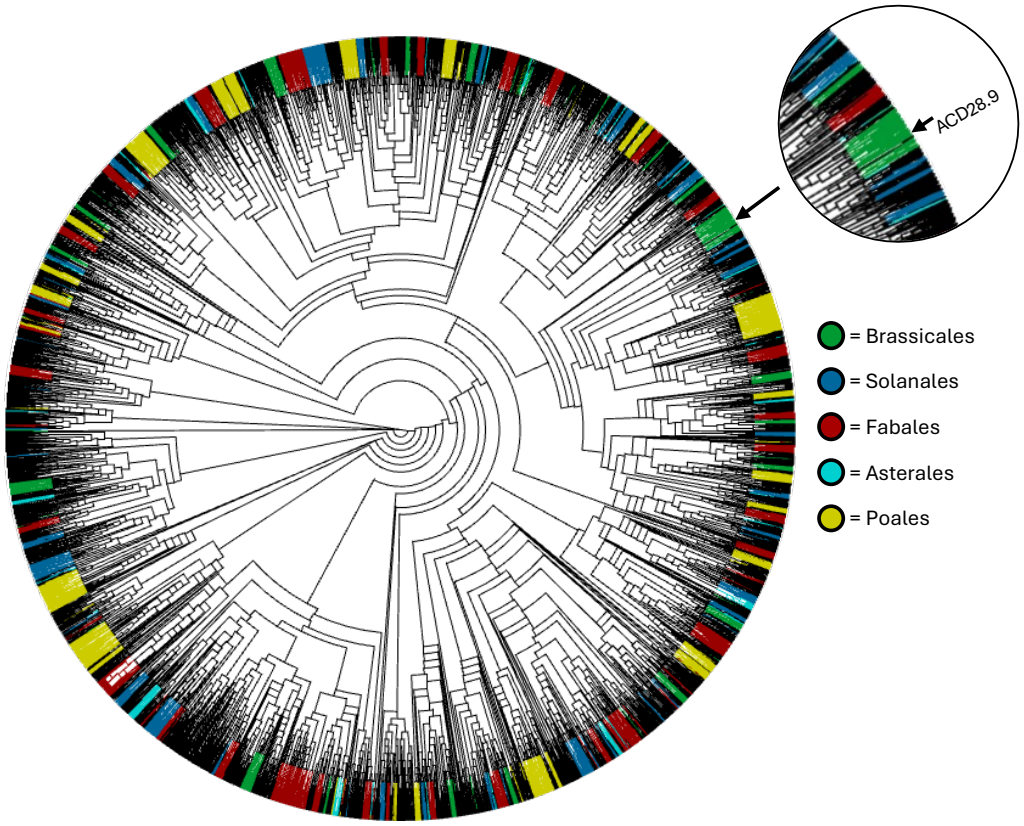


Figure 4.5: Plant ACD proteins show an evolutionary history of lineage-specific expansions.
Phylogenetic tree of length-filtered ACD proteins across the 180 plant species in the Ref-seq database. The tree remained unrooted due to the lack of a suitable outgroup. The tree shows several instances of lineage-specific expansions within the groups of plants determined to be colonisable by *M. persicae* (Mathers *et al.*, 2017).

Based on the observation that ACD28.9 is positioned in a subclade of the plant ACD protein phylogeny which appears to be expanded in the Brassicales, I selected a subclade with a bootstrap value of 1.00 to explore further. This subclade of the plant ACD protein phylogeny which contains 119 proteins, including ACD28.9, appears to be rosid-specific, and shows a distinct expansion in the brassicaceous plants. The four other *A. thaliana* ACD protein present in this subclade are ACD81.4, ACD114.3, ACD25.4 and ACD26.0. ACD28.9 appears to have diverged from the other *A. thaliana* ACD protein in this subclade prior to the divergence of Brassicales and other rosids (Figure 4.6). Notably, this phylogenetic analysis also revealed several ACD28.9 homologs in closely related species including *Arabidopsis lyrata*, *Brassica napus* and *Brassica oleracea*.

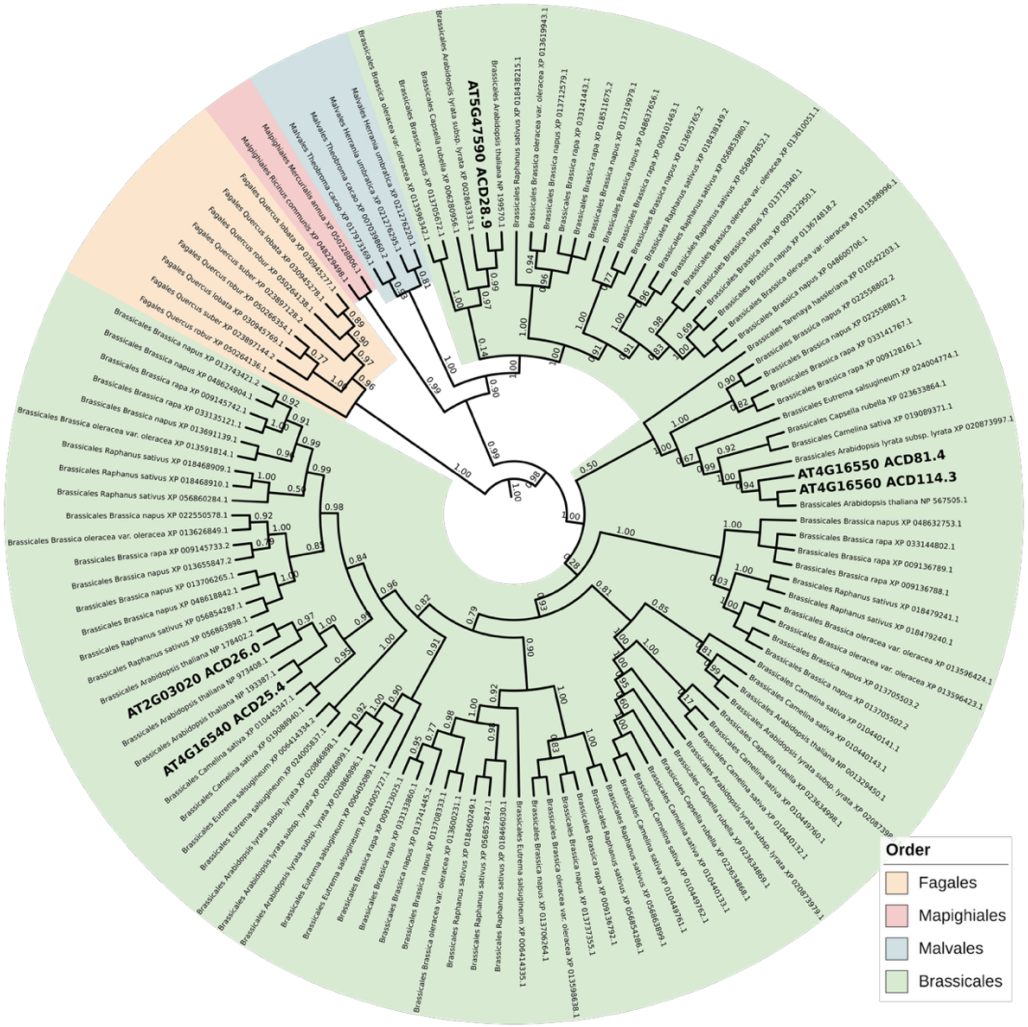


Figure 4.6: ACD28.9 lies in a rosid-specific clade of ACD proteins, which also contains four other *Arabidopsis thaliana* ACD proteins.
This clade is greatly expanded in the Brassicales order of plants. Subtree of the clade containing ACD28.9, with plant taxa labelled. This subclade contains four other Arabidopsis ACD proteins - ACD81.4, ACD114.3, ACD25.4 and ACD26.0.

4.3.4: A. thaliana proteins in the ACD28.9 subclade contain multiple ACDs.

Analysis of the domain output of HMMSearch identified the presence of multiple ACDs in a number of ACD proteins. In *A. thaliana*, 38 unique ACDs were identified by HMMSearch against the ACD/HSP20 hidden Markov model, belonging to 22 of the 23 ACD proteins named in Scharf *et al.*, 2001. 35 of these domain predictions were deemed significant (whole sequence e-value < 0.01). As such, it was determined that a number of *A. thaliana* ACD proteins contain two ACDs. This includes all members of the ACD28.9 subclade. The largest of the proteins in the ACD28.9 subclade, ACD114.3, contains three predicted ACDs (Figure 4.7). Interestingly, SLI1 (ACD54.2) also contains two ACDs, but these seem to be phylogenetically removed from the ACD28.9 subclade domains. Based on phylogenetic analysis, the ACDs in the ACD28.9-subclade HSP20s appear to have diverged prior to the expansion of the clade, forming two distinct groups in which the respective N-terminal and C-terminal ACDs show higher similarity between proteins than between ACDs in the same protein.

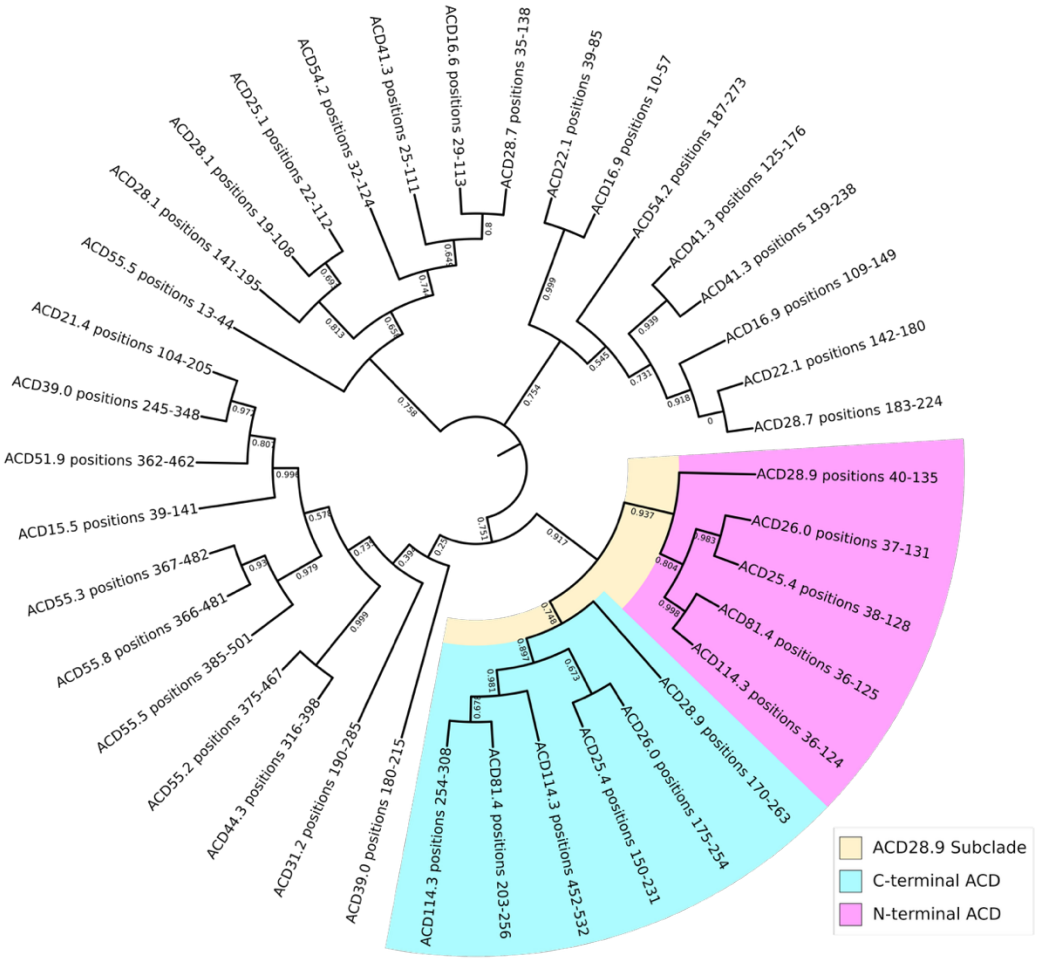


Figure 4.7: All *A. thaliana* members of the ACD28.9 subclade contain ≥2 alpha-crystallin domains. Phylogenetic tree of alignments between domains identified across previously described Arabidopsis ACD proteins (Scharf *et al.*, 2001). The ACDs of proteins in the ACD28.9 subclade are highlighted in yellow, with the N-terminal ACDs further highlighted pink, and the C-terminal ACDs blue.

Given that the two ACDs in ACD28.9 appear to be evolutionarily distinct from one another, I performed a structural prediction and sequence alignment of the two domains. Alphafold2 prediction yielded high confidence predictions for the regions corresponding to the predicted ACDs (Figure 4.8A), which appear to form the β-sandwich structure characteristic of ACDs (Figures 4.8B and 4.8C). Despite seeming to form analogous secondary structures, the sequences of the two ACDs are quite different, with only a 21% sequence identity (Figure 4.8D).

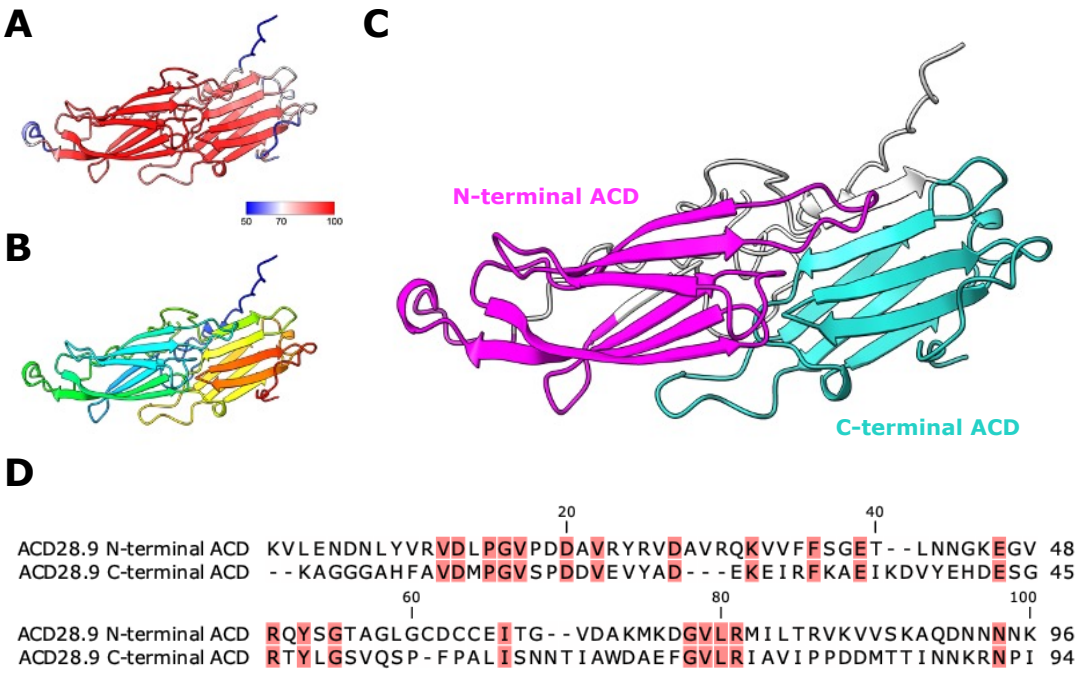


Figure 4.8: ACD28.9 contains two evolutionarily distinct ACDs.
Alphafold models of ACD28.9 show two β-sandwich structures indicative of its duplicated ACDs. **A:** Model showing prediction confidence (blue = low confidence, red = high confidence) **B:** Rainbow model displaying amino acid positioning within the ACD28.9 structure (blue = n-terminus, red = c-terminus) **C:** Model showing the two alpha-crystallin domains predicted by HMMSearch. **D:** Sequence alignment of the N- and C-terminal alpha-crystallin domains in ACD28.9, with conserved residues highlighted in red.

4.3.5: The N-terminal ACD of ACD28.9 is responsible for dimerization and binding to CathB proteins.

Yeast-2-hybrid destination plasmids were generated containing the N-terminal and C-terminal alpha-crystallin domains of ACD28.9, with which yeast was co-transformed to assess binding to CathB6 and CathB12, as well as capacity for homodimerisation. Growth of yeast on media lacking leucine, tryptophane, histidine, and adenine was rescued in yeast containing the N-terminal ACD, suggesting interactions between the N-terminal ACD of ACD28.9 with both CathB6 and CathB12 (Figure 4.9). The N-terminal ACD may also be involved in ACD28.9 homodimerisation, based on the growth of yeast double transformed with full-length ACD28.9 and the N-terminal ACD. Yeast proliferation could not be rescued at high stringency in any mutants containing only the C-terminal ACD of ACD28.9, suggesting that this domain may not be involved in the ACD28.9 interaction with CathB proteins.

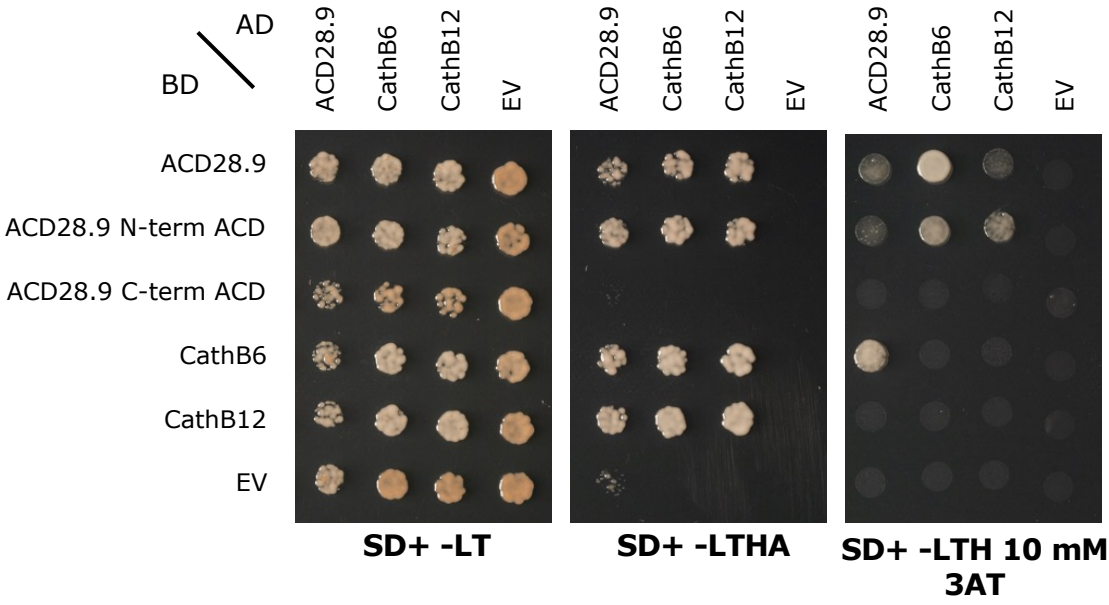


Figure 4.9: The N-terminal ACD of ACD28.9 is responsible for the interaction between ACD28.9 and CathB proteins.
Yeast-2-hybrid assay suggests that the N-terminal ACD of ACD28.9 is responsible for binding to CathB6 and CathB12. Growth on SD+ -LT media shows successful double transformation of all pairwise comparisons. Growth on SD+ -LTHA shows binding only between the N-terminal ACD with both CathB6 and CathB12. The C-terminal ACD does not appear to bind to either CathB6 or CathB12. Removing the N-terminal ACD also appears to eliminate the ability of ACD28.9 to homodimerize.

4.3.6: Members of the ACD28.9 subclade are capable of binding to CathB proteins.

Based on the previous phylogenetic analysis, I elected to explore the capability of various members of the *A. thaliana* ACD protein family. This included 4 of the 5 ACD proteins in the ACD28.9 subclade, as well as a number of proteins from outside the subclade. ACD114.3 was excluded from the yeast-2-hybrid due to its large size, which made generation of suitable plasmids a difficulty. The assay revealed that the 4 tested members of the ACD28.9 subclade are all capable of binding to CathB6 and CathB12 (Figure 4.10, middle panel), though there does appear to be a difference in the strength of binding to the CathB proteins, with ACD26.0, ACD25.4 and ACD81.4 showing no evidence of interaction at high stringency, compared to ACD28.9 which does still show evidence of binding (fig 4.10, bottom panel). This is the first evidence that the binding of CathB6 is not unique to ACD28.9, but extends to other members of the ACD protein family of *A. thaliana*.

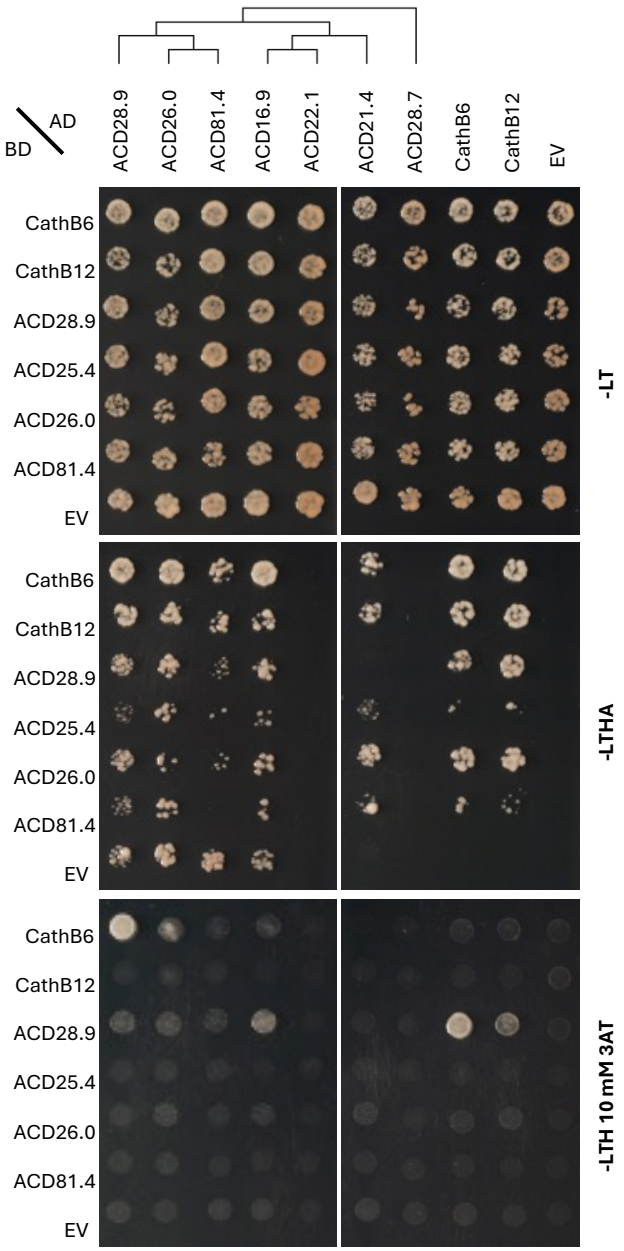


Figure 4.10: Members of the *A. thaliana* ACD28.9 subclade are capable of binding CathB proteins. Yeast-2-hybrid results show binding of CathB6 and CathB12 to a number of ACD proteins in the ACD28.9 subclade. **Top panel:** -LT media showing successful transformation. **Middle panel:** -LTHA medium stringency media for interactions shows that ACD28.9, ACD25.4, ACD26.0 and ACD81.4 all show binding to CathB6 and CathB12. **Bottom panel:** -LT media with 10 mM 3AT shows that at the highest stringency only ACD28.9 binds CathB6 and CathB12.

Following the findings that multiple *Arabidopsis* ACD proteins can bind to CathB6, and based on the knowledge that the N-terminal alpha-crystallin domain is required for CathB binding, I performed a multiple sequence alignment on the N-terminal domains of the ACD proteins in the ACD28.9 subclade (fig 4.11C). This alignment revealed that the ACD28.9 subclade show higher levels of sequence identity across N-terminal ACDs than is observed between the N- and C-terminal

domains in ACD28.9 (Figure 4.8D). The sequences have a mean sequence identity of 44.2%, with a standard deviation of ± 24.9%. The highest pairwise identity is between the N-terminal ACDs of ACD81.4 and ACD114.3, which have an 89.9% sequence identity, only differing due to an addition of an extra residue at the end of the ACD81.4 sequence. The N-terminal ACD of ACD28.9 is the most distinct within the clade, having a mean sequence identity of 27.9% (± 2.09%) with the other members of the clade.

Analysis of the conservation of N-terminal and C-terminal ACDs in the ACD28.9 subclade through pairwise comparisons of multiple sequence alignments revealed that C-terminal domains are less conserved than N-terminal domains, with a lower mean pairwise percentage identity (Figures 4.11A and 4.11B). The N-terminal domains contain regions of high sequence similarity across the domain, whereas C-terminal domains show higher degrees of divergence towards the C-terminus (Figures 4.11C and 4.11D).

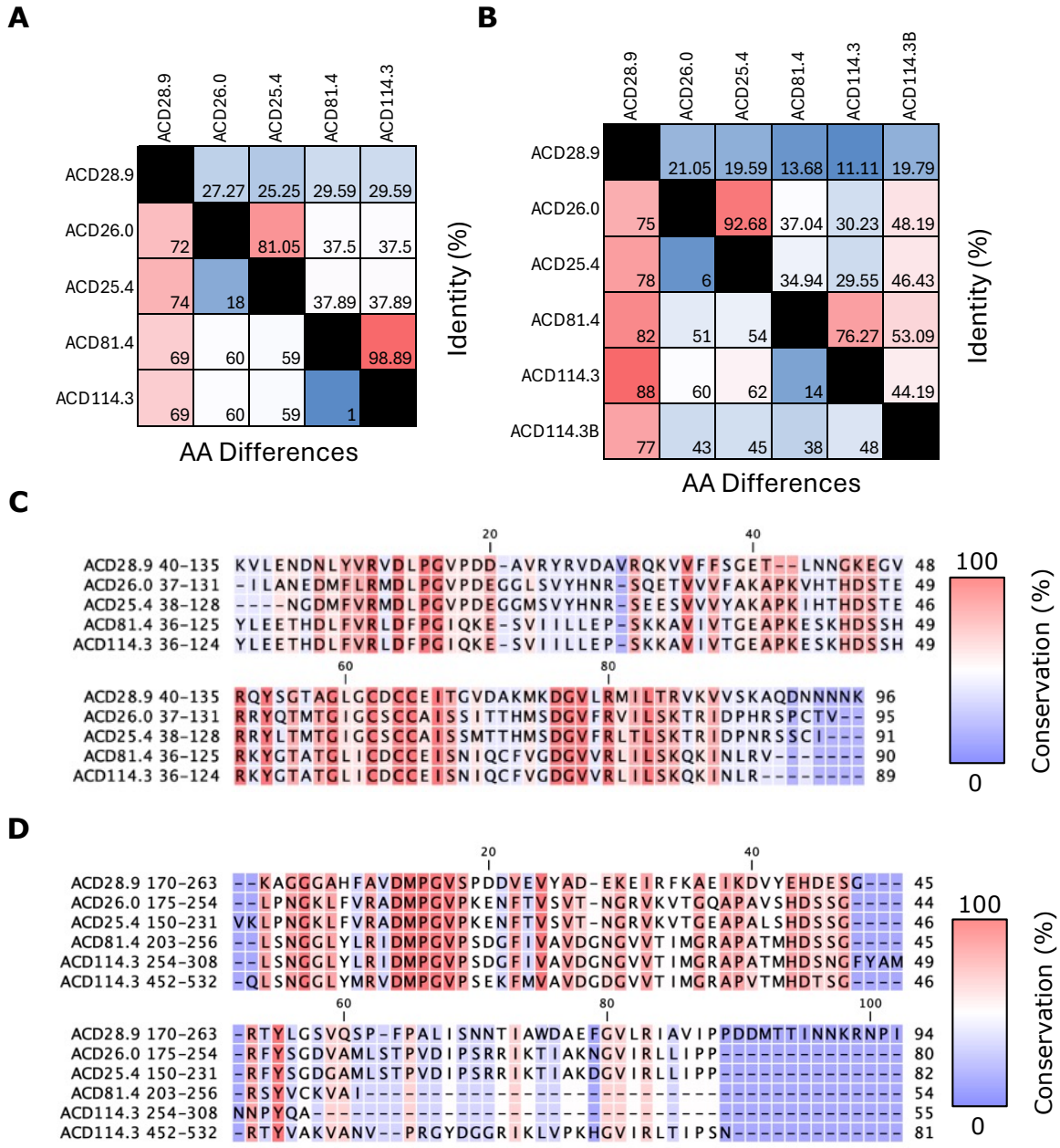


Figure 4.11: N-terminal ACDs in the ACD28.9 subclade proteins are more highly conserved than C-terminal ACDs.
A: Pairwise comparison of N-terminal ACD sequences from the ACD28.9 subclade. **B:** Pairwise comparison of C-terminal ACD sequences from the ACD28.9 subclade. **C:** Multiple sequence alignment of N-terminal ACDs from the ACD28.9 subclade. **D:** Multiple sequence alignment of C-terminal ACDs from the ACD28.9 subclade

4.4: Discussion

4.4.1: Plant sHSPs have an evolutionary history rich in lineage-specific expansion, including a Brassicales expanded clade containing 5 Arabidopsis ACD proteins.

This work presents the largest phylogeny of ACD proteins to date, covering sHSPs from 180 species from across the plant phylogeny. Based on the results of domain identification and phylogenetic analysis of ACD proteins in these species, it was possible to to identify a significant history of lineage-specific expansion across the evolutionary history of ACD proteins in plants. One such lineage-specific expansion is a Brassicales-expanded clade which contains the *A. thaliana* protein ACD28.9. Given that ACD28.9 is a known interactor of CathB proteins, and that CathB genes are upregulated by *M. persicae* when feeding on *Brassica rapa* (Mathers *et al.*, 2017), this expansion of ACD proteins in Brassicales may be driven by the pressure to efficiently bind to the distinct members of the CathB family. Expansions of ACD proteins in other orders of plants include those in Fabales, Poales, and Solanales, all of which are susceptible to infestation by *M. persicae*, as well as being hosts to a number of specialist aphids. One such interaction is between *A. pisum* and legume crops, and may have driven the expansion of a separate clade of the CathB gene family in *A. pisum* (Rispe *et al.*, 2008).

4.4.2: ACD28.9 and SLI1 belong to separate clades of Arabidopsis ACD proteins.

ACD28.9 has been shown to interact with *M. persicae* CathB proteins (Liu *et al.*, 2025), whereas SLI1, though involved in plant resistance to *M. persicae* and other herbivores (Kloth *et al.*, 2017, Kloth and Kormelink, 2020, Kloth *et al.*, 2021), does not appear to bind to any members of the *M. persicae* CathB protein family. Sequence alignment and phylogenetic analysis revealed that the two proteins show low sequence identity, and belong to distant branches of the ACD protein family in *A. thaliana*. Sequence alignment of the N-terminal alpha-crystallin domains of the two proteins, which is the region responsible for CathB binding in ACD28.9, shows that the two sequences have a low sequence identity. This finding offers insight into the reason for the differential CathB binding potential between the two proteins.

4.4.3: ACD28.9 subclade proteins contain multiple ACDs, which likely arose prior to the divergence of the clade.

The ACD proteins in the expanded ACD28.9 clade all contain at least two ACDs, which, based on phylogenetic analysis, appear to have arisen prior to the expansion of the ACD28.9 clade, potentially as a result of gene duplication. Based on high confidence protein structure predictions of ACD28.9 and ACD25.4, each ACD forms a separate β-sandwich structure, with the two being connected by a linker sequence. This is in keeping with previous literature which mark β-sandwich structures of being characteristic of ACD (Basha *et al.*, 2012). The N-terminal ACD of ACD28.9 appears to be responsible for binding CathB proteins, and though this domain is present throughout the ACD28.9 clade of *A. thaliana* ACD proteins, the N-terminal ACD in ACD28.9 displays the lowest identity when compared with the N-terminal ACDs of the other members of the ACD28.9 clade. This difference in functionality may explain the heightened level of conservation between the N-terminal domains of the ACD28.9 subclade, as this is the domain which appears to be involved in ACD protein dimerisation, as well as the recognition of CathB proteins.

4.4.4: ACD28.9 subclade proteins are capable of binding CathB proteins in yeast.

Based on yeast-2-hybrid results, the other members of the ACD28.9 clade that were tested were all capable of binding to CathB6 and CathB12. While there is no current experimental evidence that the N-terminal ACDs of the other proteins are involved in CathB binding in the same way as the ACD28.9 N-terminal ACD, given that the N-terminal ACDs of these proteins are relatively highly conserved, it is likely that they bind to CathB proteins at a similar interface. The small number of highly conserved residues may give an insight into potential CathB-binding regions of these ACD-containing proteins.

4.4.5: Advances and future potential.

The above evidence points to an evolutionary arms race between *A. thaliana* and *M. persicae*, in which the two associated protein families evolve dynamically in response to the evolution of the opposing side. While the CathB proteins evolve to avoid detection by plant ACD proteins, the plant ACD proteins evolve to ensure that CathB detection and neutralisation is still maintained. Using this newly

generated plant ACD protein phylogeny, it is now possible to identify homologs of CathB-binding ACD proteins in other species, which could be used to identify the residues in ACD28.9 which drive interaction with CathB proteins. This could also give rise to novel insight into the mechanisms of the extreme generalism observed in *M. persicae*.

Though predictions like the above are valuable in determining the structure of sHSP monomers, there is plenty of evidence that ACD-containing proteins exist natively as oligomers. For example, an X-ray structure of HSP16.9 from *Triticum aestivum* shows that the protein assembles to form an oligomer containing 12 monomeric subunits (van Montfort *et al.*, 2001), with evidence that other ACD-containing proteins from *T. aestivum, A. thaliana*, and *Pisum sativum* also form dodecamers (Basha *et al.*, 2010). While there is now evidence from yeast-2-hybrid assays that ACD28.9 is capable of assembling to form oligomers, the ultimate number of subunits in an ACD28.9 oligomer is still currently unknown, and an avenue for future research.

Chapter 5 – General Discussion

5.1: Research context

Aphids like *Myzus persicae* cause substantial damage to a broad range of plant species, including several of significant agricultural importance, leading to marked economic losses. These losses arise as a result of direct damage from the feeding aphid itself, as well as indirect damage from the impacts of aphid-vectored plant pathogens. *Myzus persicae* is particularly destructive due to its extensive polyphagy, with the potential to colonise plant species from over 40 families spanning the phylogeny of angiosperms, and vectoring over 100 plant viruses (Ramsey *et al.*, 2007).

Plants must respond to this threat from aphids and their associated pathogens using a complex immune response beginning with the perception of conserved molecular patterns associated with invasion using cell-surface bound pattern recognition receptors (Jaouannet *et al.*, 2014, Hogenhout and Bos, 2011). The downstream processes of this perception is termed pattern triggered immunity (PTI), and is typically sufficient to confer resistance to non-adapted pathogens (Jones and Dangl, 2006).

To successfully overcome the resistance conferred by PTI, aphids like *M. persicae* must deploy a repertoire of effectors that modulate this response to allow for successful colonisation in a process termed effector triggered susceptibility (ETS) (van Bel and Will, 2016, Hogenhout and Bos, 2011). In recent years, using a combination of genomic and molecular approaches, an increasing number of aphid effectors have been identified and functionally characterised, with wide ranging evolutionary and functional characteristics. In *M. persicae* alone, effectors have been identified which are ancient highly conserved singletons like Mp10 (Bos *et al.*, 2010, Rodriguez *et al.*, 2014, Deshoux *et al.*, 2022, Gravino *et al.*, 2025), members of protein families like the Cathepsin B (Mathers *et al.*, 2017, Guo *et al.*, 2020, Liu *et al.*, 2025), as well as members of lncRNA gene families (Chen *et al.*, 2020). This broad spectrum of virulence factors perhaps speaks to the unparalleled generalism displayed by *M. persicae*.

In spite of the increasing level of understanding of the identity and functions of *M. persicae* effectors, their physiological origins have largely been assumed based on a dogma of salivary gland synthesis and salivary secretion. Additionally, the

methods by which plant hosts respond to *M. persicae* effectors in order to diminish their deleterious effects are still largely uncharted. In this work, I have sought to address both of these lines of investigation, to further understand the complex transcriptional dynamics of effectors deployed by *M. persicae*, as well as the myriad effects that these effectors have *in planta*.

5.2: Summary of results

5.2.1: Organ-specific RNA-seq identified contributions of the foregut to M. persicae oral secretions.

Transcriptomic resources for *M. persicae*, though extensive when compared to that of other aphid species, lacked the resolution to determine the physiological origins of saliva, and so saliva-derived effectors were traditionally thought to be synthesised predominantly in the salivary glands. As such, in Chapter 2, I generated an RNA-seq dataset to examine gene expression across seven *M. persicae* organs. In tandem with mass spectrometry of aphid saliva, I identified foregut-specific expression of a group of Cathepsin B (CathB) proteins found in the saliva. Gene ontology analysis of organ-specific libraries and the salivary proteome showed an overlap in enriched gene ontologies between the foregut and orally secreted proteins, suggesting that oral secretions are comprised of salivary and foregut-derived elements. This goes against the central dogma in which orally secreted proteins are synthesised exclusively in the salivary glands, instead favouring a hypothesis in which oral secretions are at least partially comprised of aphid regurgitant.

Though these results have been accommodated by the increased transcriptomic resolution provided by organ-specific RNA-seq of adult aphids, the resolution remains limited by the physical limitations of the person dissecting and isolating tissues. For example, it is known that the salivary glands are comprised of two morphologically distinct groups of cells, the primary and accessory salivary glands (Mutti *et al.*, 2007). While conducting the dissections for this study, I was unable to separate these two structures while maintaining their structural integrity. As such, it is likely that there is still a degree of heterogeneity within the still composite structures of the extracted organs sequenced for the generation of these libraries.

In light of advances in single-cell RNA-sequencing (Shapiro *et al.*, 2013, Sun *et al.*, 2023, Wang *et al.*, 2024, McLaughlin *et al.*, 2022), it is possible that such an approach would allow for further disentanglement of effector expression in the aphid salivary glands and other composite structures. However, there are several factors on account of which single-nuclei sequencing would not be suitable for *Myzus persicae* with the currently available resources. Firstly, protocols described so far are still largely based on single nuclei sequencing, which would miss, for instance, RNAs present in p-bodies, which are also described in insects, such as in the intestines (Buddika *et al.*, 2022). Moreover, analysis of these large datasets typically relies on cell- or tissue-specific genetic markers for determination of cell or tissue type, the likes of which currently do not exist for the majority of *Myzus persicae* tissues. Obtaining tissue-specific markers necessitates identifying genes predominantly expressed in an organ/tissue relatively to other organs, a process that requires isolation of specific organs followed by RNA-seq. Another drawback of single-nuclei transcriptomics is its trade-off between sequencing breadth and sequencing depth. A larger number of nuclei necessitates shallower sequencing, potentially overlooking less highly expressed genes within each nucleus—an oversight which would be particularly deleterious for purposes such as the tracking or identification of new candidate effectors, which may not be the most highly expressed genes in a given tissue. Reducing the breadth of the sequencing to focus on read depth, a process which involves restricting the number of nuclei for which libraries are created, means that it is likely that rare cell types or cells derived from small organs, such as the salivary glands, may be excluded from the datasets during cell sorting.

Most RNA-seq data generated so far are derived from wingless, asexually reproducing females, which are the most abundant morph during infestation. As such, this morph is the most tractable for analysis which requires large numbers of individuals. It has also allowed for comparability between transcriptomic data and other generated data such as proteomics (Liu *et al.*, 2024). However, there are a number of alternative morphs which serve different functions in the maintenance and dispersal of aphid colonies. For example, winged morphs encounter a significantly wider range of host plants in the process of selecting suitable hosts, and as such it could be hypothesised that winged aphids express different effector genes to wingless individuals, which are typically restricted to a

single host plant species. There is evidence for morph-specific gene expression in aphids (Gu *et al.*, 2013), though studies on this topic have primarily been restricted to chemosensory genes, and so further study needs to be undertaken to examine the extent to which the effectorome of winged individuals varies from that of wingless individuals. This could be achieved leveraging the methods described in Chapter 2 could be performed on other stages of the *M. persicae* life history, providing further insight into this morph-specific effectorome diversity. The data generated in this study provide valuable resources to advance future research, including candidate markers for single-cell transcriptomics and a reference dataset for comparative analyses among *M. persicae* morphs.

Analyses of the transcriptomic and proteomic data conducted herein utilised the *M. persicae* Clone O v2.1 genome annotation (Mathers *et al.*, 2017), which, though high quality, is not totally complete. Analyses were also conducted using the annotated chromosome scale assembly (Liu *et al.*, 2024), which is of a very high standard at the time of these analyses. However, the generation of a *M. persicae* Clone O v2.1 telomere-to-telomere assembly is on the way and analysing the RNA-seq data generated herein against this assembly may identify more candidate effector genes.

5.2.2: M. persicae CathB proteins have diversified to avoid detection by ACD28.9.

Recent research has highlighted the function of one orally secreted and foregut-enriched CathB protein as an effector targeting plant immune regulator EDS1 (Liu *et al.*, 2025). This mode of action is suppressed in the presence of the *A. thaliana* alpha-crytallin domain (ACD) containing protein ACD28.9. Based on these findings, Chapter 3 explored the ways in which additional members of the *M. persicae* CathB family interact with ACD28.9. A series of yeast-2-hybrid and *in planta* assays revealed that while many members of the CathB family bind ACD28.9 in yeast, there are a number of proteins in the family which do not bind ACD28.9. The differential binding between CathB family members and ACD28.9 correspond to differences in the subcellular localisations of CathB proteins in the presence of ACD28.9 within *Nicotiana benthamiana* plant cells. Of particular interest, proteins within the recently expanded clade of the *M. persicae* CathB family show varying ACD28.9-binding capacity. This points to an intriguing

possibility that requires further investigation is that CathB members in the recently expanded clade have evolved to avoid interactions with plant ACD28.9.

Comparisons of CathB6 and CathB12 binding to ACD28.9 via yeast-2-hybrid and *in planta* assays suggest that CathB6 binding to ACD28.9 is dependent on two amino acids. It was previously shown that ACD28.9 binds to a region of CathB6 ranging from amino acids 263 to 333 (Liu *et al.*, 2025). Within this region, CathB6 and CathB12 differ by only two amino acids, and exchanging these two amino acids between CathB6 and CathB12 by creating single and double mutants revealed that these amino acids are required for the rescue of CathB6 from p-bodies by ACD28.9 (Figure 3.10B). The CathB12 mutant that carries the two CathB6 amino acids showed a gain-of-function phenotype, as the CathB12 mutant is now removed from the p-bodies in the presence of ACD28.9 (Figure 3.10C). Given that CathB6 and CathB12 are the closest paralogs within the recently expanded clade of CathB gene family, these data point to a recent evolutionary arms race between these CathB family members and ACD28.9 that involves gene duplication of CathB genes, followed by the fixation of specific mutations that leads to the prevention of CathB binding to ACD28.9.

The yeast 2-hybrid partially align with the cell biology data. Whereas CathB12 shows lower binding affinity to ACD28.9 compared to CathB6 in the yeast two-hybrid assays (Figure 3.6), consistent with the cell biology data, the CathB6 and CathB12 single and double amino acid mutants did not bind ACD28.9 at all in these assays. The reason of this unexpected behaviour of the mutants is unclear. It is possible that this experiment requires some additional optimisation to assure proper translation and folding of generated mutant proteins in the yeast system.

Given the multiple instances of closest CathB paralogs exhibiting different capacity to bind ACD28.9 (CathB6 and CathB12, CathB14 and CathB15), it should also be possible to perform analysis to identify signs of positive selection in the gene sequences of CathB genes by examining the ratio of synonymous and non-synonymous mutations between these pairs of genes. Given current evidence, I would hypothesise that there would be signs of positive selection for codons encoding residues homologous to CathB6 307A and CathB6 314F, which are the two residues responsible for ACD28.9 binding in CathB6. Intriguingly, there is evidence for positive selection in CathB genes in the pea aphid (Rispe *et al.*, 2008).

CathB gene expansions in pea aphids involve a different branch in the CathB gene family, as the expanded pea aphid CathB genes are more similar to CathB1, B2 and B3 in *M. persicae* (see phylogeny in Figure 2.10). Curiously, these CathB genes do not appear to be expressed to a significant degree in *M. persicae* (Figure 2.10).

5.2.3: A. thaliana ACD proteins in a Brassicales-expanded clade are capable of binding CathB.

ACD28.9 also belongs to an expansive family of ACD proteins in plants. To assess whether additional members of this family play a role in CathB-mediated plant immunity, in Chapter 4 I generated a phylogeny of ACD proteins from 180 plant species using a database of plant proteomes and an HMM for the characteristic ACD. Interestingly, ACD28.9 and four other *A. thaliana* proteins belong to a subclade of plant ACD proteins which has undergone expansion in the Brassicales. These proteins contain multiple ACDs which appear to have arisen through intragenic duplication prior to the expansion of the subclade, and also interact with CathB6 in yeast, though none as strongly as ACD28.9. In ACD28.9, only the N-terminal domain appears to be involved in the binding of CathB6, as well as ACD28.9 dimerisation. It remains to be investigated whether these other ACD proteins also rescue CathB6 or other *M. persicae* CathB proteins, from plant p-bodies.

5.3: Focus for future investigations

5.3.1: How are foregut-borne proteins translocated into plants?

The conclusion drawn in Chapter 2 that the foregut contributes to aphid oral secretions is based on elevated expression of genes encoding multiple orally secreted proteins in the foregut compared to other tissues. While this is strong evidence that orally secreted proteins are synthesised in the foregut, there is still currently little known about the mechanisms by which the contents of the foregut reach the plant cell.

There is, however, precedent for the deposition of foregut derived molecules in oral secretions of aphids. Viruses that adhere to the foregut of aphids, such as *Cauliflower mosaic virus* (CaMV) and closteroviruses, are readily transmitted by

aphids (fig. 2.11, reviewed in Ng and Falk, 2006), indicating that aphids can also introduce proteins from the foregut into plant cells. Curiously, soldier aphids from of the social aphid species *Tuberaphis styraci* use CathB proteins as toxins to kill enemies – these soldier aphids introduce these toxins via their stylets into the bodies of their enemies via their stylets (Kutsukake *et al.*, 2004, Kutsukake *et al.*, 2008).

5.3.2: What is the composition of the p-body and what is its role in plant immunity?

Experiments in Chapter 3 revealed that diverse CathB proteins localize within p-bodies of *N. benthamiana* cells in transient expression assays. P-bodies are membrane-less condensates of protein complexes and plant mRNAs (Xu and Chua, 2011). It is unclear which protein(s) or RNAs may recruit CathB to the p-bodies. Further analyses of the proximity interactome of CathB (Lui *et al.*, 2024b; Liu *et al.*, 2025) may reveal which A. thaliana proteins recruit CathB, and EDS1 plus partners, to p-bodies. P-bodies are dynamic complexes that move within the cytoplasm of plant cells, as can be seen in CathB-associated p-bodies (Liu et al, 2025). These bodies assemble or disassemble based on environmental conditions, such as the level of light exposure or time of day (Jang *et al.*, 2019). As such, CathB proteins may be temporary passengers of p-bodies.

Recent work has detailed the effects of effectors from other plant pathogens on the dynamics of plant p-bodies. Effectors from *Pseudomonas syringae* have been shown to induce the formation of p-bodies in infected plant cells (González-Fuente *et al.*, 2025). This includes the effector HopN1, which is a cysteine protease (Rodríguez-Herva *et al.*, 2012) like the *M. persicae* CathB effectors described in this work. However, whereas cysteine protease activity of CathB proteins is not required for p-body localization and promotion of aphid fecundity (Liu *et al.*, 2025), the formation of p-bodies by HopN1 appears to be reliant on the enzymatic activity of the cysteine protease. Another cysteine protease effector from *Phytophthora parasitica* relies on its enzymatic activity to negatively regulate host immunity (Dong *et al.*, 2025), but there is no current evidence that this process involves the modulation of p-body formation.

5.3.3: What is the native structure of ACD28.9, and does this change in response to CathB?

Several plant proteins containing ACDs have been shown to form high order oligomers in order to fulfil their role as chaperone proteins (van Montfort *et al.*, 2001). Given that the binding assays displayed in Chapters 3 and 4 show evidence homodimerization of ACD28.9 (Figure 3.6), it is possible that ACD28.9 also exists as oligomers in the plant cell. There is evidence that the number of subunits in an ACD protein oligomer is dynamic, with oligomers diverging from standard structure in response to stress (Stengel *et al.*, 2010). As such, it would be valuable to decipher the native structure of ACD28.9 oligomers, both in the presence and absence of *M. persicae* CathB effectors.

To accomplish this, tandem mass spectrometry, an approach which has previously highlighted the structural plasticity of PsHSP18.1 in response to heat stress (Stengel *et al.*, 2010), could be performed on ACD28.9 and CathB mixtures, to elucidate the number of subunits in ACD28.9-CathB complexes. This knowledge would be valuable for leveraging computational modelling of these interactions, which are currently unable to confidently predict the structure of the ACD28.9-CathB interaction. In turn, these predictions could be used to identify ACD28.9 residues responsible for binding CathB proteins, and potentially other CathB amino acids that bind ACD28.9, as well as subsequent experiments that involve the generation of relevant binding or non-binding mutants.

5.3.4: How can these findings be used to generate aphid-resistant crops?

ACD28.9 appears to play a role in plant immunity to aphids, with *A. thaliana* ACD28.9 mutants being less resistant to aphids. Additionally, CathB6 and its interactors EDS1, PAD4 and ADR1, are removed from p-bodies in the presence of ACD28.9 (Lui *et al.*, 2025). However, how ACD28.9 contributes to plant immunity to aphids is still largely unclear. An approach to investigate this further is to better understand CathB interactions with other ACD proteins, specifically those that belong to the same expanded clade as ACD28.9 of the Brassicales. The work presented in Chapter 4 provides a number of ACD28.9 homologs in closely related Brassica plant species (fig. 4.6), which could be tested for CathB binding affinity. Analysis of CathB-binding and non-binding ACD28.9 homologs could identify

residues involved in binding, and possibly the identification of more ACDs that bind aphid CathB proteins, thereby increasing plant resistance to aphids.

5.4: Conclusions

Taken together, the work described in this thesis provides the first steps in disentangling what appears to be a complex evolutionary arms race between a family of aphid effectors and a family of plant immune proteins. Specifically, the work provides data to further study how foregut-borne orally secreted CathB members of a *Myzus*-lineage expanded clade that upregulate in the aphids on Brassicaceous plants (Mathers *et al.*, 2017) have evolved to avoid detection and neutralisation by a plant ACD protein that also belongs to an expanded protein family in the Brassicales.

Using the data generated and progress made during this studentship, future investigations should be able to more comprehensively understand the mechanisms by which *M. persicae* exploits plant immune responses to facilitate successful colonisation of plant hosts. This data also provides a framework for exploring the role of diverse ACD proteins in plant responses to aphids, and particularly how CathB effectors play a role in this. A goal may be to identify ACD proteins that bind more aphid CathB protein effectors thereby increasing plant resistance to *M. persicae*, and possibly other aphid species.

Bibliography

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Abril, J. F., Agbayani, A., An, H. J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., De Pablos, B., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M. H., Ibegwam, C., *et al.* 2000. The genome sequence of *Drosophila melanogaster*. *Science,* 287**,** 2185-95.

Akman Gündüz, E. & Douglas, A. E. 2009. Symbiotic bacteria enable insect to use a nutritionally inadequate diet. *Proc Biol Sci,* 276**,** 987-91.

Atamian, H. S., Chaudhary, R., Cin, V. D., Bao, E., Girke, T. & Kaloshian, I. 2013. In Planta Expression or Delivery of Potato Aphid *Macrosiphum euphorbiae* Effectors Me10 and Me23 Enhances Aphid Fecundity. *Molecular Plant-Microbe Interactions®,* 26**,** 67-74.

Backus, E. A., Serrano, M. S. & Ranger, C. M. 2005. Mechanisms of hopperburn: An overview of insect taxonomy, behavior, and physiology. *Annual Review of Entomology,* 50**,** 125-151.

Bai, X., Correa, V. R., Toruño, T. Y., Ammar, E.-D., Kamoun, S. & Hogenhout, S. A. 2009. AY-WB Phytoplasma Secretes a Protein That Targets Plant Cell Nuclei. *Molecular Plant-Microbe Interactions®,* 22**,** 18-30.

Basha, E., Jones, C., Wysocki, V. & Vierling, E. 2010. Mechanistic Differences between Two Conserved Classes of Small Heat Shock Proteins Found in the Plant Cytosol. *Journal of Biological Chemistry,* 285**,** 11489-11497.

Basha, E., O'neill, H. & Vierling, E. 2012. Small heat shock proteins and α-crystallins: dynamic proteins with flexible functions. *Trends Biochem Sci,* 37**,** 106-17.

Baumann, P., Baumann, L., Clark, M. A. & Thao, M. L. 1998. *Buchnera aphidicola*: The endosymbiont of aphids. *ASM News,* 64**,** 203-209.

Benedetti, M., Pontiggia, D., Raggi, S., Cheng, Z., Scaloni, F., Ferrari, S., Ausubel, F. M., Cervone, F. & De Lorenzo, G. 2015. Plant immunity triggered by engineered in vivo release of oligogalacturonides, damage-associated molecular patterns. *Proceedings of the National Academy of Sciences,* 112**,** 5533-5538.

Bentham, A. R., De La Concepcion, J. C., Mukhi, N., Zdrzałek, R., Draeger, M., Gorenkin, D., Hughes, R. K. & Banfield, M. J. 2020. A molecular roadmap to the plant immune system. *Journal of Biological Chemistry,* 295**,** 14916-14935.

Biello, R., Singh, A., Godfrey, C. J., Fernandez, F. F., Mugford, S. T., Powell, G., Hogenhout, S. A. & Mathers, T. C. 2021. A chromosome-level genome assembly of the woolly apple aphid, *Eriosoma lanigerum Hausmann* (Hemiptera: Aphididae). *Mol Ecol Resour,* 21**,** 316-326.

Boissot, N., Schoeny, A. & Vanlerberghe-Masutti, F. 2016. Vat, an Amazing Gene Conferring Resistance to Aphids and Viruses They Carry: From Molecular Structure to Field Effects. *Front Plant Sci,* 7**,** 1420.

Boone, B. A., Ichino, L., Wang, S., Gardiner, J., Yun, J., Jami-Alahmadi, Y., Sha, J., Mendoza, C. P., Steelman, B. J., Van Aardenne, A., Kira-Lucas, S., Trentchev, I., Wohlschlegel, J. A. & Jacobsen, S. E. 2023. ACD15, ACD21, and SLN regulate the accumulation and mobility of MBD6 to silence genes and transposable elements. *Science Advances,* 9**,** eadi9036.

Boone, B. A., Mendoza, C. P., Behrendt, N. J. & Jacobsen, S. E. 2024. α-Crystalline Domains and Intrinsically Disordered Regions Can Work in Parallel to Induce Accumulation of MBD6 at Chromocenters in *Arabidopsis thaliana*. *Epigenomes,* 8**,** 33.

Bos, J. I., Prince, D., Pitino, M., Maffei, M. E., Win, J. & Hogenhout, S. A. 2010. A functional genomics approach identifies candidate effectors from the aphid species *Myzus persicae* (green peach aphid). *PLoS Genet,* 6**,** e1001216.

Bourgine, B. & Guihur, A. 2021. Heat Shock Signaling in Land Plants: From Plasma Membrane Sensing to the Transcription of Small Heat Shock Proteins. *Frontiers in Plant Science*, Volume 12 - 2021.

Breeze, T. D., Bailey, A. P., Balcombe, K. G. & Potts, S. G. 2011. Pollination services in the UK: How important are honeybees? *Agriculture, Ecosystems & Environment,* 142**,** 137-143.

Buddika, K., Huang, Y. T., Ariyapala, I. S., Butrum-Griffith, A., Norrell, S. A., O'connor, A. M., Patel, V. K., Rector, S. A., Slovan, M., Sokolowski, M., Kato, Y., Nakamura, A. & Sokol, N. S. 2022. Coordinated repression of pro-

differentiation genes via P-bodies and transcription maintains Drosophila intestinal stem cell identity. *Curr Biol,* 32**,** 386-397.e6.

Campos, M. L., Kang, J. H. & Howe, G. A. 2014. Jasmonate-triggered plant immunity. *J Chem Ecol,* 40**,** 657-75.

Carolan, J. C., Fitzroy, C. I., Ashton, P. D., Douglas, A. E. & Wilkinson, T. L. 2009. The secreted salivary proteome of the pea aphid *Acyrthosiphon pisum* characterised by mass spectrometry. *Proteomics,* 9**,** 2457-67.

Caspers, G.-J., Leunissen, J. a. M. & De Jong, W. W. 1995. The expanding small heat-shock protein family, and structure predictions of the conserved "α-crystallin domain". *Journal of Molecular Evolution,* 40**,** 238-248.

Chaudhary, R., Atamian, H. S., Shen, Z., Briggs, S. P. & Kaloshian, I. 2014. GroEL from the endosymbiont *Buchnera aphidicola* betrays the aphid by triggering plant defense. *Proc Natl Acad Sci U S A,* 111**,** 8919-24.

Chaudhary, R., Atamian, H. S., Shen, Z. X., Briggs, S. P. & Kaloshian, I. 2015. Potato Aphid Salivary Proteome: Enhanced Salivation Using Resorcinol and Identification of Aphid Phosphoproteins. *Journal of Proteome Research,* 14**,** 1762-1778.

Chen, Y., Singh, A., Kaithakottil, G. G., Mathers, T. C., Gravino, M., Mugford, S. T., Van Oosterhout, C., Swarbreck, D. & Hogenhout, S. A. 2020. An aphid RNA transcript migrates systemically within plants and is a virulence factor. *Proceedings of the National Academy of Sciences,* 117**,** 12763-12771.

Cheng, X., Zhu, L. & He, G. 2013. Towards Understanding of Molecular Interactions between Rice and the Brown Planthopper. *Molecular Plant,* 6**,** 621-634.

Chinchilla, D., Bauer, Z., Regenass, M., Boller, T. & Felix, G. 2006. The Arabidopsis receptor kinase FLS2 binds flg22 and determines the specificity of flagellin perception. *Plant Cell,* 18**,** 465-76.

Decroocq, V., Sicard, O., Alamillo, J. M., Lansac, M., Eyquard, J. P., García, J. A., Candresse, T., Le Gall, O. & Revers, F. 2006. Multiple Resistance Traits Control Plum pox virus Infection in *Arabidopsis thaliana*. *Molecular Plant-Microbe Interactions®,* 19**,** 541-549.

Deshoux, M., Monsion, B., Pichon, E., Jimenez, J., Moreno, A., Cayrol, B., Thebaud, G., Mugford, S. T., Hogenhout, S. A., Blanc, S., Fereres, A. & Uzest, M. 2022. Role of Acrostyle Cuticular Proteins in the Retention of an Aphid Salivary Effector. *Int J Mol Sci,* 23**,** 15337.

Dogimont, C., Bendahmane, A., Chovelon, V. & Boissot, N. 2010. Host plant resistance to aphids in cultivated crops: genetic and molecular bases, and interactions with aphid populations. *C R Biol,* 333**,** 566-73.

Dong, J., Li, W., Yang, Y., Liu, S., Li, Y., Meng, Y. & Shan, W. 2025. The cysteine protease RD19C suppresses plant immunity to Phytophthora by modulating copper chaperone ATX1 stability. *The Plant Journal,* 122**,** e70120.

Dongus, J. A. & Parker, J. E. 2021. EDS1 signalling: At the nexus of intracellular and surface receptor immunity. *Current Opinion in Plant Biology,* 62**,** 102039.

Douglas, A. E. 2018. Strategies for Enhanced Crop Resistance to Insect Pests. *Annu Rev Plant Biol,* 69**,** 637-660.

Eddy, S. R. 2011. Accelerated Profile HMM Searches. *PLOS Computational Biology,* 7**,** e1002195.

Elzinga, D. A., De Vos, M. & Jander, G. 2014. Suppression of Plant Defenses by a *Myzus persicae* (Green Peach Aphid) Salivary Effector Protein. *Molecular Plant-Microbe Interactions®,* 27**,** 747-756.

Emmanuelle, J., Gwenaelle, G. & Armelle, C. 2010. Evolutionary lability of a complex life cycle in the aphid genus Brachycaudus. *BMC Evol Biol,* 10**,** 295.

Engel, M. S. 2015. Insect evolution. *Current Biology,* 25**,** R868-R872.

Escudero-Martinez, C., Rodriguez, P. A., Liu, S., Santos, P. A., Stephens, J. & Bos, J. I. B. 2020. An aphid effector promotes barley susceptibility through suppression of defence gene expression. *Journal of Experimental Botany,* 71**,** 2796-2807.

Eyres, I., Jaquiery, J., Sugio, A., Duvaux, L., Gharbi, K., Zhou, J. J., Legeai, F., Nelson, M., Simon, J. C., Smadja, C. M., Butlin, R. & Ferrari, J. 2016. Differential gene expression according to race and host plant in the pea aphid. *Mol Ecol,* 25**,** 4197-215.

Fereres, A. & Moreno, A. 2009. Behavioural aspects influencing plant virus transmission by homopteran insects. *Virus Res,* 141**,** 158-68.

Giolai, M. 2019. *Spatially resolved transcriptomics reveals plant host responses to the aphid pest Myzus persicae.* Doctoral, University of East Anglia.

González-Fuente, M., Schulz, N., Abdrakhmanov, A., Izzati, G., Zhu, S., Langin, G., Gouguet, P., Franz-Wachtel, M., Macek, B., Hafrén, A., Dagdas, Y. & Üstün, S. 2025. Bacteria use processing body condensates to attenuate host translation during infection. *bioRxiv***,** 2025.01.09.632196.

Gosset, V., Harmel, N., Göbel, C., Francis, F., Haubruge, E., Wathelet, J.-P., Du Jardin, P., Feussner, I. & Fauconnier, M.-L. 2009. Attacks by a piercing-sucking insect (*Myzus persicae* Sultzer) or a chewing insect (*Leptinotarsa decemlineata* Say) on potato plants (*Solanum tuberosum L.*) induce differential changes in volatile compound release and oxylipin synthesis. *Journal of Experimental Botany,* 60**,** 1231-1240.

Gravino, M., Locci, F., Tundo, S., Cervone, F., Savatin, D. V. & De Lorenzo, G. 2017. Immune responses induced by oligogalacturonides are differentially affected by AvrPto and loss of BAK1/BKK1 and PEPR1/PEPR2. *Mol Plant Pathol,* 18**,** 582-595.

Gravino, M., Mugford, S. T., Pontiggia, D., Joyce, J., Drurey, C., Prince, D. C., Cervone, F., De Lorenzo, G. & Hogenhout, S. A. 2025. Aphid effector Mp10 balances immune suppression and defence activation through EDS1-dependent modulation of plant DAMP responses. *New Phytol*.

Gu, S.-H., Wu, K.-M., Guo, Y.-Y., Field, L. M., Pickett, J. A., Zhang, Y.-J. & Zhou, J.-J. 2013. Identification and Expression Profiling of Odorant Binding Proteins and Chemosensory Proteins between Two Wingless Morphs and a Winged Morph of the Cotton Aphid *Aphis gossypii* Glover. *PLOS ONE,* 8**,** e73524.

Guo, H., Zhang, Y., Tong, J., Ge, P., Wang, Q., Zhao, Z., Zhu-Salzman, K., Hogenhout, S. A., Ge, F. & Sun, Y. 2020. An Aphid-Secreted Salivary Protease Activates Plant Defense in Phloem. *Current Biology,* 30**,** 4826-4836.e7.

Guy, E., Boulain, H., Aigu, Y., Le Pennec, C., Chawki, K., Morlière, S., Schädel, K., Kunert, G., Simon, J. C. & Sugio, A. 2016. Optimization of Agroinfiltration in *Pisum sativum* Provides a New Tool for Studying the Salivary Protein Functions in the Pea Aphid Complex. *Frontiers in Plant Science,* 7.

Hewer, A., Becker, A. & Van Bel, A. J. 2011. An aphid's Odyssey--the cortical quest for the vascular bundle. *J Exp Biol,* 214**,** 3868-79.

Hogenhout, S. A., Ammar El, D., Whitfield, A. E. & Redinbaugh, M. G. 2008. Insect vector interactions with persistently transmitted viruses. *Annu Rev Phytopathol,* 46**,** 327-59.

Hogenhout, S. A. & Bos, J. I. B. 2011. Effector proteins that modulate plant-insect interactions. *Current Opinion in Plant Biology,* 14**,** 422-428.

Hoshiba, H. & Sasaki, M. 2008. Perspectives of multi-modal contribution of honeybee resources to our life. *Entomological Research,* 38**,** S15-S21.

Howe, G. A. & Jander, G. 2008. Plant immunity to insect herbivores. *Annu Rev Plant Biol,* 59**,** 41-66.

Hu, L., Gui, W., Chen, B. & Chen, L. 2019. Transcriptome profiling of maternal stress-induced wing dimorphism in pea aphids. *Ecol Evol,* 9**,** 11848-11862.

Hu, Y., Zhang, T., Liu, Y., Li, Y., Wang, M., Zhu, B., Liao, D., Yun, T., Huang, W., Zhang, W. & Zhou, Y. 2021. Pumpkin (*Cucurbita moschata*) HSP20 Gene Family Identification and Expression Under Heat Stress. *Front Genet,* 12**,** 753953.

Huang, H. J., Lu, J. B., Li, Q., Bao, Y. Y. & Zhang, C. X. 2018. Combined transcriptomic/proteomic analysis of salivary gland and secreted saliva in three planthopper species. *J Proteomics,* 172**,** 25-35.

Huang, H. J., Ye, Z. X., Lu, G., Zhang, C. X., Chen, J. P. & Li, J. M. 2021. Identification of salivary proteins in the whitefly *Bemisia tabaci* by transcriptomic and LC-MS/MS analyses. *Insect Sci,* 28**,** 1369-1381.

Huang, J., Hai, Z., Wang, R., Yu, Y., Chen, X., Liang, W. & Wang, H. 2022. Genome-wide analysis of HSP20 gene family and expression patterns under heat stress in cucumber (*Cucumis sativus L.*). *Frontiers in Plant Science,* Volume 13 - 2022.

Huang, Z., Xu, Z., Liu, X., Chen, G., Hu, C., Chen, M. & Liu, Y. 2024. Exploring the Role of the Processing Body in Plant Abiotic Stress Response. *Current Issues in Molecular Biology,* 46**,** 9844-9855.

Ingolia, T. D. & Craig, E. A. 1982. Four small Drosophila heat shock proteins are related to each other and to mammalian alpha-crystallin. *Proc Natl Acad Sci U S A,* 79**,** 2360-4.

Ishikawa, A., Ogawa, K., Gotoh, H., Walsh, T. K., Tagu, D., Brisson, J. A., Rispe, C., Jaubert-Possamai, S., Kanbe, T., Tsubota, T., Shiotsuki, T. & Miura, T. 2012. Juvenile hormone titre and related gene expression during the change of reproductive modes in the pea aphid. *Insect Mol Biol,* 21**,** 49-60.

Jang, G.-J., Yang, J.-Y., Hsieh, H.-L. & Wu, S.-H. 2019. Processing bodies control the selective translation for optimal development of Arabidopsis young seedlings. *Proceedings of the National Academy of Sciences,* 116**,** 6451-6456.

Jaouannet, M., Rodriguez, P. A., Thorpe, P., Lenoir, C. J. G., Macleod, R., Escudero-Martinez, C. & Bos, J. I. B. 2014. Plant immunity in plant–aphid interactions. *Frontiers in Plant Science,* 5.

Ji, R., Fu, J., Shi, Y., Li, J., Jing, M., Wang, L., Yang, S., Tian, T., Wang, L., Ju, J., Guo, H., Liu, B., Dou, D., Hoffmann, A. A., Zhu-Salzman, K. & Fang, J. 2021. Vitellogenin from planthopper oral secretion acts as a novel effector to impair plant defenses. *New Phytologist,* 232**,** 802-817.

Smith, C.M. (ed.) (2005) 'Inheritance of Arthropod Resistance', in Plant Resistance to Arthropods: Molecular and Conventional Approaches. Dordrecht: Springer Netherlands, pp. 219–267. Available at: https://doi.org/10.1007/1-4020-3702-3_8.

Jones, J. D. & Dangl, J. L. 2006. The plant immune system. *Nature,* 444**,** 323-9.

Jones, J. D. G., Staskawicz, B. J. & Dangl, J. L. 2024. The plant immune system: From discovery to deployment. *Cell,* 187**,** 2095-2116.

Kaloshian, I. & Walling, L. L. 2016. Hemipteran and dipteran pests: Effectors and plant host immune regulators. *Journal of Integrative Plant Biology,* 58**,** 350-361.

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**,** 3059-66.

Kearly, A., Nelson, A. D. L., Skirycz, A. & Chodasiewicz, M. 2024. Composition and function of stress granules and P-bodies in plants. *Seminars in Cell & Developmental Biology,* 156**,** 167-175.

Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology,* 37**,** 907-+.

Kim, G., Lee, S., Levy Karin, E., Kim, H., Moriwaki, Y., Ovchinnikov, S., Steinegger, M. & Mirdita, M. 2025. Easy and accurate protein structure prediction using ColabFold. *Nature Protocols,* 20**,** 620-642.

Kloth, K. J., Busscher-Lange, J., Wiegers, G. L., Kruijer, W., Buijs, G., Meyer, R. C., Albrectsen, B. R., Bouwmeester, H. J., Dicke, M. & Jongsma, M. A. 2017. SIEVE ELEMENT-LINING CHAPERONE1 Restricts Aphid Feeding on Arabidopsis during Heat Stress. *Plant Cell,* 29**,** 2450-2464.

Kloth, K. J. & Kormelink, R. 2020. Defenses against Virus and Vector: A Phloem-Biological Perspective on RTM- and SLI1-Mediated Resistance to Potyviruses and Aphids. *Viruses,* 12**,** 129.

Kloth, K. J., Shah, P., Broekgaarden, C., Ström, C., Albrectsen, B. R. & Dicke, M. 2021. SLI1 confers broad-spectrum resistance to phloem-feeding insects. *Plant Cell Environ,* 44**,** 2765-2776.

Kolberg, L., Raudvere, U., Kuzmin, I., Adler, P., Vilo, J. & Peterson, H. 2023. g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Research,* 51**,** W207-W212.

Kolde, R. 2025. *pheatmap: Pretty Heatmaps* [Online]. Available: https://github.com/raivokolde/pheatmap.

Kutsukake, M., Nikoh, N., Shibao, H., Rispe, C., Simon, J.-C. & Fukatsu, T. 2008. Evolution of Soldier-Specific Venomous Protease in Social Aphids. *Molecular Biology and Evolution,* 25**,** 2627-2641.

Kutsukake, M., Shibao, H., Nikoh, N., Morioka, M., Tamura, T., Hoshino, T., Ohgiya, S. & Fukatsu, T. 2004. Venomous protease of aphid soldier for colony defense. *Proceedings of the National Academy of Sciences,* 101**,** 11338-11343.

Lehnert, M. S., Myers, K. O. & Reiter, K. E. 2025. The Right Tool for the Job: A Review of Insect Mouthparts as a Tool Kit for Biomimetic Studies. *Biomimetics (Basel),* 10.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Proc, G. P. D. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-2079.

Lian, X., Wang, Q., Li, T., Gao, H., Li, H., Zheng, X., Wang, X., Zhang, H., Cheng, J., Wang, W., Ye, X., Li, J., Tan, B. & Feng, J. 2022. Phylogenetic and Transcriptional Analyses of the HSP20 Gene Family in Peach Revealed That PpHSP20-32 Is Involved in Plant Height and Heat Tolerance. *Int J Mol Sci,* 23.

Liu, Q., Goldberg, J. K., Mugford, S. T., Saalbach, G., Martins, C., Singh, A., Kaithakottil, G. G., Swarbreck, D. & Hogenhout, S. A. 2024. The salivary proteome of the green peach aphid/peach-potato aphid (*Myzus persicae*) (Sulzer, 1776) (Hemiptera, Aphididae). v1 ed. Zenodo.

Liu, Q., Neefjes, A. C. M., Singh, A., Kobylinska, R., Mugford, S. T., Marzo, M., Canham, J., Schuster, M., Van Der Hoorn, R. a. L., Chen, Y. & Hogenhout, S. A. 2025. Aphid effectors suppress plant immunity via recruiting defense proteins to processing bodies. *Science Advances,* 11**,** eadv1447.

Love, M. I., Huber, W. & Anders, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology,* 15**,** 550.

Loxdale, H. D. & Balog, A. 2018. Aphid specialism as an example of ecological–evolutionary divergence. *Biological Reviews,* 93**,** 642-657.

Maclean, A. M., Orlovskis, Z., Kowitwanich, K., Zdziarska, A. M., Angenent, G. C., Immink, R. G. H. & Hogenhout, S. A. 2014. Phytoplasma Effector SAP54 Hijacks Plant Reproduction by Degrading MADS-box Proteins and Promotes Insect Colonization in a RAD23-Dependent Manner. *PLOS Biology,* 12**,** e1001835.

Maclean, A. M., Sugio, A., Makarova, O. V., Findlay, K. C., Grieve, V. M., Tóth, R., Nicolaisen, M. & Hogenhout, S. A. 2011. Phytoplasma effector SAP54 induces indeterminate leaf-like flower development in Arabidopsis plants. *Plant Physiol,* 157**,** 831-41.

Maimbo, M., Ohnishi, K., Hikichi, Y., Yoshioka, H. & Kiba, A. 2007. Induction of a Small Heat Shock Protein and Its Functional Roles in Nicotiana Plants in the Defense Response against *Ralstonia solanacearum Plant Physiology,* 145**,** 1588-1599.

Maldonado-Bonilla, L. D. 2014. Composition and function of P bodies in *Arabidopsis thaliana*. *Frontiers in Plant Science,* Volume 5 - 2014.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011,* 17**,** 3.

Mathers, T. C. 2020a. Improved Genome Assembly and Annotation of the Soybean Aphid (*Aphis glycines Matsumura*). *G3 (Bethesda),* 10**,** 899-906.

Mathers, T. C., Chen, Y., Kaithakottil, G., Legeai, F., Mugford, S. T., Baa-Puyoulet, P., Bretaudeau, A., Clavijo, B., Colella, S., Collin, O., Dalmay, T., Derrien, T., Feng, H., Gabaldon, T., Jordan, A., Julca, I., Kettles, G. J., Kowitwanich, K., Lavenier, D., Lenzi, P., Lopez-Gomollon, S., Loska, D., Mapleson, D., Maumus, F., Moxon, S., Price, D. R., Sugio, A., Van Munster, M., Uzest, M., Waite, D., Jander, G., Tagu, D., Wilson, A. C., Van Oosterhout, C., Swarbreck, D. & Hogenhout, S. A. 2017. Rapid transcriptional plasticity of duplicated gene clusters enables a clonally reproducing aphid to colonise diverse plant species. *Genome Biol,* 18**,** 27.

Mathers, T. C., Mugford, S. T., Hogenhout, S. A. & Tripathi, L. 2020b. Genome Sequence of the Banana Aphid, *Pentalonia nigronervosa* Coquerel (Hemiptera: Aphididae) and Its Symbionts. *G3 (Bethesda),* 10**,** 4315-4321.

Mathers, T. C., Mugford, S. T., Percival-Alwyn, L., Chen, Y., Kaithakottil, G., Swarbreck, D., Hogenhout, S. A. & Van Oosterhout, C. 2019. Sex-specific changes in the aphid DNA methylation landscape. *Mol Ecol,* 28**,** 4228-4241.

Mathers, T. C., Wouters, R. H. M., Mugford, S. T., Biello, R., Van Oosterhout, C. & Hogenhout, S. A. 2023. Hybridisation has shaped a recent radiation of grass-feeding aphids. *BMC Biol,* 21**,** 157.

Mathers, T. C., Wouters, R. H. M., Mugford, S. T., Swarbreck, D., Van Oosterhout, C. & Hogenhout, S. A. 2021. Chromosome-Scale Genome Assemblies of Aphids Reveal Extensively Rearranged Autosomes and Long-Term Conservation of the X Chromosome. *Mol Biol Evol,* 38**,** 856-875.

Mayhew, P. J. 2007. Why are there so many insect species? Perspectives from fossils and phylogenies. *Biological Reviews,* 82**,** 425-454.

Mclaughlin, C. N., Qi, Y., Quake, S. R., Luo, L. & Li, H. 2022. Isolation and RNA sequencing of single nuclei from Drosophila tissues. *STAR Protoc,* 3**,** 101417.

Mehrparvar, M., Zytynska, S. E. & Weisser, W. W. 2013. Multiple Cues for Winged Morph Production in an Aphid Metacommunity. *PLOS ONE,* 8**,** e58323.

Meng, E. C., Goddard, T. D., Pettersen, E. F., Couch, G. S., Pearson, Z. J., Morris, J. H. & Ferrin, T. E. 2023. UCSF ChimeraX: Tools for structure building and analysis. *Protein Science,* 32**,** e4792.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S. & Steinegger, M. 2022. ColabFold: making protein folding accessible to all. *Nature Methods,* 19**,** 679-682.

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., Frandsen, P. B., Ware, J., Flouri, T., Beutel, R. G., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A. J., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T. R., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L. S., Kawahara, A. Y., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., Mckenna, D. D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J. L., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., Von Reumont, B. M., Schütte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N. U., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M. G., Wiegmann, B. M., Wilbrandt, J., Wipfler, B., Wong, T. K. F., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D. K., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K. M., *et al.* 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science,* 346**,** 763-767.

Moreno, A., Garzo, E., Fernandez-Mata, G., Kassem, M., Aranda, M. A. & Fereres, A. 2011. Aphids secrete watery saliva into plant tissues from the onset of stylet penetration. *Entomologia Experimentalis Et Applicata,* 139**,** 145-153.

Moreno, A., Palacios, I., Blanc, S. & Fereres, A. 2005. Intracellular Salivation Is the Mechanism Involved in the Inoculation of *Cauliflower Mosaic Virus* by Its Major Vectors *Brevicoryne brassicae* and *Myzus persicae*. *Annals of the Entomological Society of America,* 98**,** 763-769.

Mostafa, S., Wang, Y., Zeng, W. & Jin, B. 2022. Plant Responses to Herbivory, Wounding, and Infection. *International Journal of Molecular Sciences,* 23.

Mugford, S. T., Barclay, E., Drurey, C., Findlay, K. C. & Hogenhout, S. A. 2016. An Immuno-Suppressive Aphid Saliva Protein Is Delivered into the Cytosol of Plant Mesophyll Cells During Feeding. *Molecular Plant-Microbe Interactions®,* 29**,** 854-861.

Murano, K., Ogawa, K., Kaji, T. & Miura, T. 2018. Pheromone gland development and monoterpenoid synthesis specific to oviparous females in the pea aphid. *Zoological Letters,* 4**,** 9.

Musaqaf, N., Jorgensen, H. J. L. & Sigsgaard, L. 2023. Plant resistance induced by hemipterans - Effects on insect herbivores and pathogens. *Crop Protection,* 163.

Muthusamy, S. K., Dalal, M., Chinnusamy, V. & Bansal, K. C. 2017. Genome-wide identification and analysis of biotic and abiotic stress regulation of small heat shock protein (HSP20) family genes in bread wheat. *Journal of Plant Physiology,* 211**,** 100-113.

Mutti, N. S., Louis, J., Pappan, L. K., Pappan, K., Begum, K., Chen, M. S., Park, Y., Dittmer, N., Marshall, J., Reese, J. C. & Reeck, G. R. 2008. A protein from the salivary glands of the pea aphid, Acyrthosiphon pisum, is essential in feeding on a host plant. *Proc Natl Acad Sci U S A,* 105**,** 9965-9.

Nagaraju, M., Reddy, P. S., Kumar, S. A., Kumar, A., Rajasheker, G., Rao, D. M. & Kavi Kishor, P. B. 2020. Genome-wide identification and transcriptional profiling of small heat shock protein gene family under diverse abiotic stress conditions in *Sorghum bicolor* (L.). *International Journal of Biological Macromolecules,* 142**,** 822-834.

Nakabachi, A. & Ishikawa, H. 1999. Provision of riboflavin to the host aphid, *Acyrthosiphon pisum*, by endosymbiotic bacteria, *Buchnera*. *Journal of Insect Physiology,* 45**,** 1-6.

Nakabachi, A., Shigenobu, S., Sakazume, N., Shiraki, T., Hayashizaki, Y., Carninci, P., Ishikawa, H., Kudo, T. & Fukatsu, T. 2005. Transcriptome analysis of the aphid bacteriocyte, the symbiotic host cell that harbors an endocellular mutualistic bacterium, *Buchnera*. *Proceedings of the National Academy of Sciences,* 102**,** 5477-5482.

Ng, J. C. & Falk, B. W. 2006. Virus-vector interactions mediating nonpersistent and semipersistent transmission of plant viruses. *Annu Rev Phytopathol,* 44**,** 183-212.

Orantes, L. C., Zhang, W., Mian, M. a. R. & Michel, A. P. 2012. Maintaining genetic diversity and population panmixia through dispersal and not gene flow in a holocyclic heteroecious aphid species. *Heredity,* 109**,** 127-134.

Pan, C., He, X., Xia, L., Wei, K., Niu, Y. & Han, B. 2024. Proteomic Analysis of Salivary Secretions from the Tea Green Leafhopper, *Empoasca flavescens* Fabrecius. *Insects,* 15**,** 296.

Paul, A., Rao, S. & Mathur, S. 2016. The α-Crystallin Domain Containing Genes: Identification, Phylogeny and Expression Profiling in Abiotic Stress, Phytohormone Response and Development in Tomato (*Solanum lycopersicum*). *Frontiers in Plant Science,* Volume 7 - 2016.

Peters, R. S., Meusemann, K., Petersen, M., Mayer, C., Wilbrandt, J., Ziesmann, T., Donath, A., Kjer, K. M., Aspock, U., Aspock, H., Aberer, A., Stamatakis, A., Friedrich, F., Hunefeld, F., Niehuis, O., Beutel, R. G. & Misof, B. 2014. The evolutionary history of holometabolous insects inferred from transcriptome-based phylogeny and comprehensive morphological data. *BMC Evol Biol,* 14**,** 52.

Pieterse, C. M., Van Der Does, D., Zamioudis, C., Leon-Reyes, A. & Van Wees, S. C. 2012. Hormonal modulation of plant immunity. *Annu Rev Cell Dev Biol,* 28**,** 489-521.

Pitino, M. & Hogenhout, S. A. 2013. Aphid Protein Effectors Promote Aphid Colonization in a Plant Species-Specific Manner. *Molecular Plant-Microbe Interactions®,* 26**,** 130-139.

Price, M. N., Dehal, P. S. & Arkin, A. P. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol,* 26**,** 1641-50.

Pruitt, K. D., Tatusova, T. & Maglott, D. R. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res,* 35**,** D61-5.

Pruitt, R. N., Gust, A. A. & Nürnberger, T. 2021. Plant immunity unified. *Nature Plants,* 7**,** 382-383.

QIAGEN CLC Genomics Workbench 24.0.1 (https://digitalinsights.qiagen.com/)

Ramsey, J. S., Wilson, A. C., De Vos, M., Sun, Q., Tamborindeguy, C., Winfield, A., Malloch, G., Smith, D. M., Fenton, B., Gray, S. M. & Jander, G. 2007.

Genomic resources for *Myzus persicae*: EST sequencing, SNP identification, and microarray design. *BMC Genomics,* 8**,** 423.

Ray, S. & Casteel, C. L. 2022. Effector-mediated plant-virus-vector interactions. *Plant Cell,* 34**,** 1514-1531.

Rcoreteam 2023. R: A language and environment for statistical computing. 4.3.1 ed.

Rispe, C., Kutsukake, M., Doublet, V., Hudaverdian, S., Legeai, F., Simon, J. C., Tagu, D. & Fukatsu, T. 2008. Large gene family expansion and variable selective pressures for cathepsin B in aphids. *Mol Biol Evol,* 25**,** 5-17.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. 2011. Integrative genomics viewer. *Nature Biotechnology,* 29**,** 24-26.

Rodriguez, P. A., Escudero-Martinez, C. & Bos, J. I. B. 2017. An Aphid Effector Targets Trafficking Protein VPS52 in a Host-Specific Manner to Promote Virulence. *Plant Physiology,* 173**,** 1892-1903.

Rodriguez, P. A., Stam, R., Warbroek, T. & Bos, J. I. B. 2014. Mp10 and Mp42 from the Aphid Species *Myzus persicae* Trigger Plant Defenses in *Nicotiana benthamiana* Through Different Activities. *Molecular Plant-Microbe Interactions®,* 27**,** 30-39.

Rodríguez-Herva, J. J., González-Melendi, P., Cuartas-Lanza, R., Antúnez-Lamas, M., Río-Alvarez, I., Li, Z., López-Torrejón, G., Díaz, I., Del Pozo, J. C., Chakravarthy, S., Collmer, A., Rodríguez-Palenzuela, P. & López-Solanilla, E. 2012. A bacterial cysteine protease effector protein interferes with photosynthesis to suppress plant innate immune responses. *Cellular Microbiology,* 14**,** 669-681.

Rossi, M., Goggin, F. L., Milligan, S. B., Kaloshian, I., Ullman, D. E. & Williamson, V. M. 1998. The nematode resistance gene *Mi* of tomato confers resistance against the potato aphid. *Proceedings of the National Academy of Sciences of the United States of America,* 95**,** 9750-9754.

Roy, M., Bhakta, K. & Ghosh, A. 2022. Minimal Yet Powerful: The Role of Archaeal Small Heat Shock Proteins in Maintaining Protein Homeostasis. *Frontiers in Molecular Biosciences*, Volume 9 - 2022.

Saikhedkar, N., Summanwar, A., Joshi, R. & Giri, A. 2015. Cathepsins of lepidopteran insects: Aspects and prospects. *Insect Biochem Mol Biol,* 64**,** 51-9.

Sarkar, N. K., Kim, Y.-K. & Grover, A. 2009. Rice sHsp genes: genomic organization and expression profiling under stress and development. *BMC Genomics,* 10**,** 393.

Scharf, K. D., Siddique, M. & Vierling, E. 2001. The expanding family of Arabidopsis thaliana small heat stress proteins and a new family of proteins containing alpha-crystallin domains (ACD proteins). *Cell Stress Chaperones,* 6**,** 225-37.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P. & Cardona, A. 2012. Fiji: an open-source platform for biological-image analysis. *Nature Methods,* 9**,** 676-682.

Shang, F., Ding, B.-Y., Zhang, Y.-T., Wu, J.-J., Pan, S.-T. & Wang, J.-J. 2021. Genome-wide analysis of long non-coding RNAs and their association with wing development in *Aphis citricidus* (Hemiptera: Aphididae). *Insect Biochemistry and Molecular Biology,* 139**,** 103666.

Shapiro, E., Biezuner, T. & Linnarsson, S. 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics,* 14**,** 618-630.

Shen, W., Sipos, B. & Zhao, L. 2024. SeqKit2: A Swiss army knife for sequence and alignment processing. *iMeta,* 3**,** e191.

Shi, L., Kang, Y., Ding, L., Xu, L., Liu, X., Yu, A., Liu, A. & Li, P. 2025. Comprehensive characterization of poplar HSP20 gene family: genome-wide identification, stress-induced expression profiling, and protein interaction verifications. *BMC Plant Biology,* 25**,** 251.

Shibuya, N. & Minami, E. 2001. Oligosaccharide signalling for defence responses in plant. *Physiological and Molecular Plant Pathology,* 59**,** 223-233.

Shigenobu, S. & Yorimoto, S. 2022. Aphid hologenomics: current status and future challenges. *Current Opinion in Insect Science,* 50.

Shih, P. Y., Le Bras, S., Ollivier, R., Boulain, H., Morliere, S., Outreman, Y., Simon, J. C. & Sugio, A. 2025. A Salivary Effector of the Pea Aphid Interacts with Pea Proteins and Enhances Its Performance on the Host Plant. *Mol Plant Microbe Interact*.

Silva, A. X., Jander, G., Samaniego, H., Ramsey, J. S. & Figueroa, C. C. 2012. Insecticide Resistance Mechanisms in the Green Peach Aphid *Myzus persicae* (Hemiptera: Aphididae) I: A Transcriptomic Survey. *Plos One,* 7.

Simon, J.-C., Delmotte, F., Rispe, C. & Crease, T. 2003. Phylogenetic relationships between parthenogens and their sexual relatives: the possible routes to parthenogenesis in animals. *Biological Journal of the Linnean Society,* 79**,** 151-163.

Simon, J.-C., Rispe, C. & Sunnucks, P. 2002. Ecology and evolution of sex in aphids. *Trends in Ecology & Evolution,* 17**,** 34-39.

Simon, J.-C., Stoeckel, S. & Tagu, D. 2010. Evolutionary and functional insights into reproductive strategies of aphids. *Comptes Rendus Biologies,* 333**,** 488-496.

Singh, K. S., Cordeiro, E. M. G., Troczka, B. J., Pym, A., Mackisack, J., Mathers, T. C., Duarte, A., Legeai, F., Robin, S., Bielza, P., Burrack, H. J., Charaabi, K., Denholm, I., Figueroa, C. C., Ffrench-Constant, R. H., Jander, G., Margaritopoulos, J. T., Mazzoni, E., Nauen, R., Ramírez, C. C., Ren, G., Stepanyan, I., Umina, P. A., Voronova, N. V., Vontas, J., Williamson, M. S., Wilson, A. C. C., Xi-Wu, G., Youn, Y.-N., Zimmer, C. T., Simon, J.-C., Hayward, A. & Bass, C. 2021. Global patterns in genomic diversity underpinning the evolution of insecticide resistance in the aphid crop pest *Myzus persicae*. *Communications Biology,* 4**,** 847.

Skendžić, S., Zovko, M., Živković, I. P., Lešić, V. & Lemić, D. 2021. The Impact of Climate Change on Agricultural Insect Pests. *Insects,* 12.

Snoeck, S., Guayazán-Palacios, N. & Steinbrenner, A. D. 2022. Molecular tug-of-war: Plant immune recognition of herbivory. *Plant Cell,* 34**,** 1497-1513.

Stengel, F., Baldwin, A. J., Painter, A. J., Jaya, N., Basha, E., Kay, L. E., Vierling, E., Robinson, C. V. & Benesch, J. L. P. 2010. Quaternary dynamics and plasticity underlie small heat shock protein chaperone function. *Proceedings of the National Academy of Sciences,* 107**,** 2007-2012.

Stork, N. E., Mcbroom, J., Gely, C. & Hamilton, A. J. 2015. New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. *Proceedings of the National Academy of Sciences of the United States of America,* 112**,** 7519-7523.

Stout, J. 2013. Reevaluating the conceptual framework for applied research on host-plant resistance. *Insect Science*, 20, 263–272.

Sugio, A., Kingdom, H. N., Maclean, A. M., Grieve, V. M. & Hogenhout, S. A. 2011. Phytoplasma protein effector SAP11 enhances insect vector reproduction by manipulating plant development and defense hormone biosynthesis. *Proc Natl Acad Sci U S A,* 108**,** E1254-63.

Sun, C., Shao, Y. & Iqbal, J. 2023. Insect Insights at the Single-Cell Level: Technologies and Applications. *Cells,* 13.

Swindell, W. R., Huebner, M. & Weber, A. P. 2007. Transcriptional profiling of *Arabidopsis* heat shock proteins and transcription factors reveals extensive overlap between heat and non-heat stress response pathways. *BMC Genomics*, 8, 125.

Tao, P., Guo, W. L., Li, B. Y., Wang, W. H., Yue, Z. C., Lei, J. L. & Zhong, X. M. 2015. Genome-wide identification, classification, and expression analysis of sHSP genes in Chinese cabbage (*Brassica rapa* ssp pekinensis). *Genet Mol Res,* 14**,** 11975-93.

The International Aphid Genomics Consortium 2010. Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biol,* 8**,** e1000313.

Thompson, G. A. & Goggin, F. L. 2006. Transcriptomics and functional genomics of plant defence induction by phloem-feeding insects. *Journal of Experimental Botany,* 57**,** 755-766.

Tjallingii, W. F. 2006. Salivary secretions by aphids interacting with proteins of phloem wound responses. *Journal of Experimental Botany,* 57**,** 739-745.

Tobin, P. C. 2018. Managing invasive species. *F1000Res,* 7.

Toghani, A. & Kamoun, S. 2024. Functional annotation of 180 RefSeq reference plant proteomes reveals a dataset of 113,684 NLR proteins. Zenodo.

Torres Ascurra, Y. C., Zhang, L., Toghani, A., Hua, C., Rangegowda, N. J., Posbeyikian, A., Pai, H., Lin, X., Wolters, P. J., Wouters, D., De Blok, R., Steigenga, N., Paillart, M. J. M., Visser, R. G. F., Kamoun, S., Nürnberger, T. & Vleeshouwers, V. G. a. A. 2023. Functional diversification of a wild potato immune receptor at its center of origin. *Science,* 381**,** 891-897.

Turk, B., Turk, D. & Turk, V. 2000. Lysosomal cysteine proteases: more than scavengers. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology,* 1477**,** 98-111.

Uzest, M., Gargani, D., Dombrovsky, A., Cazevieille, C., Cot, D. & Blanc, S. 2010. The "acrostyle": a newly described anatomical structure in aphid stylets. *Arthropod Struct Dev,* 39**,** 221-9.

Uzest, M., Gargani, D., Drucker, M., Hébrard, E., Garzo, E., Candresse, T., Fereres, A. & Blanc, S. 2007. A protein key to plant virus transmission at the tip of the insect vector stylet. *Proceedings of the National Academy of Sciences,* 104**,** 17959-17964.

Van Bel, A. J. E. & Will, T. 2016. Functional Evaluation of Proteins in Watery and Gel Saliva of Aphids. *Frontiers in Plant Science,* Volume 7 - 2016.

Van Emden, H. F. & Harrington, R. 2007. *Aphids as crop pests*, Cabi Publishing.

Van Montfort, R. L. M., Basha, E., Friedrich, K. L., Slingsby, C. & Vierling, E. 2001. Crystal structure and assembly of a eukaryotic small heat shock protein. *Nature Structural Biology,* 8**,** 1025-1030.

Verchot, J. 2012. Cellular chaperones and folding enzymes are vital contributors to membrane bound replication and movement complexes during plant RNA virus infection. *Frontiers in Plant Science,* Volume 3 - 2012.

Wan, J., Zhang, X. C. & Stacey, G. 2008. Chitin signaling and plant disease resistance. *Plant Signal Behav,* 3**,** 831-3.

Wang, Q., Yuan, E., Ling, X., Zhu-Salzman, K., Guo, H., Ge, F. & Sun, Y. 2020. An aphid facultative symbiont suppresses plant defence by manipulating aphid gene expression in salivary glands. *Plant Cell Environ,* 43**,** 2311-2322.

Wang, W., Dai, H., Zhang, Y., Chandrasekar, R., Luo, L., Hiromasa, Y., Sheng, C., Peng, G., Chen, S., Tomich, J. M., Reese, J., Edwards, O., Kang, L., Reeck, G. & Cui, F. 2015. Armet is an effector protein mediating aphid-plant interactions. *The FASEB Journal,* 29**,** 2032-2045.

Wang, X., Wang, R., Ma, C., Shi, X., Liu, Z., Wang, Z., Sun, Q., Cao, J. & Xu, S. 2017. Massive expansion and differential evolution of small heat shock proteins with wheat (*Triticum aestivum* L.) polyploidization. *Scientific Reports,* 7**,** 2581.

Wang, X., Zhai, Y. & Zheng, H. 2024. Deciphering the cellular heterogeneity of the insect brain with single-cell RNA sequencing. *Insect Science,* 31**,** 314-327.

Wari, D., Kabir, M. A., Mujiono, K., Hojo, Y., Shinya, T., Tani, A., Nakatani, H. & Galis, I. 2019. Honeydew-associated microbes elicit defense responses against brown planthopper in rice. *Journal of Experimental Botany,* 70**,** 1683-1696.

Waters, E. R. 2012. The evolution, function, structure, and expression of the plant sHSPs. *Journal of Experimental Botany,* 64**,** 391-403.

Webster, C. G., Thillier, M., Pirolles, E., Cayrol, B., Blanc, S. & Uzest, M. 2017. Proteomic composition of the acrostyle: Novel approaches to identify cuticular proteins involved in virus-insect interactions. *Insect Science,* 24**,** 990-1002.

Weston, D. J., Karve, A. A., Gunter, L. E., Jawdy, S. S., Yang, X., Allen, S. M. & Wullschleger, S. D. 2011. Comparative physiology and transcriptional networks underlying the heat shock response in *Populus trichocarpa*, *Arabidopsis thaliana* and *Glycine max*. *Plant, Cell & Environment*, 34, 1488-1506.

Whitham, S. A., Anderberg, R. J., Chisholm, S. T. & Carrington, J. C. 2000. Arabidopsis RTM2 Gene Is Necessary for Specific Restriction of *Tobacco Etch Virus* and Encodes an Unusual Small Heat Shock–like Protein. *The Plant Cell,* 12**,** 569-582.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*, Springer Charm.

Will, T., Kornemann, S. R., Furch, A. C. U., Tjallingii, W. F. & Van Bel, A. J. E. 2009. Aphid watery saliva counteracts sieve-tube occlusion: a universal phenomenon? *Journal of Experimental Biology,* 212**,** 3305-3312.

Will, T., Tjallingii, W. F., Thönnessen, A. & Van Bel, A. J. E. 2007. Molecular sabotage of plant defense by aphid saliva. *Proceedings of the National Academy of Sciences of the United States of America,* 104**,** 10536-10541.

Will, T. & Vilcinskas, A. 2015. The structural sheath protein of aphids is required for phloem feeding. *Insect Biochemistry and Molecular Biology,* 57**,** 34-40.

Wilson, A. C. C., Ashton, P. D., Calevro, F., Charles, H., Colella, S., Febvay, G., Jander, G., Kushlan, P. F., Macdonald, S. J., Schwartz, J. F., Thomas, G. H. & Douglas, A. E. 2010. Genomic insight into the amino acid relations of the pea aphid, *Acyrthosiphon pisum*, with its symbiotic bacterium *Buchnera aphidicola*. *Insect Molecular Biology,* 19**,** 249-258.

Wouters, R. 2021. *The genetic diversity of natural populations of Myzus persicae, a polyphagous aphid species.* Doctoral, University of East Anglia.

Yu, J., Cheng, Y., Feng, K., Ruan, M., Ye, Q., Wang, R., Li, Z., Zhou, G., Yao, Z., Yang, Y. & Wan, H. 2016. Genome-Wide Identification and Expression Profiling of Tomato Hsp20 Gene Family in Response to Biotic and Abiotic Stresses. *Frontiers in Plant Science,* Volume 7 - 2016.

Zhang, J. & Zhou, J. M. 2010. Plant Immunity Triggered by Microbial Molecular Signatures. *Molecular Plant,* 3**,** 783-793.

Zhang, Q., Dai, B., Fan, M., Yang, L., Li, C., Hou, G., Wang, X., Gao, H. & Li, J. 2024. Genome-wide profile analysis of the Hsp20 family in lettuce and identification of its response to drought stress. *Frontiers in Plant Science,* Volume 15 - 2024.

Zhang, Y., Fu, Y., Liu, X., Francis, F., Fan, J., Liu, H., Wang, Q., Sun, Y., Zhang, Y. & Chen, J. 2023. SmCSP4 from aphid saliva stimulates salicylic acid-mediated defence responses in wheat by interacting with transcription factor TaWKRY76. *Plant Biotechnol J,* 21**,** 2389-2407.

Zhao, P., Wang, D., Wang, R., Kong, N., Zhang, C., Yang, C., Wu, W., Ma, H. & Chen, Q. 2018. Genome-wide analysis of the potato Hsp20 gene family: identification, genomic organization and expression profiles in response to heat stress. *BMC Genomics,* 19**,** 61.

Appendix

```
## Quality Control of raw reads
[seddonr@NBI-HPC quality_control]$ while read id; do ~/scripts/submit-slurm_v1.1.pl -q jic-medium -m 25000 -c 16 -t 1-00:00 -e -j ${id}_qc -i  "source fastqc-0.11.8; fastqc reads/${id}_1.fq.gz reads/${id}_2.fq.gz"; done < all_file_prefixes.txt
[seddonr@NBI-HPC quality_control]$ multiqc .

## Trimming of raw reads
[seddonr@NBI-HPC tissue_specific_RNAseq_all_data]$ mkdir trimmed_reads
[seddonr@NBI-HPC tissue_specific_RNAseq_all_data]$ while read id; do ~/scripts/submit-slurm_v1.1.pl -q jic-medium -m 25000 -c 16 -t 1-00:00 -e -j ${id}_trimmming -i "source trim_galore-0.6.10_CBG; trim_galore --output_dir trimmed_reads --quality 20 --length 20 --stranded_illumina --paired --phred33 reads/${id}_1.fq.gz reads/${id}_2.fq.gz"; done < all_file_prefixes.txt

## Quality Control of trimmed reads

[seddonr@NBI-HPC tissue_specific_RNAseq_all_data]$ while read id; do ~/scripts/submit-slurm_v1.1.pl -q jic-medium -m 25000 -c 16 -t 1-00:00 -e -j ${id}_qc -i  "source fastqc-0.11.8; fastqc trimmed_reads/${id}_1_val_1.fq.gz trimmed_reads/${id}_2_val_2.fq.gz"; done < all_file_prefixes.txt

## Genome index

[seddonr@NBI-HPC tissue_specific_RNAseq_all_data]$ ~/scripts/submit-slurm_v1.1.pl -q jic-medium -m 25000 -c 16 -t 1-00:00 -e -j Mpersicae_Genome_Index -i "source hisat2-2.2.1_CBG; hisat2-build tissue_specific_RNAseq/Myzus_persicae_O_v2.0.scaffolds.fa MP_index"

## Alignment of reads to indexed genome

[seddonr@NBI-HPC tissue_specific_RNAseq_all_data]$ while read id; do ~/scripts/submit-slurm_v1.1.pl -q jic-medium -m 100000 -c 32 -t 1-00:00 -e -j ${id}_mapping -i "source switch-institute ei;source HISAT-2.0.5; source samtools-1.3;hisat2 -q --dta-cufflinks --rna-strandness RF -p 32 -x MP_index -1 trimmed_reads/${id}_1_val_1.fq.gz -2 trimmed_reads/${id}_2_val_2.fq.gz | samtools view -b -o ${id}.bam -; mkdir -p ${id}_tmp; samtools sort -o ${id}.sorted.bam -T ./${id}_tmp/aln.sorted --threads 32 -m 3G ${id}.bam; samtools index -c ${id}.sorted.bam; samtools index -b ${id}.sorted.bam;rm ${id}.bam"; done < all_file_prefixes.txt

## Checking alignment rates

[seddonr@NBI-HPC tissue_specific_RNAseq_all_data]$ while read id; do ls ${id}_mapping.err.slurm >> concat_read_alignments.txt; tail -n 15 ${id}_mapping.err.slurm >> concat_read_alignments.txt; done < all_file_prefixes.txt

## Quantification of reads

[seddonr@NBI-HPC read_counts]$ pwd
read_counts

[seddonr@NBI-HPC read_counts]$ while read id; do ~/scripts/submit-slurm_v1.1.pl -q jic-medium -m 100000 -c 32 -t 1-00:00 -e -j ${id}_count -i "source htseq-0.12.4_CBG; htseq-count -s reverse --nonunique fraction sorted_bam/${id}.sorted.bam practice_DE_analysis/MYZPE13164_O_EIv2.1.annotation.gff3.gtf -c ${id}_read_count_rev.tsv"; done < all_file_prefixes.txt
```

Supplementary Figure 1: RNA-seq analysis was performed using an established multi-tool pipelined for the alignment and counting of reads.

Raw reads were checked for quality using FastQC and then trimmed using TrimGalore. The *M. persicae* genome was indexed and reads were aligned using the HISAT2 tool. Counts were then generated using HTSeq-count. Resulting read counts were used in downstream expression analyses in RStudio.

Sample	Batch
Antenna_1	1
Antenna_2	1
Antenna_6	1
Antenna_8	1
Antenna_11	2
Foregut_1	1
Foregut_2	1
Foregut_6	1
Foregut_8	1
Foregut_9	2
Foregut_11	2
Foregut_12	2
Hindgut_1	1
Hindgut_2	1
Hindgut_6	1
Hindgut_8	1
Leg_1	1
Leg_2	1

Sample	Batch
Leg_6	1
Leg_8	1
Leg_10	2
Ovary_9	2
Ovary_10	2
Ovary_11	2
Ovary_12	2
SGland_1	1
SGland_2	1
SGland_6	1
SGland_8	1
SGland_10	2
SGland_12	2
Stylet_1	1
Stylet_2	1
Stylet_6	1
Stylet_8	1
Stylet_9	2

Supplementary Table 1: RNA sequencing was performed in two batches.
Total RNA was sequenced by Novogene across two batches. These batches were added as metadata in downstream differential expression analysis.

Organism Name	Assembly Name
Arabidopsis thaliana	TAIR10.1
Micromonas commoda	ASM9098v2
Vitis vinifera	ASM3070453v1
Citrus sinensis	DVS_A1.0
Ziziphus jujuba	ASM3175591v1
Ostreococcus lucimarinus CCE9901	ASM9206v1
Daucus carota subsp. sativus	DH1 v3.0
Solanum pennellii	SPENNV200
Corylus avellana	CavTom2PMs-1.0
Ostreococcus tauri	version 140606
Rhododendron vialii	ASM3025357v1
Bathycoccus prasinos	ASM222023v1
Ipomoea triloba	ASM357664v1
Zea mays	Zm-B73-REFERENCE-NAM-5.0
Solanum lycopersicum	SL3.1
Triticum aestivum	IWGSC CS RefSeq v2.1
Glycine max	Glycine_max_v4.0
Gossypium hirsutum	Gossypium_hirsutum_v2.1
Hordeum vulgare subsp. vulgare	MorexV3_pseudomolecules_assembly
Brassica napus	Da-Ae
Brassica rapa	CAAS_Brap_v3.01
Sorghum bicolor	Sorghum_bicolor_NCBIv3
Malus domestica	ASM211411v1
Medicago truncatula	MtrunA17r5.0-ANR
Chlamydomonas reinhardtii	Chlamydomonas_reinhardtii_v5.5
Cucumis sativus	Cucumber_9930_V3
Arachis hypogaea	arahy.Tifrunner.gnm1.KYV3
Capsicum annuum	UCD10Xv1.1
Brassica oleracea var. oleracea	BOL

Organism Name	Assembly Name
Prunus persica	Prunus_persica_NCBIv2
Phaseolus vulgaris	PhaVulg1_0
Cucumis melo	USDA_Cmelo_AY_1.0
Physcomitrium patens	Phypa V3
Cicer arietinum	ASM33114v1
Manihot esculenta	M.esculenta_v8
Brachypodium distachyon	Brachypodium_distachyon_v3.0
Cannabis sativa	ASM2916894v1
Lactuca sativa	Lsat_Salinas_v11
Hevea brasiliensis	ASM3005281v1
Populus trichocarpa	P.trichocarpa_v4.1
Musa acuminata AAA Group	Cavendish_Baxijiao_AAA
Pisum sativum	CAAS_Psat_ZW6_1.0
Helianthus annuus	HanXRQr2.0-SUNRISE
Olea europaea var. sylvestris	O_europaea_v1
Setaria italica	Setaria_italica_v2.0
Raphanus sativus	ASM80110v3
Mangifera indica	CATAS_Mindica_2.1
Elaeis guineensis	EG5
Beta vulgaris subsp. vulgaris	EL10.2
Lotus japonicus	LjGifu_v1.2
Salvia miltiorrhiza	IMPLAD_Smil_shh
Sesamum indicum	S_indicum_v1.0
Gossypium arboreum	ASM2569848v2
Panicum virgatum	P.virgatum_v5
Fragaria vesca subsp. vesca	FraVesHawaii_1.0
Coffea arabica	Cara_1.0
Lolium perenne	MPB_Lper_Kyuss_1697
Theobroma cacao	Criollo_cocoa_genome_V2
Vigna radiata var. radiata	Vradiata_ver6
Punica granatum	ASM765513v2
Juglans regia	Walnut 2.0

Organism Name	Assembly Name
Vigna unguiculata	ASM411807v2
Ananas comosus	ASM154086v1
Phoenix dactylifera	palm_55x_up_171113_PBpolish2nd_filt_p
Eucalyptus grandis	ASM1654582v1
Cajanus cajan	C.cajan_V1.1
Ricinus communis	ASM1957865v1
Spinacia oleracea	BTI_SOV_V1
Aegilops tauschii subsp. strangulata	Aet v5.0
Rosa chinensis	RchiOBHm-V2
Setaria viridis	Setaria_viridis_v2.0
Cryptomeria japonica	Sugi_1.0
Cucurbita pepo subsp. pepo	ASM280686v2
Zingiber officinale	Zo_v1.1
Quercus robur	dhQueRobu3.1
Camelina sativa	Cs
Prunus dulcis	ALMONDv2
Prunus mume	P.mume_V1.0
Humulus lupulus	drHumLupu1.1
Lycium barbarum	ASM1917538v2
Asparagus officinalis	Aspof.V1
Vigna angularis	ASM1680809v1
Carya illinoinensis	C.illinoinensisPawnee_v1
Populus nigra	ddPopNigr1.1
Glycine soja	ASM419377v2
Papaver somniferum	ASM357369v1
Oryza glaberrima	OglaRS2
Gossypium raimondii	ASM2569854v1
Trifolium pratense	ARS_RC_1.1
Phragmites australis	lpPhrAust1.1
Triticum urartu	Tu2.1
Triticum dicoccoides	WEW_v2.1
Lupinus angustifolius	LupAngTanjil_v1.0
Benincasa hispida	ASM972705v1
Cynara cardunculus var. scolymus	CcrdV1.1
Vitis riparia	EGFV_Vit.rip_1.0
Nicotiana attenuata	NIATTr2

Organism Name	Assembly Name
Arachis duranensis	aradu.V14167.gnm2.J7QH
Oryza brachyantha	ObraRS2
Tripterygium wilfordii	ASM1340144v1
Erigeron canadensis	C_canadensis_v1
Salvia hispanica	UniMelb_Shisp_WGS_1.0
Solanum dulcamara	daSolDulc1.2
Diospyros lotus	ASM1463336v1
Panicum hallii	PHallii_v3.1
Rosa rugosa	drRosRugo1.1
Euphorbia lathyris	ddEupLath1.1
Argentina anserina	drPotAnse1.1
Malania oleifera	ASM2987363v1
Salvia splendens	SspV2
Arachis stenosperma	arast.V10309.gnm1.PFL2
Quercus lobata	ValleyOak3.2
Coffea eugenioides	Ceug_1.0
Alnus glutinosa	dhAlnGlut1.1
Malus sylvestris	drMalSylv7.2
Actinidia eriantha	MaoHua_MHT
Amaranthus tricolor	ASM2621246v1
Macadamia integrifolia	SCU_Mint_v3
Arachis ipaensis	Araip1.1
Magnolia sinica	MsV1
Rhodamnia argentea	ASM2092103v1
Juglans microcarpa x Juglans regia	Jm3101_v1.0
Impatiens glandulifera	dImpGla2.1
Dioscorea cayenensis subsp. rotundata	TDr96_F1_v2_PseudoChromosome.rev07_lg8_w22 25.fasta
Lolium rigidum	APGP_CSIRO_Lrig_0.1
Solanum stenotomum	ASM1918654v1
Mercurialis annua	ddMerAnnu1.2
Telopea speciosissima	Tspe_v1
Nymphaea colorata	ASM883128v2
Vicia villosa	Vvil1.0

Organism Name	Assembly Name
Lycium ferocissimum	AGI_CSIRO_Lferr_CH_V1
Carica papaya	Papaya1.0
Monoraphidium neglectum	mono_v1
Nicotiana tabacum	Ntab-TN90
Solanum tuberosum	SolTub_3.0
Camellia sinensis	AHAU_CSS_1
Jatropha curcas	RJC1_Hi-C
Prunus avium	PAV_r1.0
Chenopodium quinoa	ASM168347v1
Pyrus x bretschneideri	Pyrus_bretschneideri_v1
Nelumbo nucifera	Chinese Lotus 1.1
Arabidopsis lyrata subsp. lyrata	v.1.0
Erythranthe guttata	Mimgu1_0
Eutrema salsugineum	Eutsalg1_0
Momordica charantia	ASM199503v1
Quercus suber	Cork oak 2.0
Andrographis paniculata	ASM980555v1
Populus alba	ASM523922v1
Citrus x clementina	Citrus_clementina_v1.0
Capsella rubella	Caprub1_0
Cucurbita moschata	Cmos_1.0
Populus euphratica	PopEup_1.0
Selaginella moellendorffii	v1.0
Vigna umbellata	ASM1883591v1
Tarenaya hassleriana	ASM46358v1
Volvox carteri f. nagariensis	v1.0
Cucurbita maxima	Cmax_1.0
Auxenochlorella protothecoides	ASM73321v1
Amborella trichopoda	AMTR1.0
Phalaenopsis equestris	ASM126359v1
Dendrobium catenatum	ASM160598v2
Morus notabilis	ASM41409v2
Chlorella variabilis	v 1.0
Ipomoea nil	Asagao_1.1
Pistacia vera	PisVer_v2
Durio zibethinus	Duzib1.0
Hibiscus syriacus	ASM638163v2
Coccomyxa subellipsoidea C-169	Coccomyxa subellipsoidae v2.0
Micromonas pusilla CCMP1545	Micromonas pusilla CCMP1545 v2.0
Prosopis alba	ASM479914v2
Nicotiana sylvestris	Nsyl
Cornus florida	ASM3098733v1
Prosopis cineraria	ASM2901754v1
Nicotiana tomentosiformis	Ntom_v01
Abrus precatorius	Abrus_2018
Syzygium oleosum	Sole_v2.1
Solanum verrucosum	falcon-dt-bn
Gastrolobium bilobum	APGP_CSIRO_Gbil_v1
Herrania umbratica	ASM216827v2
Oryza sativa Japonica Group	ASM3414082v1

Supplementary Table 2: Species names and annotation names which were used to generate the concatenated proteomes of 180 plant species.

Genomes were downloaded by AmirAli Toghani (Kamoun Lab, The Sainsbury Laboratory) from the NCBI Genome Database by searching "Viridiplantae" and filtering reference genomes annotated using NCBI RefSeq.