Global analysis of genes involved in capsule development and seed mucilage polysaccharide production in *Plantago ovata*

Lina Herliana

MBiotech (Plant Biotechnology)

ꕥ

A thesis submitted to The University of Adelaide in fulfilment of the requirement for the degree of Doctor of Philosophy

The University of Adelaide

Faculty of Sciences

School of Agriculture, Food and Wine

October 2022

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Lina Herliana

October 2022

Acknowledgments

I am incredibly grateful to pursue my PhD study under the supervision of Prof. Rachel Burton, Dr Tina Bianco-Miotto and Dr Nathan Watson-Haigh. I could not have undertaken this journey without help from Prof. Rachel in providing a lot of things, not limited to helping me get a scholarship, funding this project, and trusting me to be involved in Plantago ovata research. Also, for giving me emotional support and providing time and energy to help complete this thesis. I admire your expertise in science and your leadership management. I am also immensely thankful to Dr Tina Bianco for the insightful discussion and time in advising and editing this thesis. Your knowledge and connection also help me successfully run my experiments. I am deeply indebted to Dr Nathan Watson-Haigh. It was a pleasure to learn bioinformatics directly from the expert. You have ways to guide and teach people from different backgrounds or levels of understanding in bioinformatics. Special thanks to my postgraduate coordinator, Associate Professor Kenneth Chalmers, for PhD consultation and for helping with the administration.

I would like to thank The University of Adelaide for awarding me the Adelaide Graduate Research Scholarship (AGRS) and Plant Energy Biology Supplementary Scholarship. Further thanks to my workplace, The National Research and Innovation Agency (BRIN-Indonesia), for permitting me to undertake study leave to enhance my skills. I would like to acknowledge institutions and people that provide the facility, services, or data for my research. The University of Adelaide provided me access to Phoenix HPC (High-Performance Computer) facility to run workflows, the Undercroft Glasshouse facility for growing plants, and the Adelaide Microscopy facility for cryosectioning and laser capture microdissecting (LCM) experiments. I would like to recognise Dr Fabien Voisin for his technical support in using the Phoenix-HPC, Dr Gwen Mayo for her assistance with microscopy and Melissa Pickering for plant care. Basil Hetzel Institute also provided a temporary facility for me to conduct LCM experiments. The Flinders Genomics Facility for RNA sequencing capsule tissues and the South Australian Genomics Centre (SAGC) for RNA quality check and sequencing of early

developing seed coat tissues. Dr Mark Van der Hoek and Dr Timothy Rudd from SAGC assisted in troubleshooting the experiment to obtain good-quality RNA. SA pathology also provided services to check RNA quality for my LCM samples. TECAN company to create custom ribodepletion kit using newly P. ovata assembled genome to target ribosomal RNA from LCM samples.

I would like to express my sincerest thanks to Associate Professor Matthew Tucker for his advice during my candidature and for providing chemical and membrane slides for troubleshooting RNA extraction and P. ovata mutant population. Words cannot express my appreciation to Dr Lisa O'Donovan for her help during my study, especially in microscopy works. You are such a caring person, and we are a good team. Thanks to Dr James Cowley for collaborating with me to publish papers and help with experiments. I acknowledge the valuable discussions about genome assembly with Aaron Phillips.

Many thanks to former and current Plant Cell Wall and RABLAB group members for creating a welcoming environment. I thank Dr Neil Shirley, Dr Julian Schwerdt, Dr Jana Phan, Tycho Neumann, Sandy Khor, Dr Renee Philips, Dr Carolyn Schultz, Dr Juanita Lauer, Tharuka Jayampathi, Ali Gill, Ghazwan Karem, Jakob Schulz, and Dr Helen Collins.

Lastly, I would like to thank my parents, who always supported me in achieving my dream. Thank you for caring for my kid, especially during my thesis writing. I am glad I managed to finish my thesis amid the Covid-19 pandemic that impacted my work.

Table of Contents

List of Publications and Expected Publications

1. Cowley JM, **Herliana L**, Neumann KA, Ciani S, Cerne V, and Burton RA (2020) A small-scale fractionation pipeline for rapid analysis of seed mucilage characteristics. Plant Methods 16:20 https://doi.org/10.1186/s13007-020- 00569-6
2. Phan JL, Cowley JM, Neumann KA, **Herliana L**, O'Donovan LA, and Burton RA (2020) The novel features of *Plantago ovata* seed mucilage accumulation, storage and release. Scientific Reports **10 11766** https://doi.org/10.1038/s41598-020-68685-w
3. **Herliana L**, Schwerdt JG, Neumann TR, Severn-Ellis, Phan JL, Cowley JM, Shirley NJ, Tucker MR, Bianco-Miotto T, Batley Jacqueline, Watson-Haigh NS and Burton RA (Unpublished) A chromosome-level genome of *Plantago ovata*. Submitted for Publication to Scientific Reports.
4. **Herliana L**, Cowley JM, O'Donovan LA, Neumann TR, Khor SF, Watson-Haigh NS, Bianco-Miotto, and Burton RA (Unpublished) Morphological and transcriptomic analysis of developing *Plantago ovata* seed capsules reveal structural features and key gene network.

Chapter 1

General Introduction

ꕥ•ꕥ

Thesis Introduction

Plantago ovata is an annual plant producing seed husk or psyllium that turns into gel-like material called mucilage when wetted, which has many downstream applications. It has long been known as a medicinal plant as it helps reduce constipation and lower blood cholesterol levels. Psyllium is also widely used as a gluten replacement in bread making. As industry usage grows, the demand for psyllium increases. It is reported that conventional breeding approaches have not made a significant improvement to yield. To improve the quality and quantity of psyllium, we need to identify the problems and then tailor the most appropriate approach to solve them.

One of the most problematic issues in cultivating *P. ovata* is capsule shattering. Capsules are dry fruits that release the seeds because the pericarp splits open when reaching maturity. Heavy rains and strong winds increase shattering events leading to high yield loss. Capsule maturation is not synchronised in this plant, adding another layer of complexity in harvesting. To prevent high yield loss, farmers harvest the plants before all capsules are mature by manually pulling up the entire plant. This action leads to mixed quality of seeds. Potentially good quality seeds could be harvested simultaneously using a machine if the capsules were still intact or indehiscent. Understanding capsule development and finding the candidate genes controlling this mechanism will help to address the shattering problem.

The husk accounts for only a quarter of the total seed weight, and most of the seed parts are used for other purposes with less investment return for the growers, like animal feed. The question is how to increase the husk yield, and thus the proportion of seed coat producing mucilage, without affecting embryo development. While seed mucilage accumulation, storage, and release mechanisms in *P. ovata* have been published, the candidate genes controlling these mechanisms are mostly still unknown.

While diversity is critical to crop breeding programs, *P. ovata* commercial germplasm lacks this feature. Gamma irradiation has been reported to introduce variation that successfully led to a patented cultivar. A previous study in Adelaide, Australia has generated thousands of mutants that can be used to identify candidate genes for traits of interest. Here, two different mutants were used to study development of the capsule (shattering) and the outer seed layers (mucilage). As a reference genome was unavailable, a genome assembly with high-quality annotation was also generated during this study.

Thesis Structure

This thesis contains seven chapters comprised of a general thesis introduction (Chapter 1), a review of the literature (Chapter 2), and two research chapters in publication format (Chapters 3 and 4) where one of them has been submitted for publication, two research chapters in conventional format (Chapters 5 and 6) and a final general discussion with suggestions for future directions (Chapter 7). A schematic overview of the thesis structure is presented in Figure 1.

The review of the literature, Chapter 2, aims to introduce readers to *Plantago ovata* market value and factors affecting psyllium production, current breeding progress, and the status of the genetic understanding of seed shattering and mucilage production from model species and other crops, as well as for *P. ovata*.

In Chapter 3, a universal reference genome is presented. *De novo* contig assembly was performed on genomic long-reads and scaffolded using genomic Hi-C short reads. RNA-sequencing data was used to annotate the assembled genome. It is clear that the *P. ovata* genome contains a very high amount of repeat sequences and we hypothesize that this has hindered the development of effective marker-assisted breeding.

In Chapter 4, experiments are presented that were designed to better understand the *P. ovata* shattering mechanism and identify candidate genes controlling capsule shattering. Transcripts

were extracted from capsule tissues from two genotypes; the wild type and a mutant called *accelerato* (*ace*), at four different time points. The transcriptome profiles in the capsule tissues were compared between wild type and *ace* to identify differences both between the two lines and across development.

In Chapter 5, results of a screen of about 201 gamma-irradiated mutant lines searching for a candidate mutant that yields a higher quantity of mucilage than the wild type are presented. The mutant *raya (ray)* is introduced, with more mucilage but no other obvious deleterious seed defects, providing a comparative genotype for wild type in the following RNA sequencing analysis in Chapter 6.

In Chapter 6, the study aims to identify candidate genes controlling mucilage production and to shed more light on events occurring in developing seed tissues. Transcripts were quantified from integument (INT) and mucilage secretory cells (MSCs) from two genotypes (wild type and *raya* (*ray*)) at three different time points. This allowed a comparison of the two tissues which are predicted to have unique developmental pathways, a comparison of the wild type versus mutant tissues and mapping of important synthesis and metabolism pathways in each of the tissues individually through time.

A summary is provided in Chapter 7 with suggestions for further research.

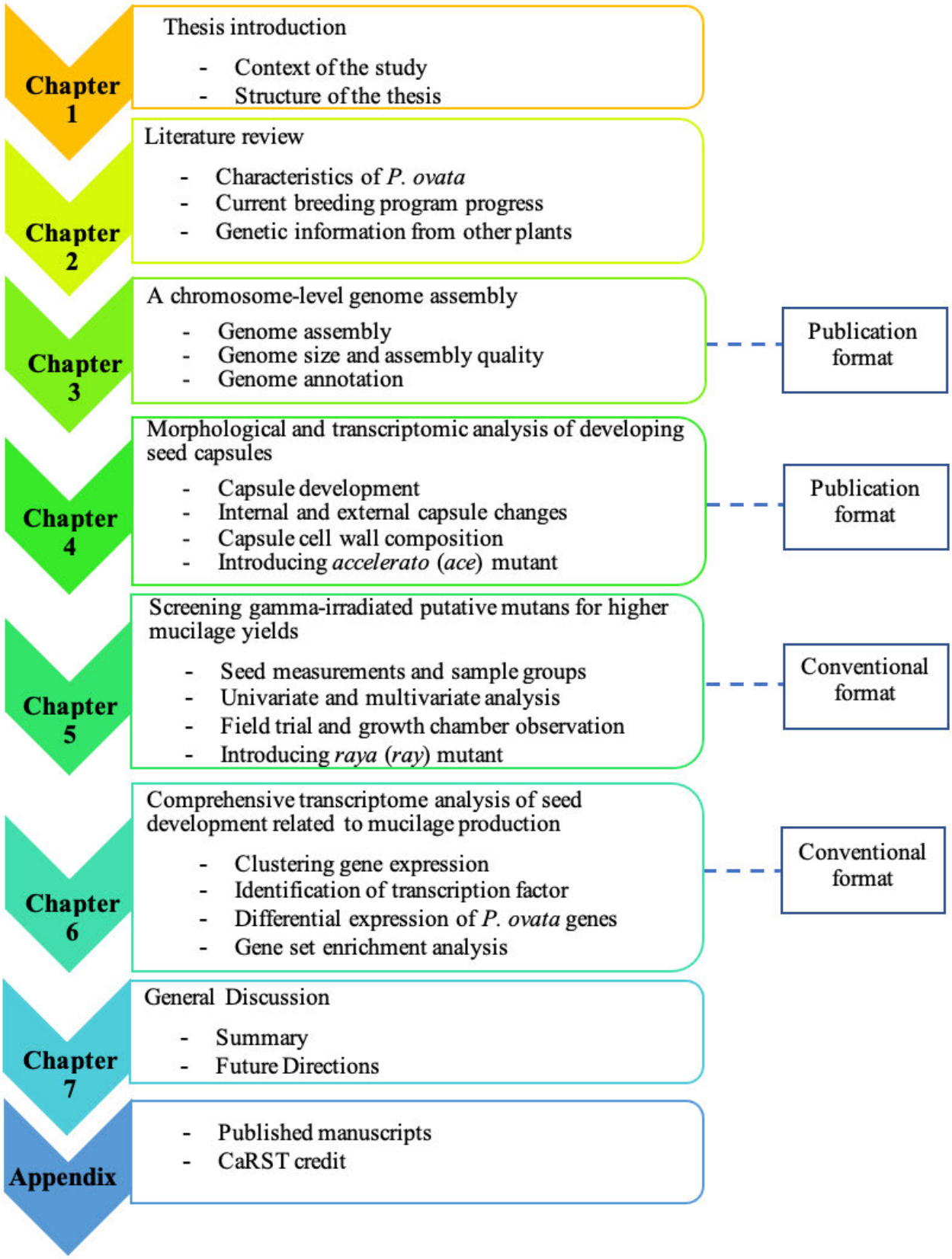


Figure 1. A schematic overview of the thesis structure.

Chapter 2

Plantago ovata capsule and seed developmental traits influencing yield and industrial end uses: Literature review

ꕥ

Introduction

Plantago ovata Forsk is an annual herb propagated through seeds (spermatophytes) that belongs to the family Plantaginaceae with more than 200 close relatives (Rønstadt et al., 2002; Dhar et al., 2005; Gonçalves and Romano, 2016). *P. ovata* is native to the Mediterranean and West Asia (Kumar, 2015). The plants are cultivated mainly for their seed coats (husks), known as isabgol or psyllium (Gonçalves and Romano, 2016). Upon hydration, the seeds absorb water, forming mucilage, in a process called myxospermy (Cowley et al., 2020; Phan et al., 2020). Many studies and review papers have shown the benefits and added value of psyllium mucilage for health supplements, food ingredients, and industrial uses (Verma and Mogra, 2013; Gonçalves and Romano, 2016; Franco et al., 2020; Cowley and Burton, 2021). For example, psyllium seeds and their seed husks do not contain gluten, so they can be used in making gluten-free products such as bread (Cappa et al., 2013; Franco et al., 2020). Gluten can damage people's health who have celiac disease and gluten sensitivity as undigested gluten can cause gut inflammation and trigger an immune response (Volta and De Giorgio, 2012). In addition, a growing population, especially young adults, believe and choose gluten-free products for healthier diets (Cappa et al., 2013; Christoph et al., 2018). It can be seen that the market size of gluten-free food keeps expanding and is expected to reach USD 8.3 billion in 2025 from 5.6 billion in 2020 (https://www.statista.com/statistics/248467/global-gluten-free-food-market-size/). India is the biggest grower and exporter of psyllium (Verma and Mogra, 2013; Kumar, 2015; Franco et al., 2020). Export of psyllium seeds and husks has fluctuated but shows an upward trend (Figure 1). Within a decade (2013-2022), total exports of husks and seeds increased by 1.35 times (39.7 to 53.8 billion tonnes) and 3.13 times (0.56 to 1.77 billion tonnes), respectively (Figure 1). Given the broad range of psyllium applications, the demand for psyllium is expected to increase. Here, the characteristics of *P. ovata*, current breeding program progress, and genetic information from model plants and other crops that could be used for future *P. ovata* genetic improvement, will be summarised.

Morphological characteristics and response to environmental conditions

P. ovata (Figures 2a-m) does not have a broad tolerance range of environmental conditions. Adverse climatic conditions have a high risk of disrupting the psyllium market. Ghaderi-Far et al. (2012) evaluated several factors affecting *P. ovata* seed germination and emergence. Psyllium seeds require a temperature between 10 and 25 ºC to reach maximum germination (84-92%), with 21.24 ºC being the optimum condition; no germinated seeds were observed at 35 ºC. The seeds can cope with salinity up to 200 mM before the germination rate declines and reaches 50% inhibition at 328 mM. Psyllium seeds germinated better in acidic soil with a pH between 4-6 (germination rate of 87-93%) compared to alkaline soil at pH 7-9 (52-61% germination). *P. ovata* tends to cope with water stress or drought as it needs an osmotic potential of -1.24 MPa to inhibit germination by 50% (Ghaderi-Far et al., 2012). Cowley et al. (2022) reported that seeds that experienced 26 mm rain before harvesting germinated slower than unaffected seeds. Karimzadeh and Omidbaigi (2004) observed that sowing date and nitrogen application influenced the growth and seed characteristics. They found that psyllium seeds were cold-sensitive, so early spring sowing should be avoided (Karimzadeh and Omidbaigi, 2004). Thus, unseasonal conditions may inhibit seed germination, increasing production costs.

The plant starts flowering at approximately 60 days from the sowing date (Dhar et al., 2005). Flowers are arranged in spikes (Figures 2a-b). Both male and female organs are in the same flower, which is a hermaphrodite (Dhar et al., 2005) (Figures 2b-c). The stigma (female reproductive organ) matures or becomes receptive earlier, at least by 14 hours, than pollen/anther dehiscence (male reproductive cells), which is called protogyny (Patel et al., 1980). Sharma et al. (1992) reported another type of pollination where synchronised stigma and anther maturation occur. Both types of plants exhibit the same degree of inbreeding or are highly self-pollinated (Sharma et al., 1992; Dhar et al., 2005). Sharma et al. (1992) observed a shifting in reproduction from cross- to self-pollination during *Plantago* domestication. This transition is common in domestication and found frequently in flowering plants; however, only

10-15% of seed plants are mainly self-fertilising (Wright et al., 2013). Wright et al. (2013) explained that selfing could be beneficial in the short term, such as offering reproductive assurance, which increases the chance of pollinating the stigma. However, it can lead to inbreeding depression in the long term, where reduced fitness or the lower ability to adapt by inbred progeny leads to higher extinction rates than outbred progeny (Wright et al., 2013). Therefore, genetic diversity is crucial in maintaining psyllium production in the future.

Genetic characteristics and breeding program strategies

To minimise the influence of environmental changes on psyllium production and quality and keep up with increasing demand, efforts have been made to create psyllium cultivars with less prone or non-shattering capsules, synchronous maturity resistance to biotic (pests and diseases) and abiotic (rain and cold) stresses, larger seed size, and higher husk content (Dhar et al., 2005; Patel et al., 2018). Several methods have been used in *P. ovata* breeding programs, including selection, hybridisation, induced mutation and polyploidy (Dhar et al., 2005; Lal et al., 2020). Among these methods, induced mutation is the most promising. Lal et al. (2020) recorded two successful varieties of *P. ovata* commercialised in India, namely Niharika, released in 1998 and Mayuri, released in 2007, both with characteristic high seed and seed husk yields. Mayuri also has an early maturing trait as a marker. Gamma radiation Co^{60} with 10-100 kR at 10 kR intervals and treatment with 0.2% EB (ethidium bromide) have been used in mutation experiments. The 0.2 EB + 20 kR treatment generated both cultivars Niharika and Mayuri (Lal et al., 2020). Tucker et al. (2017) also used Co^{60} to generate *P. ovata* mutants in Australia to obtain cultivars that suit local environments. Even though no cultivars have been released yet, Tucker et al. (2017) and Cowley et al. (2020) have shown diverse and promising mucilage related traits in putative mutants.

The genetic diversity of *P. ovata* landraces and mutants has been evaluated using RAPDs (Random Amplified Polymorphic DNA) in many studies (Pal and Raychaudhuri, 2003; Vahabi et al., 2008; Singh et al., 2009; Vala et al., 2011; Rohilla et al., 2012; Kaswan et al., 2013; Fougat et al., 2014; Kour et al., 2016). All eight studies show only slight genetic variation among the studied genotypes. For example, Singh et al. (2009) detected only 2-17% genetic variation among 80 *P. ovata* accessions even though geographically and morphologically they were distinct. The differences in morphology could be due to ecological adaptation or use of RAPD markers not able or sensitive enough to detect small changes in the DNA. Fougat et al. (2014) attempted to develop molecular markers for *P. ovata* molecular breeding. They sequenced GI-3 genomic DNA using ion torrent PGM^{TM} technology and assembled the data to generate 3,447 contigs (N50=1,346 bp) containing 249 Simple Sequence Repeats (SSR). They randomly selected 30 SSRs in six cultivars (GI-3, GI-2, RI-89, DPO13, DPO14, and EC124345) and six allies (*P. arteria*, *P. coronopus*, *P. psyllium*, *P. indica*, *P. serraria*, and *P. lanceolata*). These markers could differentiate at the genus level. Six clusters were generated from 12 species using these SSR markers. However, the results showed that all SRRs were monomorphic within *P. ovata*, meaning very low polymorphism (Fougat et al., 2014). All in all, RAPD and SSR markers show that *P. ovata* has limited genetic variability, possibly due to low chromosome number and small chromosome size, low chiasmata frequency, and a high proportion of heterochromatin (Dhar et al., 2002; Dhar et al., 2005; Dhar et al., 2006; Dhar et al., 2009). Commercial psyllium is diploid and only has eight chromosomes - two less than the model species Arabidopsis which has ten chromosomes. However, the *P. ovata* genome at ~500 Mb is three times larger than Arabidopsis (135 Mb). Dhar et al. (2009) observed a high heterochromatin portion in *P. ovata* chromosomes, predicted to be repeat regions.

As conventional breeding has not been encouraging for *P. ovata*, exploiting natural variation in the *Plantago* genus (Phan et al., 2016; Cowley et al., 2021) and utilising transgenic breeding or genetic engineering could be good options. Published molecular data from *P. ovata* that can be

utilised for the breeding program is limited. Only six studies generating RNAseq data and one for genomic data have been deposited in the NCBI database since 2010 (Table 1). *P. ovata* RNAseq data have been used to identify mucilage related genes (Jensen et al., 2011; Jensen et al., 2013; Jensen et al., 2014; Kotwal et al., 2016; Phan et al., 2016), flowering genes (Patel et al., 2020), and resistance genes against downy mildew (Ponnuchamy et al., 2020). All studies required *de novo* assembly to generate transcriptomic references as no *P. ovata* genome was available. In 2019, CSIR-Central Salt and Marine Chemicals Research Institute deposited genomic reads from the GI-20 cultivar, but no assembled nuclear genome is available yet.

Genetic basis of seed shattering from model species and other crops

Elimination of seed shattering traits is crucial in crop domestication before improving other aspects such as increasing starch or oil contents (Lin et al., 2007; Olsen, 2012). Several studies show that one or multiple genes can be responsible for acquiring shattering resistance by disrupting the abscission layer (AL) or dehiscence zone (DZ). In Arabidopsis, the silique dehiscence process involves the differentiation of specialised cell types (replum, valve, and dehiscence zone) and biochemical and molecular mechanisms are known (Figures 3h-k) (Ferrándiz, 2002; Dong and Wang, 2015). Pod shattering is also well characterised in legumes (Figures 3l-s) (Dong et al., 2014). For example, excessive deposition of secondary cell walls in FCC (fibre cap cells) caused pod shattering resistance in domesticated soybean (*Glycine max*) (Figures 3n-o) (Dong et al., 2014). Another example is the lack of a dehiscence zone (DZ) in the common vetch (*Vicia sativa*) shattering resistant accession, which prevents pod dehiscence and seed shedding (Figures 3r-s) (Dong et al., 2017).

Konishi et al. (2006) found that mutation in a QTL (quantitative trait locus) for seed shattering on chromosome 1 (*qSH1*) resulted in complete loss of the AL in a shattering resistant cultivar of rice, Nipponbare. The *qSH1* gene is an ortholog of *REPLUMLESS* (*RPL*) in Arabidopsis. Lin et al. (2007) showed that mutation in a single dominant gene, *shattering 1* (*sha-1*), leads to non-shattering rice. They found that mutation in *sha-1* was present in all 201 domesticated rice

cultivars (Japonica and Indica cultivars), but the mutation was not present in any of the 24 wild rice lines tested. Lv et al. (2018) reported that *SH3* (*Shattering 3*) and *SH4* (*Shattering 4*) controlled the degree of the shattering trait in *Oryza glaberrima* Steud. which is African domesticated rice. Both genes encode a transcription factor in the YABBY family. The loss of function of both genes leads to no abscission layer, while only one gene mutated leads to partial AL formation. *Shattering 1* (*SHA-1*) described by Lin et al. (2007) and *Shattering 4* (*SH4*) noted by Lv et al. (2018) are referred to as the exact location in chromosome 4 of rice. The earlier study suggested that *SH4* has a role in the establishment of AL at early-stage flower development and may activate the abscission process at a later stage (Li et al., 2006). However, Lin et al. (2007) reported that mutation in *SHA-1* did not eliminate the abscission layer (AL). They predicted that *SHA-1* is involved in cell wall degradation in AL instead of governing AL formation (Lin et al., 2007). Olsen (2012) showed that mutations in the *sh1* gene encoding a YABBY transcription factor resulted in a non-shattering seed phenotype in domesticated sorghum. Interestingly, *Sh1* (*Shattering1*) in sorghum is an ortholog of *SH3* in *Oryza glaberrima* (Li et al., 2020). Li et al. (2020) predicted that *Sh1* is likely to function downstream of *qSH1*. *SH4* regulates *SHATTERING ABORTION1* (*SHAT1*), encoding APETALA2 (Zhou et al., 2012). *SH5* (homologous to *qSH1*) induces *SH4* and *SHAT1* (Yoon et al., 2014). In summary, *qSH1*, *SH3/Sh1*, *SH4/SHA1*, *SHAT1*, and *SH5* are all seed shattering genes that promote abscission layer formation between pedicle and spikelet.

Unlike rice, the Arabidopsis dehiscence zone (DZ) consists of a lignified layer (LL) and a separation layer (SL) formed between the replum and valves (Ferrándiz, 2002). The formation of each part is controlled by a well-coordinated complex genetic regulatory network (Figure 4A). Valve identity is regulated by *FRUITFULL* (*FUL*/*AGL8*) (Gu et al., 1998), while replum development is influenced by *REPLUMLESS* (*RPL*) (Roeder et al., 2003). Dehiscence zones, both LL and SL, are regulated by *INDEHISCENT* (*IND*) (Liljegren et al., 2004; Dong and Wang, 2015). SL is also under the regulation of *ALCATRAZ* (*ALC*) (Rajani and Sundaresan,

2001; Dong and Wang, 2015). Both *IND* and *ALC* are positively regulated by *SHATTERPROOF1/2* (*SHP1/2*) (Liljegren et al., 2004; Dong and Wang, 2015). *RPL* and *FUL* repress the expression of *IND*, *ALC* and *SHP1/2* (Pabón-Mora et al., 2014; Dong and Wang, 2015). The expression of *RPL* and *SHP1/2* is negatively regulated by *APETALA2* (*AP2*) (Dong and Wang, 2015). Downstream of *SHP1/2* cascade, *ARABIDOPSIS DEHISCENCE ZONE POLYGALACTURONASE1/2* (*ADPG1/2*) is expressed and required for pectin degradation in the SL, while *NST1* is expressed explicitly in LL to promote lignification of valve margins (Dong and Wang, 2015). *IND* and *ALC* control *ADPG1/2*, while *NST1/3* is regulated by *IND* (Dong and Wang, 2015). Loss of function in the double mutant *shp1/2*, *nst1*, and *adpg1/2* still generated shattered pods while individual mutants *repl* and *alc* induced partially indehiscent pods (Dong and Wang, 2015; Di Vittori et al., 2019). In contrast, the *ful* mutant led to a premature bursting pod (Gu et al., 1998; Dong and Wang, 2015). Another player in shattering is *SEEDSTICK* (*STK/AGL11*), where the seed failed to detach, and no clear DZ was observed in the *stk* mutant (Mizzotti et al., 2012).

The polysaccharides cellulose and hemicellulose, plus lignin, were highly correlated with the pod shattering resistance index (SRI) in various rapeseed varieties (*Brassica napus* L.) (Kuai et al., 2016). Hemicellulose was the consistent physiological indicator affecting pod shattering resistance among these three secondary cell wall components (Kuai et al., 2016). Indeed, lignin deposition can both promote and inhibit shattering. In rice, *SH5* causes grain shattering by repressing lignin deposition in the pedicel region (Yoon et al., 2017). In Arabidopsis, *NST1* and *NST3* promote lignin deposition in DZ, leading to shattering (Mitsuda and Ohme-Takagi, 2008). In *Glycine max* (soybean), *SHAT1-5*, homologous to *NST1/2* in Arabidopsis, promotes lignification or cell wall thickening of fibre cap cells (FCC) (Dong et al., 2014). Biosynthesis of secondary cell wall components such as lignin, cellulose, and hemicellulose, share common regulatory genes, including previously described *NST1* (Figure 4B) (Nakano et al., 2015; Taylor-Teeples et al., 2015).

Shattering in *Plantago ovata*

P. ovata has a dehiscent dry fruit called a capsule with a bilocular ovary that houses the two developing seeds (Figures 2d-g and 3a) (Lamba and Veena, 1981; Phan et al., 2020). Each locule contains one ovule (developing seed) borne in axile placentation (Figures 2f and 3a) (Lamba and Gupta, 1981). At maturity, the capsule splits at the dehiscence zone and then releases or shatters its seeds when the lid or operculum comes off (Figures 2e and 3g) (Lamba and Gupta, 1981). Seed shattering can be triggered by unseasonal or heavy rain and strong winds, and it is responsible for catastrophic seed losses (Dhar et al., 2005; Patel et al., 2018; Cowley et al., 2022). Rain also affects seed husk quality. Cowley et al. (2022) observed that rain-damaged seeds were darker and greener. They also stated that rain could affect the seeds because the capsule has a dehiscence zone that exposes the seed to the environment. The seed coat/husk experienced premature hydration when the humidity or water was trapped or prevented from evaporating, eventually damaging the seed and mucilage quality after rain (Cowley et al., 2022).

Even though shattering is responsible for yield losses, the dehiscence or shattering mechanism in *P. ovata* has not been fully described or understood. General development of the dehiscence zone in psyllium has been described by Lamba and Gupta (1981). They have described the following events in development and dehiscence of the capsule. The *P. ovata* ovary comprises four layers of parenchyma cells, with the innermost layer smaller than the outer layers (Figure 3b). Cell division occurs only in the middle of the ovary, in the meristematic zone (Figure 3c). The outer three cell layers elongate while the inner layer gets compressed (Figure 3d). The meristematic zone becomes the future dehiscence zone (Figure 3e). An increase in cell wall thickening was only observed in the outer epidermis except in the future dehiscence region (Figure 3f). In a mature capsule, lignification occurs in the cell wall of the outer two layers creating a very thick cuticle (Figure 3g). Finally, the lid (upper capsule) separates from the base (bottom capsule) through the dehiscence zone facilitated by the pressure from mature seeds

(Lamba and Gupta, 1981). However, these observations have not been updated since 1981, although microscopy techniques have been greatly improved since then.

Genetic basis of mucilage polysaccharide production from model species

P. ovata seed husk/mucilage accounts for only 25% of the seed weight (Kumar, 2015) but is the most economically valuable fraction. The accumulation of mucilage polysaccharides in the seed was first thoroughly described by Hyde (1970). Phan et al. (2020) then comprehensively explored mucilage polysaccharide synthesis, deposition, desiccation, and release processes. Mucilage polysaccharides accumulate in the outer layer of the ovule integument called mucilage secretory cells (MSCs). The other integument layers become crushed and progressively compressed as the embryo expands and are hardly visible in the mature seeds. Cowley and Burton (2021) emphasised the different mucilage-related features of *P. ovata* from Arabidopsis and *Linum usitatissimum* (flax). Mature *P. ovata* seeds do not have an intact MSC layer as the cell walls also degrade as the seed matures, leaving the desiccated mucilage polysaccharides in place on the seed outer surface as the husk, so there is no cellular rupture during rehydration and expansion of the mucilage, as seen in the classic model from Arabidopsis. In fact, neither *P. ovata* nor flax possess a columella in the mature MSC. Cowley and Burton (2021) illustrated *P. ovata* mucilage secretory cell (MSC) differentiation and mucilage production in Figure 5.

Many genes encoding transcription factors (TFs) and enzymes controlling mucilage polysaccharide production have been intensively studied in Arabidopsis (Western et al., 2001; Western et al., 2004; Gonzalez et al., 2009; Arsovski et al., 2010; Saez-Aguayo et al., 2013; North et al., 2014; Voiniciuc et al., 2015; Golz et al., 2018). Voiniciuc et al. (2015) and Golz et al. (2018) have suggested different ways to group these genes. Voiniciuc et al. (2015) have six groups sequentially corresponding to developmental stages (Figure 6). Group 1, belonging to outer ovule integument development and group 2 defining seed coat epidermal cell differentiation, both contain transcription factors, while groups 3 to 6 are enzymes or other

proteins. Group 3 enzymes involved in synthesis of mucilage components were further separated into pectin, hemicellulose, surface proteins and cellulose. Groups 4 and 5 included mucilage secretion and wall modification, respectively. Genes in the last group, corresponding to columella synthesis may have different functions in other species that do not have a columella, like *P. ovata* (Figure 6). Golz et al. (2018) assigned transcription factors into a hierarchy, with tiers 1-3 according to mutant phenotypes, relationship with other regulators and known targets (Figure 7). Tier 3 contains early-acting TFs that were activated following fertilisation. TFs in tier 3 are NARS1/NARS2, AP2 and the transcriptional complex TTG1, MYB5, TT2, EGL3, TT8, and MYB23. Activation of these TFs promote differentiation of the outermost integument layer, driving distinct MSC morphology and mucilage formation. Severe defects in epidermal differentiation were observed in these mutants. MYB52, MYB61, GL2 and TTG2 were placed in tier 2. The third layer (tier 1) contains DF1, SHP1, SHP2, STK, MUM/LUH, and KNAT7. Unlike Voiniciuc et al. (2015), Golz et al. (2018) did not separate genes for mucilage production/synthesis, secretion, modification, and adhesion. The model from Golz et al. (2018) included additional information about the gibberellic acid (GA) pathway in seed coat development and mucilage production.

Regulatory genes involved in mucilage polysaccharide biosynthesis also control other pathways, such as those for tannins and fatty acids (FA). Shi et al. (2012) generated *ttg1*, *tt2*, *myb5*, *egl3*, *tt8*, *gl2*, *mum4* and *ttg2* mutants and found that only *ttg2* did not have an increased seed oil content compared to the wild type. This finding was also reported by Chen et al. (2012) and Wang et al. (2014) for *tt2*, Chen et al. (2015) and Li et al. (2018) for *ttg1*, and Shen et al. (2006) and Shi et al. (2012) for *gl2* and *mum4*. Li et al. (2018) proposed that phosphorylation of *TTG1* by *SK11/SK12* (*SHAGGY-like kinases 11/12*) at serine 215 prevents interaction between *TTG1* and *TT2* and decreases the transcript of *GL2*, so fatty acid biosynthesis increases. However, the production of mucilage polysaccharides and tannin is inhibited. However, Johnson et al. (2002) showed that the tannin level in *gl2* is similar to the wild type. *TTG1*

probably interacts with *TTG2*, not *GL2*, in affecting pigmentation. Other studies show that tannin production increased in seeds carrying *ttg1* (Debeaujon et al., 2003), *tt2* (Johnson et al., 2002; Debeaujon et al., 2003; Doughty et al., 2014), and *ttg2* (Johnson et al., 2002; Debeaujon et al., 2003). All these studies suggest that some mucilage regulatory genes can act as negative regulators for FA biosynthesis and positive regulators for PA biosynthesis.

Mucilage properties are determined by polysaccharide composition and their molecular structures, particularly substitution (Phan et al., 2016; Cowley and Burton, 2021). The primary components of seed mucilage are pectins and heteroxylans, and their relative contents differ among species (Western et al., 2000; North et al., 2014; Cowley and Burton, 2021). *Arabidopsis* mucilage mainly comprises pectic unbranched RG I (Naran et al., 2008; Arsovski et al., 2010; Cowley and Burton, 2021), while *P. ovata* is rich in complex heteroxylans (Fischer et al., 2004; Guo et al., 2008). Heteroxylan is composed of a backbone of xylose residues decorated with a variety of side chains typically comprised of arabinose (Ara), xylose (Xyl), glucuronic acid (GlcA), and traces of other sugars (Fischer et al., 2004; Ebringerová, 2005; Yu et al., 2017; Cowley and Burton, 2021). Genes encoding enzymes for xylan backbone formation belong to two families, which are glycosyltransferase (GT) 43 (*IRX 9*, *IRX 9L*, *IRX14*, and *IRX14 like(L)*) and GT47 (*IRX10 and IRX10L*) (Jensen et al., 2013; Jensen et al., 2014). The expression of genes in these families is variable among species and even in different parts of the same plant. For example, the expression level of *IRX10/IRX10L* is higher in *Plantago* seed integuments, but very low in *Plantago* stem. In contrast, *Arabidopsis* seeds show high expression levels of *IRX14/IRX14L* but low expression of *IRX10/IRX10L* (Jensen et al., 2013). Candidate genes for arabinose and xylose substitution belong to the GT61 family, with three clades defined (Anders et al., 2012). Clade A is likely involved in arabinosyltransferase activity (arabinose substitution). Clade C in xylosyltransferase activity and enzymes in clade B are currently uncharacterized (Figure 8) (Anders et al., 2012; Voiniciuc et al., 2015). Phan et al. (2016) identified *GT61* genes co-expressed with *IRX10* in seed tissues of *P. ovata* and *P. cunninghamii*.

However, they did not determine whether these GT61 proteins act as xylosyltransferases and/or arabinosyltransferases.

Conclusions

Psyllium seeds and husks are in high demand as they are used in many applications, especially for health and industrial products. Adverse environmental conditions are responsible for yield losses by increasing shattering events and decreasing husk/mucilage quality. Many factors that limit the genetic improvement of *P. ovata* include

- a lack of genetic diversity among *P. ovata* cultivars,
- lack of fundamental knowledge about seed and capsule development and associated shattering mechanisms,
- lack of a reference genome, and
- limited publicly available genomic and transcriptomic data.

Proposed study

Data mining, bioinformatics, microscopy, chemical, and transcriptomic analysis were performed to investigate *P. ovata* shattering and mucilage polysaccharide production mechanisms. Publically available genomic and transcriptomic data from the Sequence Read Archive (SRA) at the National Center for Biotechnology Information (NCBI) and unpublished data from the long-term *P. ovata* project at the University of Adelaide were collected and evaluated. These data were used to *de novo* assemble and annotate the *P. ovata* genome. This genome was deposited into SRA NCBI and used as a reference genome for two RNAseq experiments. Several databases were generated from this genome, including lists of ribosomal RNA for depleting rRNA in RNAseq experiments and GO term and KEGG identifiers for enrichment analysis and pathway identification. Capsules and seeds from wild type and selected putative gamma-irradiated mutants at specific developmental time points and tissue types were

compared using microscopic and chemical analysis to identify phenotypic differences in capsule and early seed development. Capsule RNAseq data were obtained from whole tissues, while RNAseq data from developing seeds were generated by combining cryosectioning and laser capture microdissection. Transcriptomic datasets were used to identify candidate genes associated with shattering and mucilage polysaccharide production. The assembled genome, candidate mutants, and identified genes and networks will be useful as targets for *P. ovata* genetic improvement and for studying cell wall organisation in crops and other non-model species.

Table 1: Summary of RNA-seq and genomic data from *P. ovata*.

Submitted year	Institution	Cultivar	Platform	Source	Paired/	Accession number
2010	DOE JOINT GENOME INSTITUTE (JGI)	Indian, *P. ovata*, Sand Mountain Herbs, AL, USA	1 LS454 (454 GS FLX)	Transcriptomic	Single	SRR066373
						SRR066374
						SRR066375
						SRR066376
2011	Research Technology Support Facility at Michigan State University (MSU-RTSF)	Indian, *P. ovata*, Sand Mountain Herbs, AL, USA	1 LS454 (454 GS FLX)	Transcriptomic	Single	SRR342350
						SRR342351
2012	University of Jammu	JUSBT-1	Illumina Genome Analyzer II	Transcriptomic	Paired	SRR629688
2013	Anand Agricultural University	EC-124345, Niharika	1 LS454 (454 GS FLX Titanium)	Transcriptomic	Single	SRR1311174
						SRR1311175
						SRR1311176
						SRR1311177
2016	The University of Adelaide	Indian P. ovata	1 ILLUMINA (Illumina HiSeq 2000)	Transcriptomic	Single	SRR3883622
						SRR3883620
						SRR3883621
						SRR3883618
						SRR3883619
						SRR3885726
						SRR3885727
						SRR3885728
2019	ICAR-Directorate of Medicinal and Aromatic Plants Research	DPO-185, DPO-14	1 ILLUMINA (NextSeq 500)	Transcriptomic	Paired	SRR5434206
						SRR5434207
						SRR5434208
						SRR5434209
						SRR5434211
						SRR5434210
						SRR5434213
						SRR5434212
2019	CSIR-Central Salt and Marine Chemicals Research Institute	GI-20	HiSeq X Ten	Genomic	Paired	SRR10076762

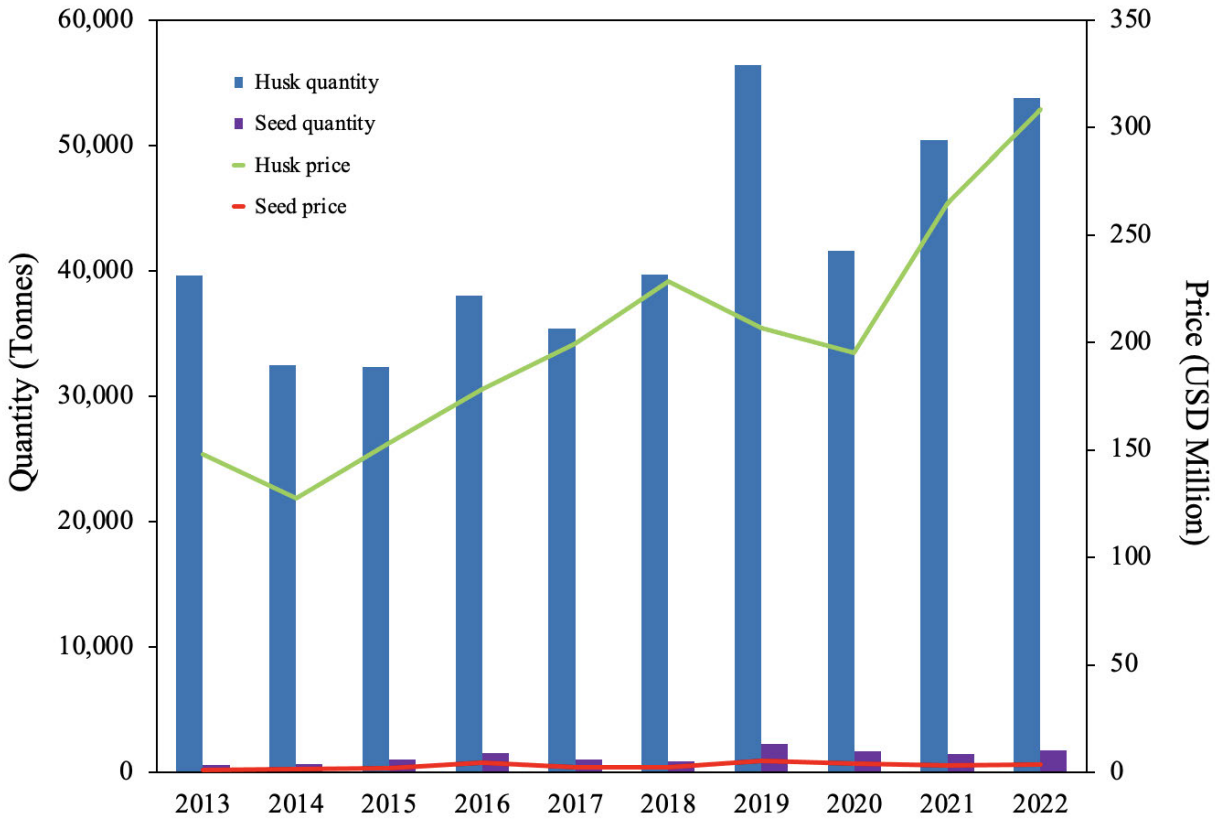


Figure 1. Export quantity and value of psyllium seeds and husks across the last decade (Adapted from Govt. India, Dept. of Commerce 2022).

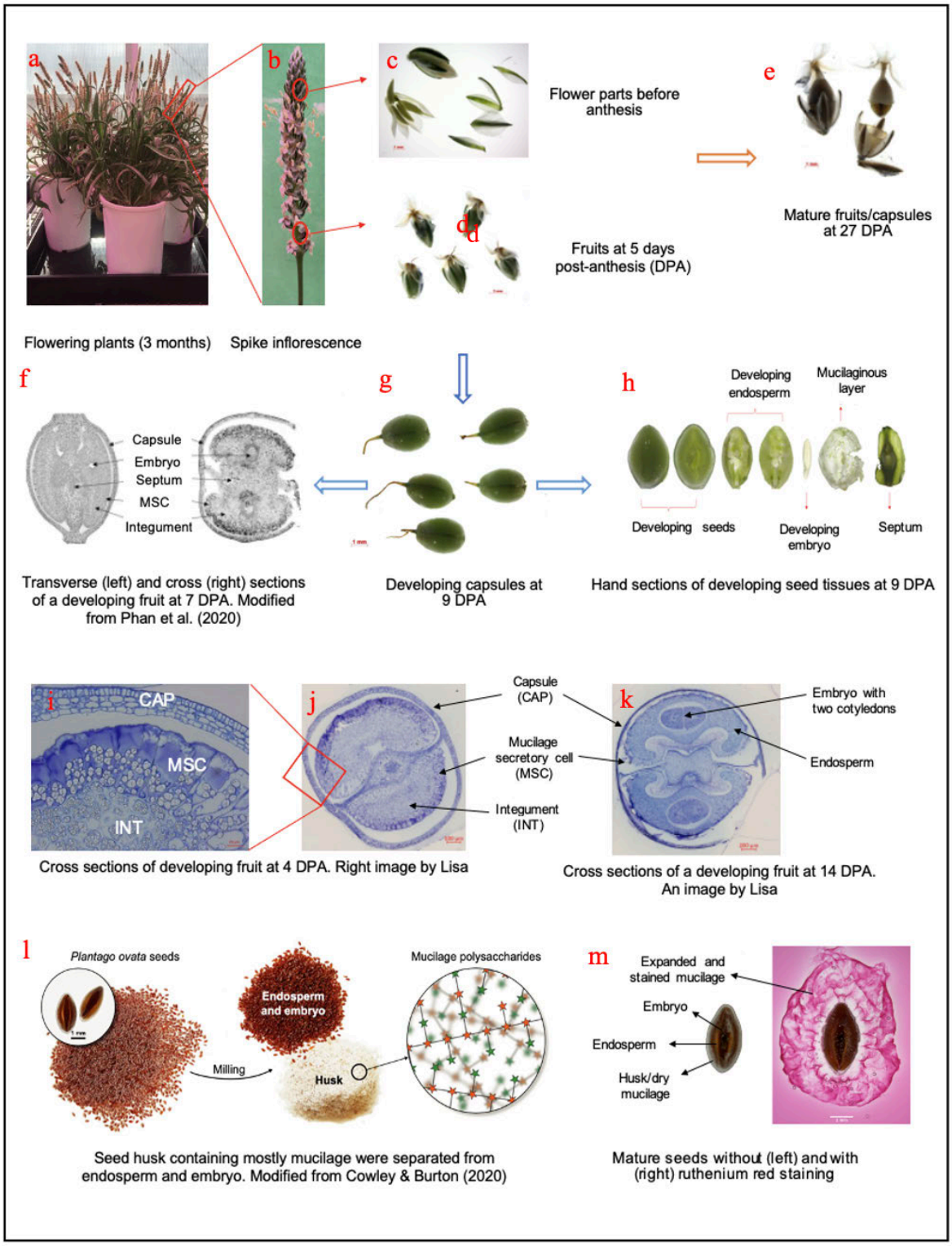


Figure 2. Morphology and anatomy of *Plantago ovata*.

Abbreviations DPA= day post-anthesis; CAP = capsule; MSC = mucilage secretory cell; INT = integument. Scale bar c = 1 mm; d = 2 mm; e = 1 mm; g = 1 mm; i = 20 mm; j = 100 mm; k = 200 mm; l-m = 1 mm. Image f was modified from Phan et al. (2020) and image l was modified from Cowley & Burton (2020).

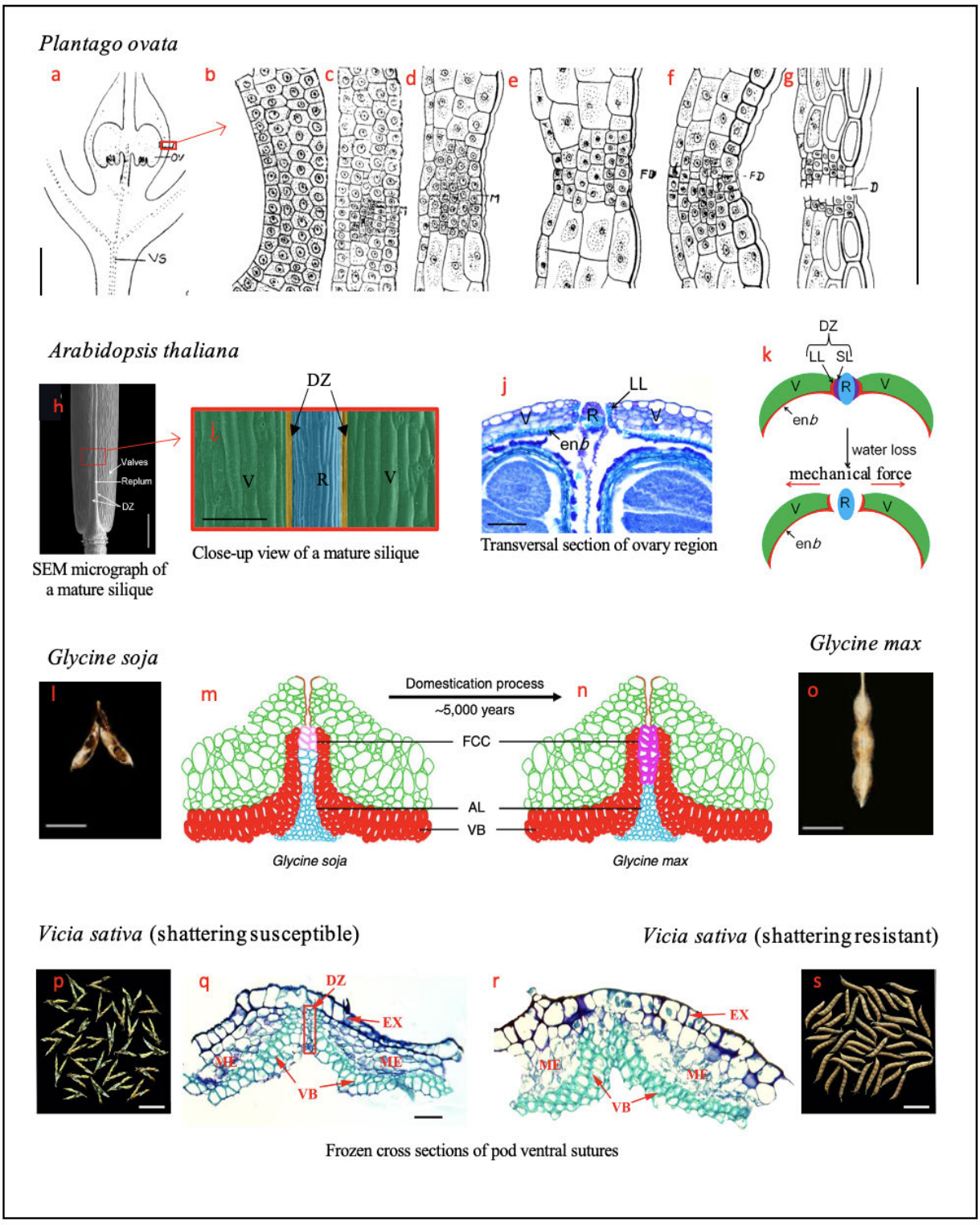


Figure 3. Anatomy of fruits (capsule/silique/pod) as related to shattering ability in six species. *Plantago ovata* has two ovaries fused (syncarpous) (a). The development of the dehiscence zone (DZ) in psyllium is shown in a-g. The main features of mature siliques in Arabidopsis are valves, the replum and DZ (h-j) and a model of the dehiscence process (k). Excessive deposition of the secondary wall in FCC led to pod shattering resistance in the legume species *Glycine*

max compared to the wild progenitor *Glycine soja* (l-o). The shattering resistant cultivar of *Vicia sativa* does not have a dehiscence zone (DZ) like as in shattering susceptible cultivar (p-s).

Abbreviations OV = ovule; VS = vascular supply; M = meristematic zone; FD = future dehiscence; D = dehiscence line; SEM = scanning electron microscope; DZ = dehiscence zone; en*b* = endocarp *b* layer; LL = lignified layer; R = replum; SL = separation layer; V = valves; FCC = fibre cap cells; AL = abscission layer; VB = vascular bundle; EX = exocarp; ME = mesocarp).

Scale bar a = 0.1 mm; b-g = 0.1 mm; h = 1.5 mm; i-j = 80 mm; l-o = 1 cm; p-s = 5 cm; q-r = 100 mm.

Modified from Lamba and Gupta (1981); Dong et al. (2014); Dong and Wang (2015); Dong et al. (2017).

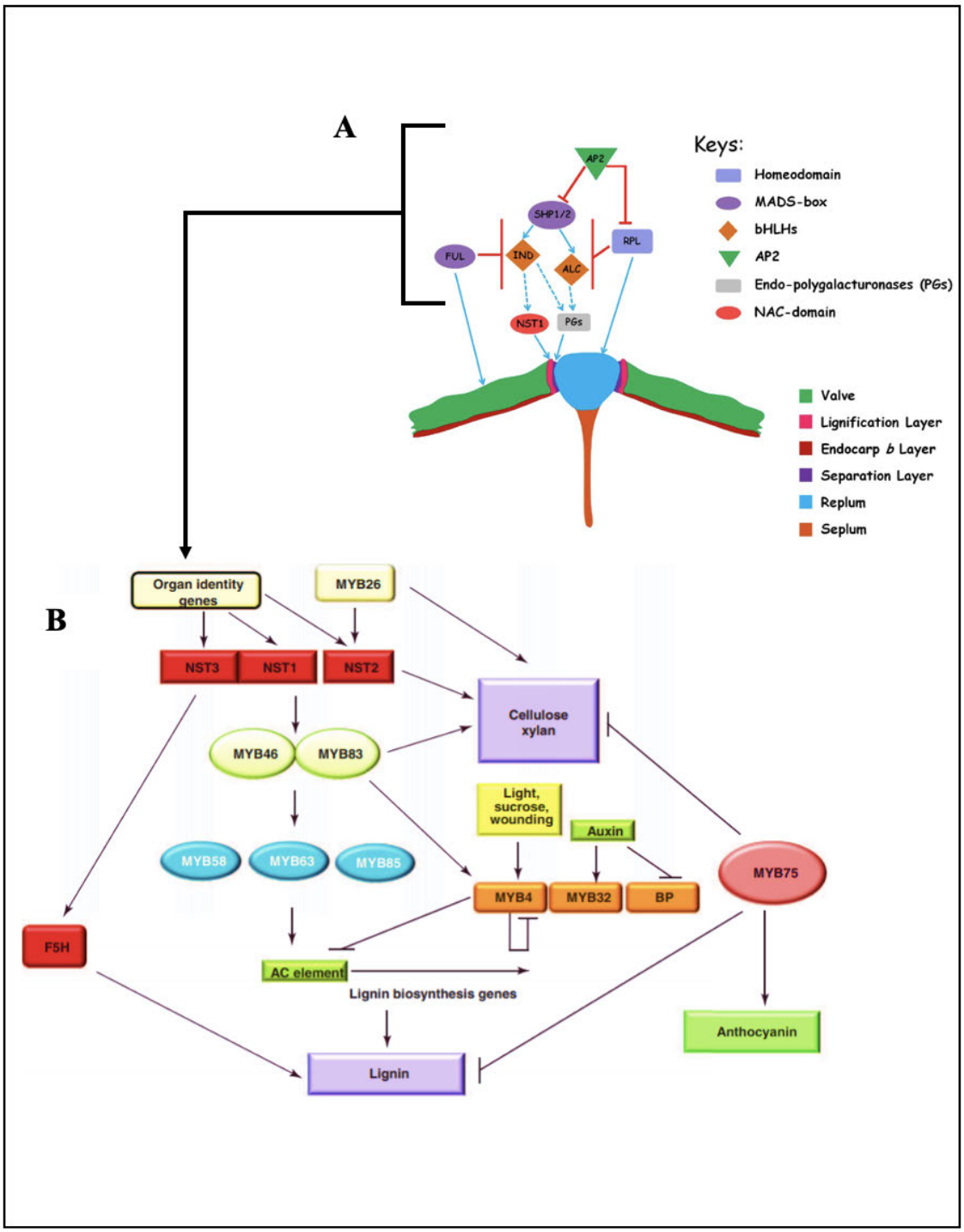


Figure 4. Transcriptional networks are involved in organ identity (A) and secondary cell wall biosynthesis (B) in *A. thaliana*. Modified from Dong and Wang (2015) and Zhao and Dixon (2011).

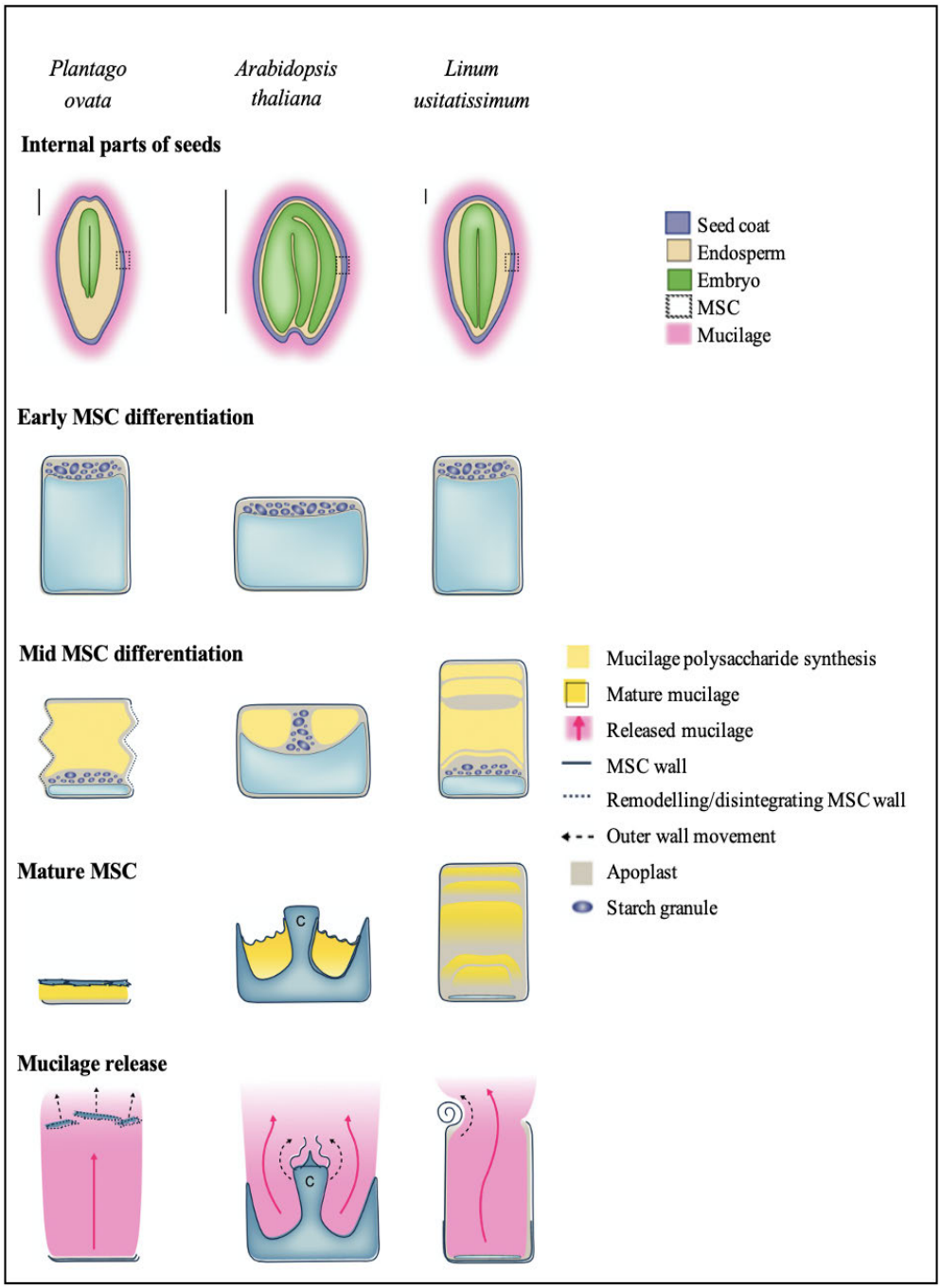


Figure 5. Comparison of seed features, mucilage secretory cell (MSC) differentiation and mucilage production in *Plantago ovata*, *Arabidopsis thaliana*, and *Linum usitatissimum.* All three species being compared are dicots; however, there are variations in endosperm and embryo proportion relative to their seed sizes. *P. ovata* retains a large portion of endosperm in mature seeds. The location of MSC is on the external seed coat surface. At the early stage, MSC features are similar between each species. All three species exhibit differences in mature MSC and mucilage release stages. C = Columella. Modified from Cowley and Burton (2021).

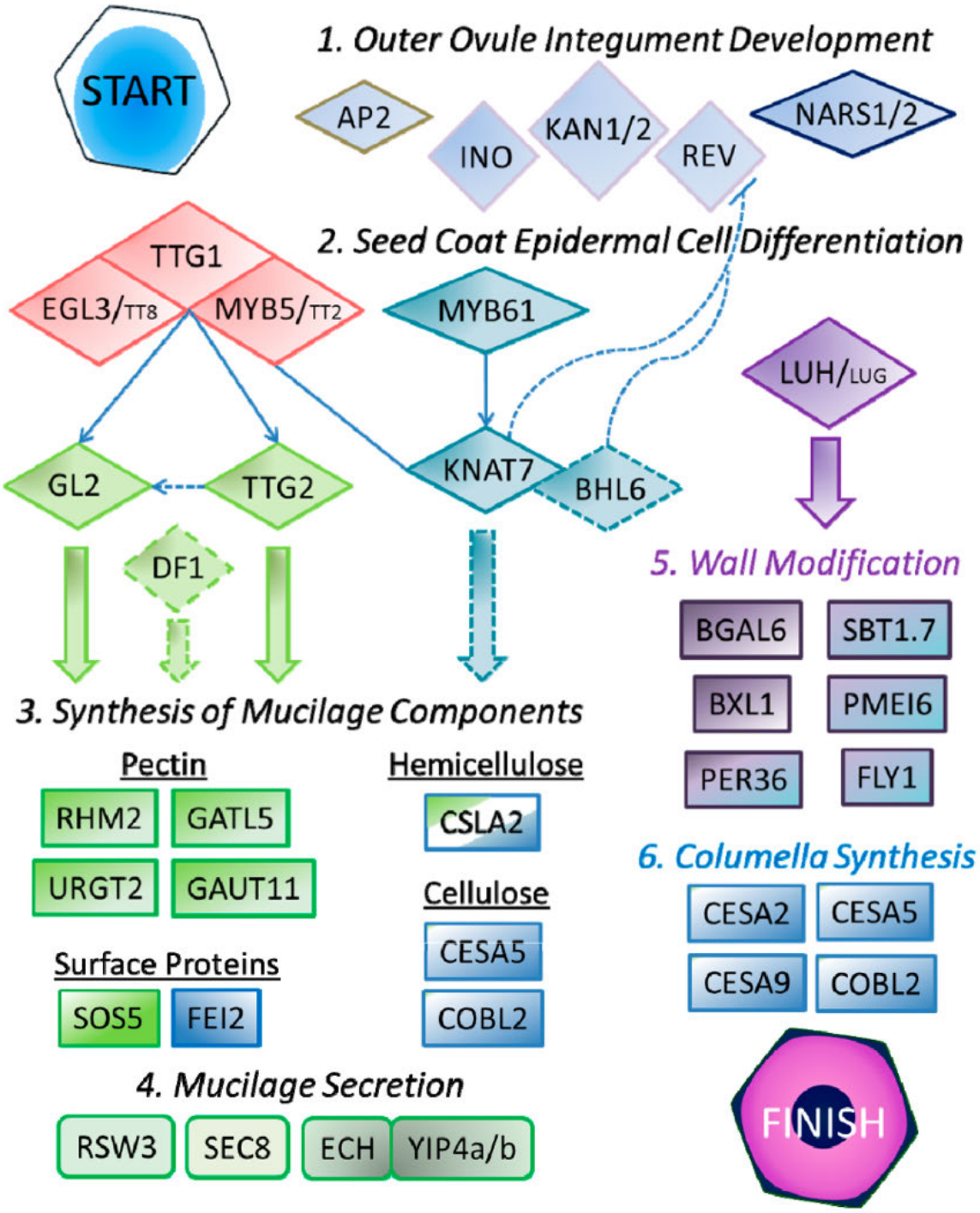


Figure 6. A transcriptional network regulating mucilage secretory cell (MSC) differentiation and mucilage production in *A. thaliana* was proposed by Voiniciuc et al. (2015).

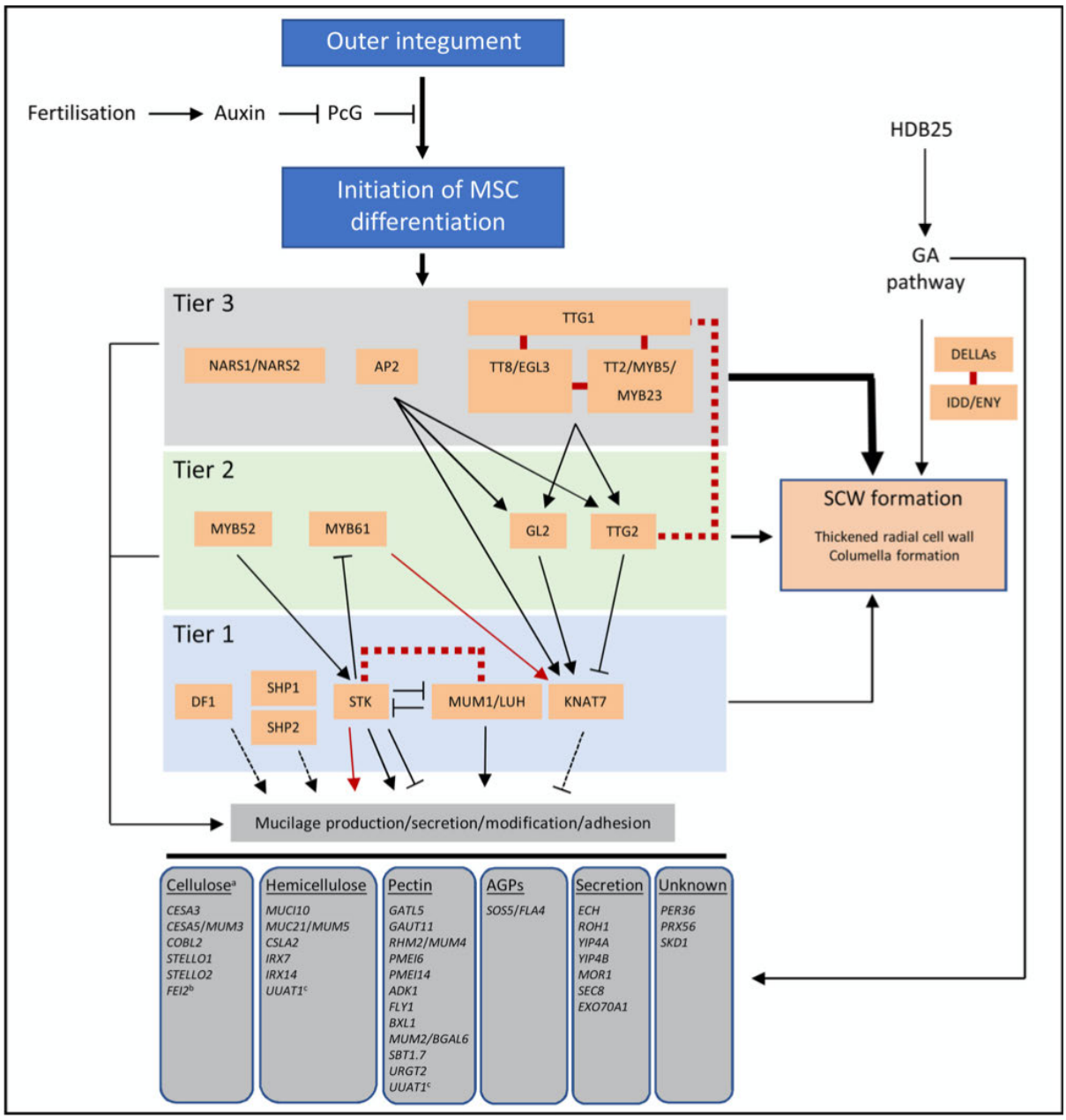


Figure 7. A transcriptional network regulating Arabidopsis mucilage secretory cell (MSC) differentiation and mucilage production was proposed by Golz et al. (2018).

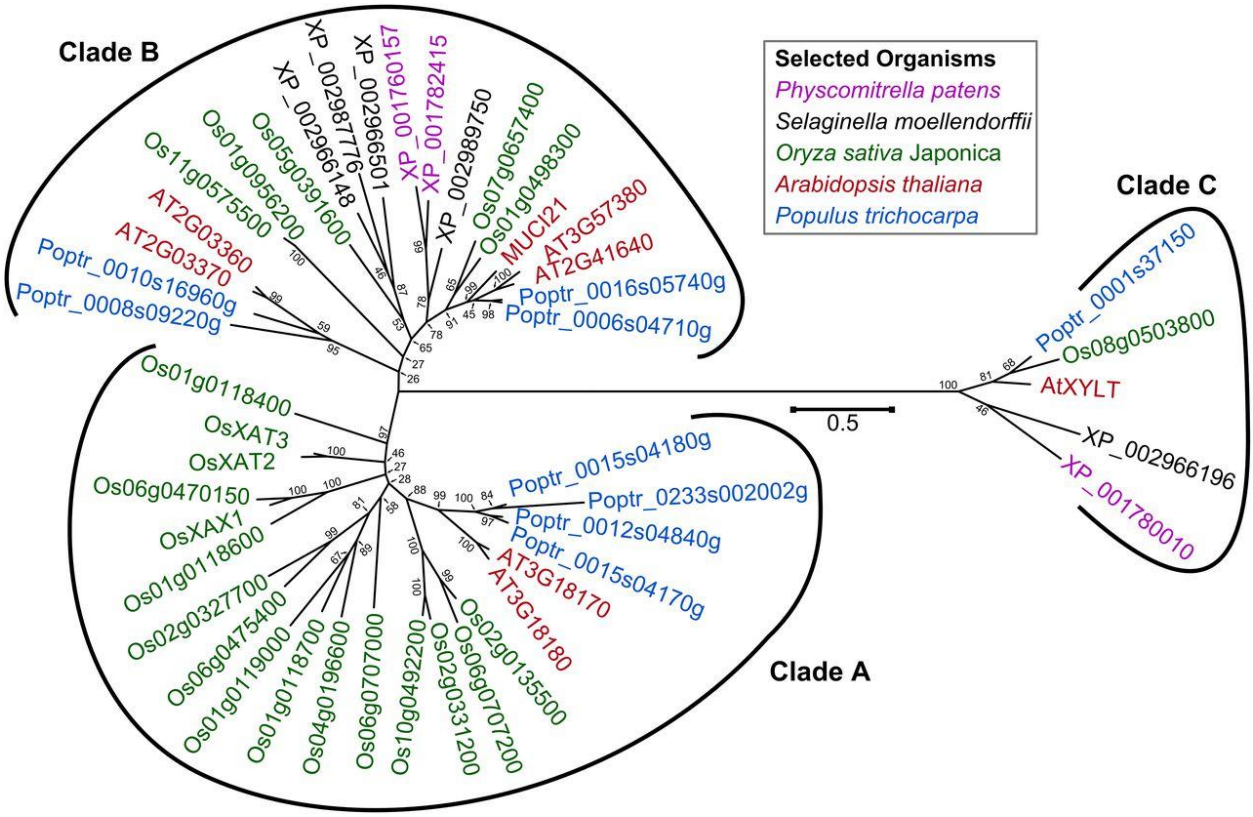


Figure 8. A phylogenetic tree of GT61 proteins from five species as shown in the box. Clade A has arabinosyltransferase activity, clade B has not been characterised, and clade C has xylosyltransferase activity (Voiniciuc et al., 2015).

References

Anders N, Wilkinson MD, Lovegrove A, Freeman J, Tryfona T, Pellny TK, Weimar T, Mortimer JC, Stott K, Baker JM, Defoin-Platel M, Shewry PR, Dupree P, Mitchell RAC (2012) Glycosyl transferases in family 61 mediate arabinofuranosyl transfer onto xylan in grasses. Proceedings of the National Academy of Sciences **109:** 989-993

Arsovski AA, Haughn GW, Western TL (2010) Seed coat mucilage cells of *Arabidopsis thaliana* as a model for plant cell wall research. Plant Signaling & Behavior **5:** 796-801

Cappa C, Lucisano M, Mariotti M (2013) Influence of Psyllium, sugar beet fibre and water on gluten-free dough properties and bread quality. Carbohydrate polymers **98:** 1657-1666

Chen M, Wang Z, Zhu Y, Li Z, Hussain N, Xuan L, Guo W, Zhang G, Jiang L (2012) The Effect of TRANSPARENT TESTA2 on seed fatty acid biosynthesis and tolerance to environmental stresses during young seedling establishment in Arabidopsis. Plant Physiology **160:** 1023

Chen M, Zhang B, Li C, Kulaveerasingam H, Chew FT, Yu H (2015) TRANSPARENT TESTA GLABRA1 regulates the accumulation of seed storage reserves in *Arabidopsis*. Plant Physiology **169:** 391-402

Chen M, Zhang B, Li C, Kulaveerasingam H, Chew FT, Yu H (2015) TRANSPARENT TESTA GLABRA1 regulates the accumulation of seed storage reserves in Arabidopsis. Plant Physiol **169:** 391-402

Christoph MJ, Larson N, Hootman KC, Miller JM, Neumark-Sztainer D (2018) Who values gluten-free? Dietary intake, behaviors, and sociodemographic characteristics of young adults who value gluten-free food. Journal of the Academy of Nutrition and Dietetics **118:** 1389-1398

Cowley JM, Burton RA (2021) The goo-d stuff: *Plantago* as a myxospermous model with modern utility. New Phytol **229:** 1917-1923

Cowley JM, Herliana L, Neumann KA, Ciani S, Cerne V, Burton RA (2020) A small-scale fractionation pipeline for rapid analysis of seed mucilage characteristics. Plant Methods **16:** 1-12

Cowley JM, McNeil DL, Lui KY, Barsby JP, Ciani S, Cerne V, Burton RA (2022) Rain events at maturity severely impact the seed quality of psyllium (*Plantago ovata* Forssk.). Journal of Agronomy and Crop Science

Cowley JM, O'Donovan LA, Burton RA (2021) The composition of Australian *Plantago* seeds highlights their potential as nutritionally-rich functional food ingredients. Sci Rep **11:** 12692

Debeaujon I, Nesi N, Perez P, Devic M, Grandjean O, Caboche M, Lepiniec L (2003) Proanthocyanidin-accumulating cells in Arabidopsis testa: regulation of differentiation and role in seed development. Plant Cell **15:** 2514-2531

Dhar M, Fuchs J, Houben A (2009) Distribution of Eu-and heterochromatin in *Plantago ovata*. Cytogenetic and genome research **125:** 235-240

Dhar M, Kaul S, Friebe B, Gill B (2002) Chromosome identification in *Plantago ovata* Forsk. through C-banding and FISH. Current Science**:** 150-152

Dhar M, Kaul S, Sareen S, Koul A (2005) *Plantago ovata*: Genetic diversity, cultivation, utilization and chemistry. Plant Genetic Resources: characterization and utilization **3:** 252-263

Dhar MK, Friebe B, Kaul S, Gill BS (2006) Characterization and physical mapping of ribosomal RNA gene families in Plantago. Annals of botany **97:** 541-548

Di Vittori V, Gioia T, Rodriguez M, Bellucci E, Bitocchi E, Nanni L, Attene G, Rau D, Papa R (2019) Convergent Evolution of the Seed Shattering Trait. Genes **10**

Dong R, Dong D, Luo D, Zhou Q, Chai X, Zhang J, Xie W, Liu W, Dong Y, Wang Y, Liu Z (2017) Transcriptome analyses reveal candidate pod shattering-associated genes involved in the pod ventral sutures of common vetch (*Vicia sativa* L.). Frontiers in Plant Science **8:** 649

Dong Y, Wang Y-Z (2015) Seed shattering: from models to crops. Frontiers in Plant Science **6:** 476

Dong Y, Yang X, Liu J, Wang B-H, Liu B-L, Wang Y-Z (2014) Pod shattering resistance associated with domestication is mediated by a NAC gene in soybean. Nature Communications **5:** 3352

Doughty J, Aljabri M, Scott RJ (2014) Flavonoids and the regulation of seed size in Arabidopsis. Biochem Soc Trans **42:** 364-369

Ebringerová A (2005) Structural diversity and application potential of hemicelluloses. Macromolecular Symposia **232:** 1-12

Ferrándiz C (2002) Regulation of fruit dehiscence in Arabidopsis. Journal of Experimental Botany **53:** 2031-2038

Fischer MH, Yu N, Gray GR, Ralph J, Anderson L, Marlett JA (2004) The gel-forming polysaccharide of psyllium husk (*Plantago ovata* Forsk). Carbohydrate Research **339:** 2009-2017

Fougat RS, Joshi C, Kulkarni K, Kumar S, Patel A, Sakure A, Mistry J (2014) Rapid development of microsatellite markers for *Plantago ovata* Forsk.: using next generation sequencing and their cross-species transferability. Agriculture **4:** 199-216

Franco EAN, Sanches-Silva A, Ribeiro-Santos R, de Melo NR (2020) Psyllium (*Plantago ovata* Forsk): From evidence of health benefits to its food application. Trends in Food Science & Technology **96:** 166-175

Ghaderi-Far F, Alimagham S, Kameli A, Jamali M (2012) Isabgol (*Plantago ovata* Forsk) seed germination and emergence as affected by environmental factors and planting depth. Journal of Agricultural Science and Technology **6:** 185-194

Golz JF, Allen PJ, Li SF, Parish RW, Jayawardana NU, Bacic A, Doblin MS (2018) Layers of regulation - Insights into the role of transcription factors controlling mucilage production in the Arabidopsis seed coat. Plant Sci **272:** 179-192

Golz JF, Allen PJ, Li SF, Parish RW, Jayawardana NU, Bacic A, Doblin MS (2018) Layers of regulation - Insights into the role of transcription factors controlling mucilage production in the Arabidopsis seed coat. Plant Science **272:** 179-192

Gonçalves S, Romano A (2016) The medicinal potential of plants from the genus *Plantago* (Plantaginaceae). Ind Crops Prod **83:** 213-226

Gonzalez A, Mendenhall J, Huo Y, Lloyd A (2009) TTG1 complex MYBs, MYB5 and TT2, control outer seed coat differentiation. Developmental Biology **325:** 412-421

Gu Q, Ferrandiz C, Yanofsky MF, Martienssen R (1998) The FRUITFULL MADS-box gene mediates cell differentiation during Arabidopsis fruit development. Development **125:** 1509-1517

Guo Q, Cui SW, Wang Q, Christopher Young J (2008) Fractionation and physicochemical characterization of psyllium gum. Carbohydrate Polymers **73:** 35-43

Hyde BB (1970) Mucilage-producing cells in the seed coat of *Plantago ovata*: developmental fine structure. American Journal of Botany**:** 1197-1206

Jensen JK, Johnson N, Wilkerson CG (2013) Discovery of diversity in xylan biosynthetic genes by transcriptional profiling of a heteroxylan containing mucilaginous tissue. Frontiers in plant science **4:** 183-183

Jensen JK, Johnson NR, Wilkerson CG (2014) *Arabidopsis thaliana* IRX10 and two related proteins from psyllium and *Physcomitrella patens* are xylan xylosyltransferases. The Plant Journal **80:** 207-215

Jensen JK, Kim H, Cocuron JC, Orler R, Ralph J, Wilkerson CG (2011) The DUF579 domain containing proteins IRX15 and IRX15-L affect xylan synthesis in *Arabidopsis*. The Plant Journal **66:** 387-400

Johnson CS, Kolevski B, Smyth DR (2002) TRANSPARENT TESTA GLABRA2 a trichome and seed coat development gene of Arabidopsis, encodes a WRKY transcription Factor. The Plant Cell **14:** 1359

Karimzadeh G, Omidbaigi R (2004) Growth and seed characteristics of isabgol (*Plantago ovata* Forsk) as influenced by some environmental factors. Journal of Agricultural Science and Technology **6:** 103-110

Kaswan V, Joshi A, Maloo S (2013) Assessment of genetic diversity in Isabgol (*Plantago ovata* Forsk.) using random amplified polymorphic DNA (RAPD) and inter-simple sequence repeat (ISSR) markers for developing crop improvement strategies. African Journal of Biotechnology **12**

Konishi S, Izawa T, Lin Shao Y, Ebana K, Fukuta Y, Sasaki T, Yano M (2006) An SNP caused loss of seed shattering during rice domestication. Science **312:** 1392-1396

Kotwal S, Kaul S, Sharma P, Gupta M, Shankar R, Jain M, Dhar MK (2016) De novo transcriptome analysis of medicinally important *Plantago ovata* using RNA-Seq. PloS one **11:** e0150273

Kour B, Kotwal S, Dhar MK, Kaul S (2016) Genetic diversity analysis in Plantago ovata and some of its wild allies using RAPD markers. Russian agricultural sciences **42:** 37-41

Kuai J, Sun Y, Liu T, Zhang P, Zhou M, Wu J, Zhou G (2016) Physiological mechanisms behind differences in pod shattering resistance in Rapeseed (Brassica napus L.) Varieties. PLOS ONE **11:** e0157341

Kumar J (2015) Good agricultural practices for Isabgol. ICAR – Directorate of Medicinal and Aromatic Plants Research, Gujarat, India

Lal RK, Chanotiya CS, Gupta P (2020) Induced mutation breeding for qualitative and quantitative traits and varietal development in medicinal and aromatic crops at CSIR-

CIMAP, Lucknow (India): past and recent accomplishment. International Journal of Radiation Biology **96:** 1513-1527

Lamba LC, Veena G (1981) Anatomy of circumscissile dehiscence in *Plantago ovata* Forsk. Current science (Bangalore) **50:** 541-543

Li C, Zhang B, Chen B, Ji L, Yu H (2018) Site-specific phosphorylation of *TRANSPARENT TESTA GLABRA1* mediates carbon partitioning in Arabidopsis seeds. Nat Commun **9:** 571

Li C, Zhang B, Chen B, Ji L, Yu H (2018) Site-specific phosphorylation of *TRANSPARENT TESTA GLABRA1* mediates carbon partitioning in *Arabidopsis* seeds. Nature Communications **9:** 571

Li C, Zhou A, Sang T (2006) Rice domestication by reducing shattering. Science **311:** 1936-1939

Li F, Komatsu A, Ohtake M, Eun H, Shimizu A, Kato H (2020) Direct identification of a mutation in OsSh1 causing non-shattering in a rice (*Oryza sativa* L.) mutant cultivar using whole-genome resequencing. Scientific Reports **10:** 14936

Liljegren SJ, Roeder AHK, Kempin SA, Gremski K, Østergaard L, Guimil S, Reyes DK, Yanofsky MF (2004) Control of fruit patterning in Arabidopsis by INDEHISCENT. Cell **116:** 843-853

Lin Z, Griffith Me Fau - Li X, Li X Fau - Zhu Z, Zhu Z Fau - Tan L, Tan L Fau - Fu Y, Fu Y Fau - Zhang W, Zhang W Fau - Wang X, Wang X Fau - Xie D, Xie D Fau - Sun C, Sun C (2007) Origin of seed shattering in rice (*Oryza sativa* L.).

Lv S, Wu W, Wang MA-O, Meyer RS, Ndjiondjop MN, Tan LA-O, Zhou H, Zhang JA-O, Fu Y, Cai HA-O, Sun C, Wing RA, Zhu ZA-O (2018) Genetic control of seed shattering during African rice domestication.

Mitsuda N, Ohme-Takagi M (2008) NAC transcription factors NST1 and NST3 regulate pod shattering in a partially redundant manner by promoting secondary wall formation after the establishment of tissue identity.

Mizzotti C, Mendes MA, Caporali E, Schnittger A, Kater MM, Battaglia R, Colombo L (2012) The MADS box genes SEEDSTICK and ARABIDOPSIS Bsister play a maternal role in fertilization and seed development. The Plant Journal **70:** 409-420

Nakano Y, Yamaguchi M, Endo H, Rejab NA, Ohtani M (2015) NAC-MYB-based transcriptional regulation of secondary cell wall biosynthesis in land plants. Frontiers in Plant Science **6:** 288

Naran R, Chen G, Carpita NC (2008) Novel rhamnogalacturonan I and arabinoxylan polysaccharides of flax seed mucilage. Plant Physiology **148:** 132-141

North HM, Berger A, Saez-Aguayo S, Ralet M-C (2014) Understanding polysaccharide production and properties using seed coat mutants: future perspectives for the exploitation of natural variants. Annals of Botany **114:** 1251-1263

Olsen KM (2012) One gene's shattering effects. Nature Genetics **44:** 616-617

Pabón-Mora N, Wong GK-S, Ambrose BA (2014) Evolution of fruit development genes in flowering plants. Frontiers in plant science **5:** 300-300

Pal MD, Raychaudhuri S (2003) Estimation of genetic variability in *Plantago ovata* cultivars. Biologia Plantarum **47:** 459-462

Patel D, Patel H, Patel P, Patel H, Amin A (2018) Evaluation of stable and non shattering isabgol cultivar-Gujarat isabgol 4. JOSAC**:** 88-90

Patel N, Sriram S, Dalal K (1980) Floral biology and stigma-pollen maturation schedule in isabgul *Plantago ovata* F. Current Science**:** 689-691

Patel S, Pachhigar K, Ganvit R, Panchal RR, Ponnuchamy M, Kumar J, Reddy NRR (2020) Exploring flowering genes in Isabgol (*Plantago ovata* Forsk.) Through Transcriptome Analysis. Plant Molecular Biology Reporter**:** 1-20

Phan JL, Cowley JM, Neumann KA, Herliana L, O'Donovan LA, Burton RA (2020) The novel features of *Plantago ovata* seed mucilage accumulation, storage and release. Sci Rep **10:** 1-14

Phan JL, Tucker MR, Khor SF, Shirley N, Lahnstein J, Beahan C, Bacic A, Burton RA (2016) Differences in glycosyltransferase family 61 accompany variation in seed coat mucilage composition in *Plantago* spp. Journal of Experimental Botany **67:** 6481-6495

Ponnuchamy M, Patel S, Mathew J, Kumar J, Reddy NRR (2020) Comparative transcriptome analysis uncovers genes and pathways relating to downy mildew resistance in Isabgol (*Plantago ovata* Forsk.).

Rajani S, Sundaresan V (2001) The Arabidopsis myc/bHLH gene ALCATRAZ enables cell separation in fruit dehiscence. Current Biology **11:** 1914-1922

Roeder AHK, Ferrándiz C, Yanofsky MF (2003) The role of the REPLUMLESS homeodomain protein in patterning the Arabidopsis fruit. Current biology **13:** 1630-1635

Rohilla AK, Kumar M, Sindhu A, Boora K (2012) Genetic diversity analysis of the medicinal herb Plantago ovata (Forsk.). African Journal of Biotechnology **11:** 15206-15213

Rønstadt N, Chase MW, Albach DC, Bello MA (2002) Phylogenetic relationships within Plantago (Plantaginaceae): evidence from nuclear ribosomal ITS and plastid trnL-F sequence data. Botanical Journal of the Linnean Society **139:** 323-338

Saez-Aguayo S, Ralet M-C, Berger A, Botran L, Ropartz D, Marion-Poll A, North HM (2013) PECTIN METHYLESTERASE INHIBITOR6 promotes *Arabidopsis* mucilage release by limiting methylesterification of homogalacturonan in seed coat epidermal cells. The Plant Cell **25:** 308

Sharma N, Koul P, Koul A (1992) Reproductive biology of *Plantago*: Shift from cross-to self-pollination. Annals of botany **69:** 7-11

Shen B, Sinkevicius KW, Selinger DA, Tarczynski MC (2006) The Homeobox Gene GLABRA2 Affects Seed Oil Content in Arabidopsis. Plant Molecular Biology **60:** 377-387

Shi L, Katavic V, Yu Y, Kunst L, Haughn G (2012) Arabidopsis glabra2 mutant seeds deficient in mucilage biosynthesis produce more oil. Plant J **69:** 37-46

Singh N, Lal RK, Shasany AK (2009) Phenotypic and RAPD diversity among 80 germplasm accessions of the medicinal plant isabgol (*Plantago ovata*, Plantaginaceae). Genet. Mol. Res **8:** 1273-1284

Taylor-Teeples M, Lin L, de Lucas M, Turco G, Toal TW, Gaudinier A, Young NF, Trabucco GM, Veling MT, Lamothe R, Handakumbura PP, Xiong G, Wang C, Corwin J, Tsoukalas A, Zhang L, Ware D, Pauly M, Kliebenstein DJ, Dehesh K,

Tagkopoulos I, Breton G, Pruneda-Paz JL, Ahnert SE, Kay SA, Hazen SP, Brady SM (2015) An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature **517:** 571-575

Tucker M, Ma C, Phan J, Neumann K, Shirley N, Hahn M, Cozzolino D, Burton R (2017) Dissecting the genetic basis for seed coat mucilage heteroxylan biosynthesis in *Plantago ovata* using gamma irradiation and infrared spectroscopy. Frontiers in Plant Science **8**

Vahabi A, Lotfi A, Solouki A, Bahrami S (2008) Molecular and morphological markers for the evaluation of diversity between Plantago ovata in Iran. Biotechnology **7:** 702-708

Vala AG, Fougat RS, Jadeja GC (2011) Genetic diversity of Plantago ovata Forsk. through RAPD markers. Electronic Journal of Plant Breeding **2:** 592-596

Verma A, Mogra R (2013) Psyllium (*Plantago ovata*) husk: A wonder food for good health.

Voiniciuc C, Günl M, Schmidt MH-W, Usadel B (2015) Highly branched xylan made by IRREGULAR XYLEM14 and MUCILAGE-RELATED21 links mucilage to *Arabidopsis* seeds. Plant physiology **169:** 2481-2495

Voiniciuc C, Yang B, Schmidt MH-W, Günl M, Usadel B (2015) Starting to gel: How *Arabidopsis* seed coat epidermal cells produce specialized secondary cell walls. International journal of molecular sciences **16:** 3452-3473

Volta U, De Giorgio R (2012) New understanding of gluten sensitivity. Nature Reviews Gastroenterology & Hepatology **9:** 295-299

Wang Z, Chen M, Chen T, Xuan L, Li Z, Du X, Zhou L, Zhang G, Jiang L (2014) TRANSPARENT TESTA2 regulates embryonic fatty acid biosynthesis by targeting FUSCA3 during the early developmental stage of Arabidopsis seeds. The Plant Journal **77:** 757-769

Western TL, Burn J, Tan WL, Skinner DJ, Martin-McCaffrey L, Moffatt BA, Haughn GW (2001) Isolation and characterization of mutants defective in seed coat mucilage secretory cell development in *Arabidopsis*. Plant Physiology **127:** 998-1011

Western TL, Skinner DJ, Haughn GW (2000) Differentiation of mucilage secretory cells of the *Arabidopsis* seed coat. Plant Physiology **122:** 345-356

Western TL, Young DS, Dean GH, Tan WL, Samuels AL, Haughn GW (2004) *MUCILAGE-MODIFIED4* encodes a putative pectin biosynthetic enzyme developmentally regulated by *APETALA2*, *TRANSPARENT TESTA GLABRA1& GLABRA2* in the *Arabidopsis* seed coat. Plant Physiology **134:** 296

Wright SI, Kalisz S, Slotte T (2013) Evolutionary consequences of self-fertilization in plants. Proceedings. Biological sciences **280:** 20130133-20130133

Yoon J, Cho L-H, Antt HW, Koh H-J, An G (2017) KNOX Protein OSH15 induces grain shattering by repressing lignin biosynthesis genes. Plant Physiology **174:** 312-325

Yoon J, Cho Lh Fau - Kim SL, Kim Sl Fau - Choi H, Choi H Fau - Koh H-J, Koh Hj Fau - An G, An G (2014) The BEL1-type homeobox gene SH5 induces seed shattering by enhancing abscission-zone development and inhibiting lignin biosynthesis.

Yu L, Yakubov GE, Zeng W, Xing X, Stenson J, Bulone V, Stokes JR (2017) Multi-layer mucilage of *Plantago ovata* seeds: Rheological differences arise from variations in arabinoxylan side chains. Carbohydrate Polymers **165:** 132-141

Zhao, Q. and Dixon, R.A (2011). Transcriptional networks for lignin biosynthesis: more complex than we thought?. Trends in Plant Science **16**(4): .227-233.

Zhou Y, Lu D, Li C, Luo J, Zhu B-F, Zhu J, Shangguan Y, Wang Z, Sang T, Zhou B, Han B (2012) Genetic control of seed shattering in rice by the APETALA2 transcription factor shattering abortion1. The Plant Cell **24:** 1034-1048

Chapter 3

A chromosome-level genome assembly of *Plantago ovata*

ꕥ

Statement of Authorship

Title of Paper	A chromosome-level genome assembly of *Plantago ovata*
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Lina Herliana[1], Julian G. Schwerdt[1], Tycho R. Neumann[1,4], Anita Severn-Ellis[5], Jana L. Phan[1], James M. Cowley[1], Neil J. Shirley[1], Matthew R. Tucker[1], Tina Bianco-Miotto[1], Jacqueline Batley[5], Nathan S. Watson-Haigh[2,3]* and Rachel A. Burton[1]*

Principal Author

Name of Principal Author (Candidate)	Lina Herliana
Contribution to the Paper	Conducted data mining, wrote and executed code/script/pipelines, assembled RNAseq data, assembled and annotated the genome, analysed and interpreted data, wrote and edited manuscripts, deposited raw and processed datasets into databases.
Overall percentage (%)	75%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 8/6/22

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Julian G. Schwerdt
Contribution to the Paper	Conceived the project and provided genomic data.
Signature	Date 8/6/22

Name of Co-Author	Tycho R. Neumann
Contribution to the Paper	Performed DNA extraction and library preparation for PacBio CLR sequencing and commented on the manuscript.
Signature	Date 8/6/22

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Anita Severn-Ellis		
Contribution to the Paper	Performed DNA extraction and library preparation for the Hi-C experiment and contributed to writing the method.		
Signature		Date	8/6/22

Name of Co-Author	Jana L. Phan		
Contribution to the Paper	Provided RNA-seq data and commented on the manuscript.		
Signature		Date	8/6/22

Name of Co-Author	James M. Cowley		
Contribution to the Paper	Provided RNA-seq data and commented on the manuscript.		
Signature		Date	8/6/22

Name of Co-Author	Neil J. Shirley		
Contribution to the Paper	Provided RNA-seq data and commented on the manuscript.		
Signature		Date	8/6/22

Name of Co-Author	Matthew R. Tucker		
Contribution to the Paper	Provided RNA-seq data and commented on the manuscript.		
Signature		Date	8/6/22

Name of Co-Author	Tina Bianco-Miotto		
Contribution to the Paper	Assisted in data analysis and contributed to preparing the manuscript.		
Signature		Date	8/6/22

Name of Co-Author	Jacqueline Batley		
Contribution to the Paper	Provided genomic Hi-C data and commented on the manuscript.		
Signature		Date	8/6/22

Name of Co-Author	Nathan S. Watson-Haigh		
Contribution to the Paper	Provided bioinformatics training, assisted in workflow development, contributed to preparing the manuscript and corresponding author.		
Signature		Date	8/6/22

Name of Co-Author	Rachel A. Burton		
Contribution to the Paper	Conceived the project, designed experiments, revised the manuscript, and corresponding author.		
Signature		Date	8/6/22

Title A chromosome-level genome of *Plantago ovata*

Authors Lina Herliana[1], Julian G. Schwerdt[1], Tycho R. Neumann[1,4], Anita Severn-Ellis[5], Jana L. Phan[1], James M. Cowley[1], Neil J. Shirley[1], Matthew R. Tucker[1], Tina Bianco-Miotto[1], Jacqueline Batley[5], Nathan S. Watson-Haigh[2,3]* and Rachel A. Burton[1]*

[1]School of Agriculture, Food and Wine, Waite Research Institute, University of Adelaide, Waite Campus, Urrbrae, SA, Australia

[2]South Australian Genomics Centre (SAGC), SA, Australia

[3]Australian Genome Research Facility, Victorian Comprehensive Cancer Centre, Melbourne, VIC 3000, Australia

[4]IP Australia, PO Box 200, Woden, ACT, 2606, Australia

[5]School of Biological Sciences, University of Western Australia, Crawley, WA 6009, Australia

Keywords *Plantago ovata*; psyllium; mucilage; reference genome; PacBio; Hi-C; chromosome; annotation

Abbreviations

AGRF	Australian Genome Research Facility Ltd
BLAST	Basic Local Alignment Search Tool
bp	base pairs
BUSCO	Benchmarking Universal Single-Copy Orthologs
BWA	Burrows-Wheeler Aligner
CDS	coding sequence
Chr	Chromosome
CLR	Continuous long read
CRL	custom repeat library
DPA	days post anthesis
FISH	fluorescence in situ hybridization
Hi-C	high-throughput chromosome conformation capture
kb	kilobase pairs
Gb	gigabase pairs
LINE	long interspersed nuclear element
lncRNA	long non-coding RNA
LAI	LTR Assembly Index
LTR	long terminal repeat
Mb	megabase pairs
N50	A metric to assess the contiguity of an assembly that is defined by the length of the shortest contig at 50 % of the assembly
NCBI	National Center for Biotechnology Information
ncRNA	Non-coding RNA
NIB	Nuclei Isolation Buffer
NUMT	Nuclear mitochondrial

NUPT	Nuclear plastid
PacBio	Pacific Biosciences
RNA-seq	RNA sequencing
rRNA	Ribosomal RNA
SMRT	Single-molecule real-time
SRA	Sequence Read Archive
TE	Transposable element
TRF	Tandem Repeats Finder
tRNA	Transfer RNA
UTR	Untranslated region

Abstract

Plantago ovata is cultivated for production of its seed husk (psyllium). When wet, the husk transforms into a mucilage with properties suitable for pharmaceutical industries, utilised in supplements for controlling blood cholesterol levels, and food industries for making gluten-free products. There has been limited success in improving husk quantity and quality through breeding approaches, partly due to the lack of a reference genome. Here we constructed the first chromosome-scale reference assembly of *P. ovata* using a combination of 5.98 million PacBio and 636.5 million Hi-C reads. We also used publicly available short-read Illumina genomic data to estimate genome size and transcripts to generate gene models. The final assembly covers ~500 Mb with 95.10% gene set completeness. A total of 97% of the sequences are anchored to four chromosomes with an N50 of ~128.87 Mb. The *P. ovata* genome contains 61.90% repeats, where 40.04% are long terminal repeats (LTRs). We identified 23,346 protein-coding genes, 411 non-coding RNAs (ncRNAs), 108 ribosomal RNAs (rRNAs), and 1,295 transfer RNAs (tRNAs). This genome will provide a resource for plant breeding programs to, for example, reduce agronomic constraints such as seed shattering, increase psyllium yield and quality, and overcome crop disease susceptibility.

Introduction

Plantago ovata (Fig. 1) seed husk, commonly called psyllium or Isabgol, has a long history of use in human health as dietary fibre when ingested[1,2] and food industries as a primary stabiliser in products such as ice cream, and as a gluten substitute in baking[3]. As a commercially valuable plant, many attempts have been made to develop higher-yielding varieties with larger seed size, higher husk content, non-shattering capsules, synchronous maturity, and resistance to abiotic (e.g., drought and frost) and biotic stresses (e.g., downy mildew)[4,5]. As the primary producer and exporter, India initiated a *P. ovata* breeding program as early as 1976 in the Pilwai tract of North Gujarat, while trials to establish best agronomic practices were undertaken in Australia in 1985 in the Ord River Irrigation Area (ORIA), Kununurra region, Western Australia[6]. However, many studies reported that conventional breeding approaches had not significantly improved seed or psyllium production[7-9]. Genetic improvement of this plant is challenging because *P. ovata* has a narrow genetic base, a small number of chromosomes (2n=8) enriched in heterochromatin, low chiasmata frequency, low recombination index and a high selfing rate[8-13]. As a result, this plant is sensitive to environmental changes that may threaten the supply chain and increase the global price of psyllium.

Exposure to gamma irradiation has been reported to successfully induce phenotypic variation in *P. ovata*[13-15]. *P. ovata* var 'Mayuri' is one example of a gamma-irradiated mutant with valuable traits, including early maturation with pigment markers guiding the right timing for harvesting, combined with high seed and husk production[14]. However, before this cultivar was patented in 2003, the evaluation period was very long, requiring three generations for selfing (M1-M3), three generations for vegetative propagation (M4-M6) and two years for pilot-scale trials[14]. This period could be significantly reduced if the candidate genes related to the favourable traits were known. Candidate genes can be targeted using CRISPR/Cas9 technology to improve the quantity and quality of psyllium. One way to identify candidate genes is to use RNA sequencing to generate transcriptomic data. Since 2010, at least six studies have deposited

P. ovata RNA-seq raw data in the Sequence Read Archive (SRA) at the National Center for Biotechnology (NCBI) (Supplementary file 1: Table S1). All the studies used *de novo* transcript assembly because no genome reference was available. Only the *P. ovata* chloroplast genome has been assembled to date[16]. This helps resolve taxonomic relationships among species but has limited application for genetic improvement. The challenge of using transcriptome assemblies is distinguishing between sequence artefacts and the genes themselves due to alternative splicing producing splice variants. In addition, there is a need to create a transcriptome assembly for every different project as transcripts are tissue and time specific. To provide a universal resource, a reference genome is required.

Here we report the process of generating and utilising a *P. ovata* chromosome level assembly. Continuous long read (CLR) data from Pacific Biosciences (PacBio) was used to create a contig assembly, while a Hi-C approach capturing chromosome conformation was used to guide the scaffolding. We gathered all publically available RNA-seq data and combined it with data generated at the University of Adelaide to predict the gene models. Since short reads have fewer sequence errors than long reads, we used publicly available Illumina whole-genome sequencing data, from SRA accession number SRR10076762, to estimate the genome size. The construction of a *P. ovata* reference genome will help genetic improvement programs for *P. ovata* as well as supporting laboratory-based experiments to better understand the seed biology of this species.

Figure 1: *Plantago ovata*. **a**. A two-and-a-half-month-old plant. **b**. Capsules containing two seeds each are fully ripened and shatter easily at around 25 days post anthesis (DPA).

Results

Genome assembly

The data pool was comprised of a total of 5.98 M (50 bp-121.17 Kb) PacBio long reads and 636.5 million (47.74 Gb) Hi-C short-reads. PacBio reads were used to assemble contigs, while Hi-C reads were used to achieve chromosome level assembly. About 14.15% (0.85 M) unwanted reads (mitochondrial and chloroplast sequences) were removed from PacBio raw reads, leaving 5.14 M (33.97 Gb) clean reads suitable for assembly. Canu generated 5,591 contigs (577.27 Mb) including 711 repeats (17.51 Mb), 1,401 bubbles (33.32 Mb) and 828,398 unassembled sequences (4.15 Gb) (Table 1). In total, the initial contig number was 6,992 (610.60 Mb, N50=184.59 Kb). Contigs were polished to improve accuracy and purged to remove alternative contigs (haplotigs) and artefacts. After polishing using clean reads, the contig number did not change, but the genome length slightly increased (611.02 Mb, N50=184.10 Kb). The bubbles and circular contigs were also removed from the polished contigs, leaving 5,591 sequences (577.67 Mb, N50=203.30 Kb). After purging and clipping, this dropped to 3,919 contigs (500.60 Mb, N50=260.25 Kb) (Table 1).

Table 1: Assembly statistics at different stages.

Stage of assembly	Seqs	Cumulative size (Mb)	N50	Min	Max (Mb)	LAI	GC content
Draft contig	6,992	610.60	184.58 Kb	1.83 Kb	9.04	-	-
Polished contig	6,992	611.02	184.10 Kb	1.84 Kb	9.04		
Purged contig	4,235	536,57	235.20 Kb	2.53 Kb	9.04	-	-
Clipped contig	3,919	500.60	260.25 Kb	2 bp	9.04	-	-
Draft scaffold (SALSA)	2,047	501.54	1.77 Mb	2 bp	53.60	-	-
Draft scaffold (3D-DNA)	900	502,30	129.19 Mb	2 bp	221.38	-	-
Curated scaffold (JBAT)	894	502.30	129.19 Mb	2 bp	138.10	-	-
NCBI criteria	876	500.94	128.87 Mb	328 bp	137.73	10.27	38.4%
Final assembly	4	487.38	128.87 Mb	106.35 Mb	137.73	-	-
	872	13.55	20.50 Kb	328 bp	0.23		

Seqs: Number of sequences; LAI: LTR Assembly Index

Two scaffolders were tested to generate a chromosome level assembly from these contigs using 298.34 million Hi-C clean reads. Using the SALSA2 pipeline[17], resolution to chromosome level could not be achieved (2,047 contigs, N50=1.77 Mb). In contrast, using the 3D-DNA pipeline, 900 sequences were obtained (502.30 Mb) with a significant increase in the N50 value (129.19 Mb). After visualising the assembly using Juicebox, this scaffold assembly shows four chromosomes. However, there were some misjoined scaffolds and so the misjoins were curated manually following the JBAT method[18]. The curated assembly (894 sequences, 502.30 Mb, N50=129.19 Mb) had fewer sequences but a similar N50 value compared to the uncurated assembly (Table 1). After adjusting NCBI criteria, the final assembly has 876 sequences (500.94 Mb, N50=128.87 Mb). The four superscaffolds comprise chromosome length accounting for 97.29% (487.38 Mb) of the total genome length and the unplaced scaffolds account only for 2.71% (13.55 Mb). Based on the lengths we labelled HiC_scaffold_1 (137.73 Mb) as chromosome 1, HiC_scaffold_2 (128.87 Mb) as chromosome 2, HiC_scaffold_3 (114.44 Mb) as chromosome 3, and HiC_scaffold_4 (106.35 Mb) as chromosome 4 (Supplementary file 2).

Genome size and assembly quality

To predict the *P. ovata* genome size, publically-available raw genomic short-read Illumina reads (HiSeq X Ten) were used. The data (SRR10076762) is 195.7 M reads (58.4 Gb). The result from *k*-mer analysis shows that the estimated haploid genome size is 398.59 Mb (21-mer) up to 583.57 Mb (100-mer) (Table 2). Our assembled genome with unknown gaps is 500.94 Mb. By excluding "N" (unknown gaps), the size is adjusted to 500.60 Mb.

The assembled genome quality was assessed in five ways. First, at 50% of the total genome length (N50), the shortest scaffold length is 128.87 Mb (Table 1). Second, the percentage of BUSCO completeness is 95.1% (Fig. 2). This percentage consists of 90.2% single-copy and 4.9% duplicated genes. Only 1% of sequences are fragmented, and 3.9% remain missing (Fig. 2). Third, the percentage of publically available genomic short-read Illumina data

(SRR10076762) mapped to our genome assembly is 95.81%, while the portion of our genomic long-read PacBio data (SRR14643405) mapped back to the assembly is 92.25%. Fourth, the mapping rate of RNA-seq data is up to 96.10% (Supplementary file 1: Table S2). Lastly, the LTR Assembly Index (LAI) score is 10.27 (Supplementary file 3).

Table 2: Estimation of genome size, repeat content, and heterozygosity using different *k*-mers.

Parameters	**21**	**31**	**39**	**45**	**50**	**70**	**100**
Heterozygosity	0.19%	0.17%	0.16%	0.15%	0.15%	0.14%	0.13%
Haploid (Mb)	398.59	439.68	465.63	482.27	494.37	531.69	583.57
Repeat (Mb)	182.27	186.26	187.92	187.86	187.12	179.83	182.36
Unique (Mb)	216.32 (54.27%)	253.47 (57.6%)	277.71 (59.64%)	294.41 (61.05%)	307.26 (62.15%)	351.86 (66.18%)	401.20 (68.74%)
Model Fit	95.96%	96.41%	96.89%	97.22%	97.54%	98.58%	99.52%
Read Error Rate	0.34%	0.28%	0.24%	0.23%	0.15%	0.18%	0.14%

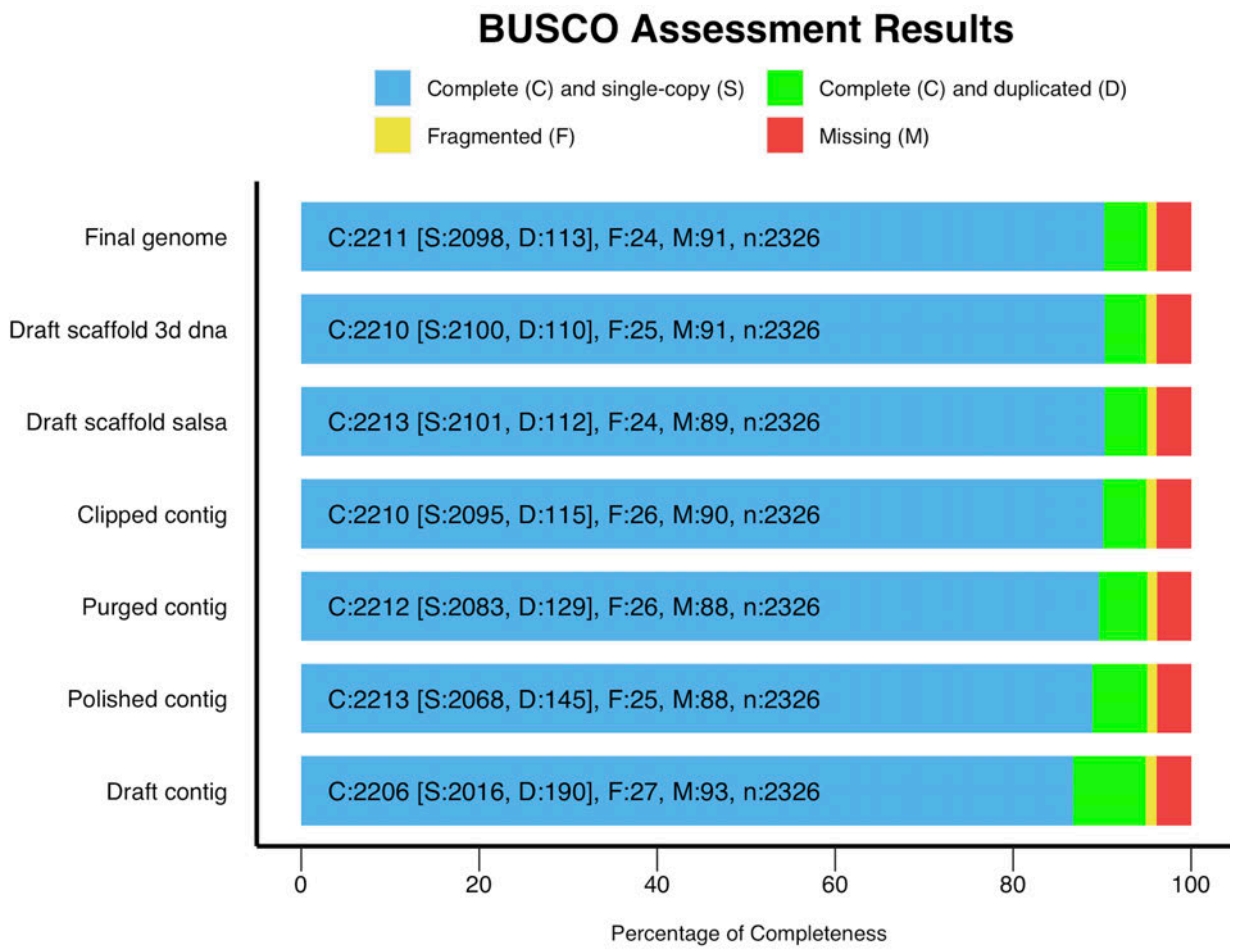


Figure 2: Assessment of BUSCO completeness at each assembly stage.

Repeat content estimation and identification

Repeat content was estimated using short and long reads. The repeat content prediction using Illumina genomic data shows that the genome contains 31.26–45.73% repetitive regions (Supplementary file 1: Table S3). Using long reads, the *P. ovata* genome appears to contain 61.90% (310.10 Mb) repeats with long terminal repeat (LTR) retrotransposons comprising the highest proportion (200.59 Mb, 40.04%) (Supplementary file 1: Table S3). Two out of three major groups of LTR retrotransposons were detected in the assembled genome. They are Ty1/Copia (98.64 Mb, 19.69%) and Gypsy/DIRS1 (101.64 Mb, 20.29%). There are 366 sequences defined as satellites (98,9 Kb, 0.02%). Less than 1% (3.49 Mb) of the repeat content is simple repeats. Simple repeats TTTAGGG identified as a typical plant telomere sequence[11],

were located at the end of all chromosomes, while AAACCCT, the canonical or reverse complement of the telomere repeat, were found at the beginning of the chromosomes. Other telomeric variants were also found in this assembly, such as TTTGGGG, TTTCGGG, TTCAGGG, TTTTAGGG and AACCCGG (Supplementary file 4).

Coding and non-coding genes

This *P. ovata* genome is estimated to contain 25,045 genes, determined using RNA-seq data (Supplementary file 1: Table S2). Approximately 93% of these encode proteins (23,346) and they cover 19.09% (~95.63 Mb) of the genome length (Table 3, Supplementary file 1: Table S4). The number of transcripts (mRNA) is equal to the number of coding sequences (CDS) at 63,228. On average, one gene can have 2-3 transcripts due to alternative splicing. The average size of each protein-encoding gene is 4,096 bp and they are distributed across all chromosomes. There are 7,203 genes on chromosome 1, 5,779 on chromosome 2, 5,412 on chromosome 3 and 4,581 on chromosome 4. About 371 genes cannot be placed onto any of the four chromosomes since they align to unplaced scaffolds. Proportions between total cumulative gene length and chromosome length are 21.38%, 18.64%, 19.47%, 17.15%, respectively, for chromosome 1 to chromosome 4 (Table 3).

Table 3: Summary of coding and non-coding genes on each chromosome.

Chr	Total chromosome length	Number of coding genes	% GC of CDSs	Total gene length	Percentage	GT61 genes	tRNA	rRNA	lncRNA
Chr 1	137,725 Kb	7,203	44.4	29,445 Kb	21.38%	5	415	64	97
Chr 2	128,867 Kb	5,779	44.7	24,019 Kb	18.64%	2	286	6	76
Chr 3	114,445 Kb	5,412	44.4	22,282 Kb	19.47%	2	301	2	86
Chr 4	106,346 Kb	4,581	44.4	18,237 Kb	17.15%	10	255	8	56
Unplaced	13,555 Kb	371	43.9	1,645 Kb	12.14%	-	38	28	13
Total	500,939 Kb	23,346	44.4	95,629 Kb	19.09%	19	1,295	108	328

Seventeen *PoGT61* genes were mapped to different locations in the *P. ovata* genome (Table 3, Fig. 3, Supplementary file1: Table S5). Only *PoGT61_4* and *PoGT61_4L* have an overlapping location, so they are now counted as one gene (*PoGT61_4*) (Supplementary file 1: Table S5). Two novel *PoGT61* genes were found, totalling 19 *GT61* genes in this assembly (Table 3, Fig. 4, Supplementary file 1: Table S5). Five genes are on chromosome 1, two genes each on chromosome 2 and chromosome 3, and ten genes in a cluster on chromosome 4 (Table 3, Fig. 3). The number of transcripts from each gene varies from one (*PoGT61_12*) to nine (*PoGT61_13*) transcripts per gene (Supplementary file 1: Table S5).

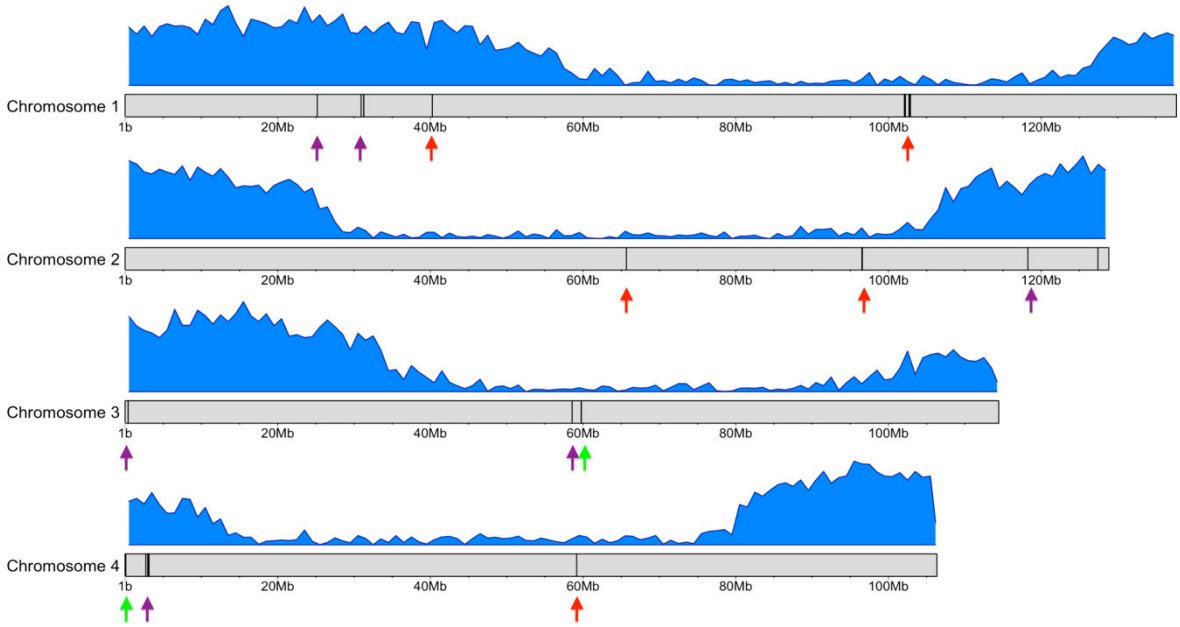


Figure 3: Gene density (blue) and distribution of 5S (red arrows), 45S (green arrows) rRNA, and *PoGT61* genes (purple arrows) in the *P. ovata* genome. The figure was generated in R using the karyoploteR library[51]. The x-axis represents genome position (Mb) and the y-axis represents gene density using a sliding window of one megabase in length.

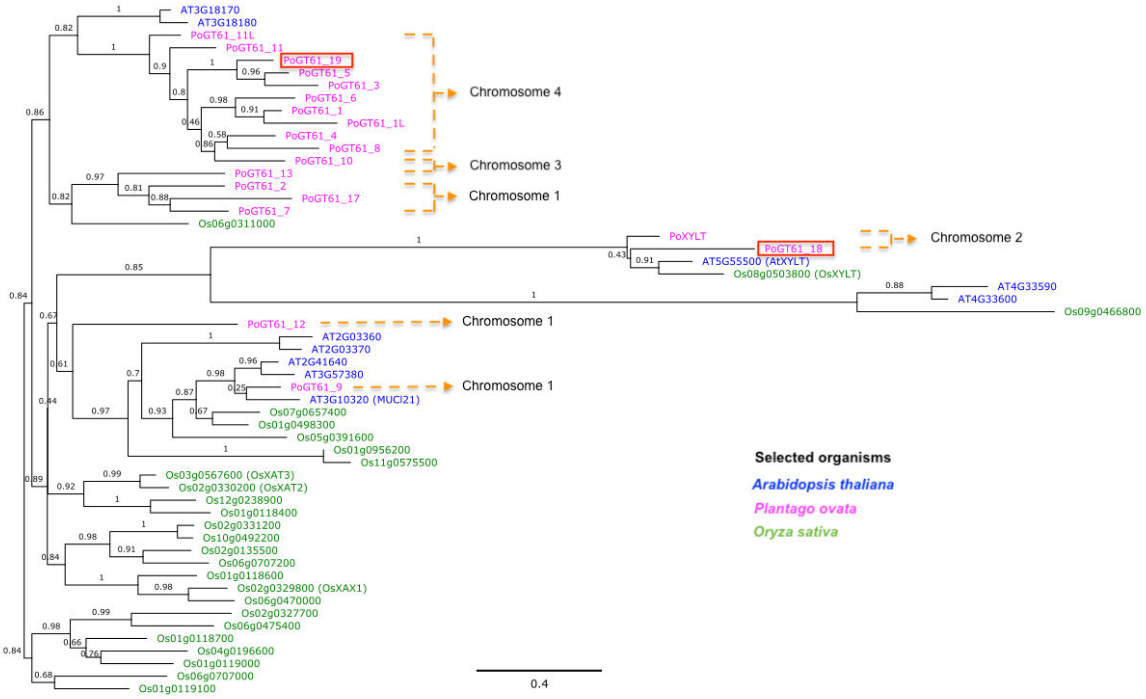


Figure 4: A phylogenetic tree of GT61 protein sequences on selected species was visualised using FigTree v1.4.4. Red boxes indicate *P. ovata* GT61 that have not been identified previously.

We identified 108 ribosomal RNAs (rRNAs), 1,295 transfer RNAs (tRNAs), and 411 non-coding RNAs (ncRNAs). The identified non-coding RNAs (ncRNAs) comprise 328 long non-coding RNAs (lncRNAs), 17 primary transcripts of microRNAs (miRNAs), 48 small nuclear RNAs (snRNAs), 12 small nucleolar RNAs (snoRNAs), 2 ribonuclease mitochondrial RNA processing (RNase MRP) RNAs, and 4 signal recognition particle (SRP) RNAs. Several types of cytoplasmic rRNA are annotated in the genome belonging to 5S, 18S, and 25S classes. The 5S sequences are clustered on chromosome 1 (63 sequences) with only six 5S sequences on chromosome 2, one on chromosome 4 and none on chromosome 3. Ribosomal 45S RNAs are found only on chromosomes 3 and 4 (Fig. 3).

In total, there are 328 lncRNAs in the *P. ovata* genome. They are distributed across four chromosomes with 97 transcripts on chromosome 1, 76 transcripts on chromosome 2, 86 transcripts on chromosome 3, 56 transcripts on chromosome 4, and 13 transcripts on unplaced sequences. Based on the locations of lncRNAs and the nearest mRNAs, we found 288 lncRNA/mRNA pairs in the assembly (Supplementary file 1: Table S6). They can be grouped into six categories as shown in Supplementary file 5 with sense and antisense intergenic groups dominating this assembly.

Discussion

The *P. ovata* genome assembly presented here is high quality as defined using a range of parameters. Despite having 876 scaffolds, the scaffold N50 value of 128.87 Mb (Table 1) is excellent and can be used to show that the shortest scaffold at half of the total genome length is the length of chromosome 2 (Table 3). Four chromosomes account for 97.29% of the total haploid genome size, indicating that the assembly is highly contiguous. The scaffold N50 value is far higher than the average length of a *P. ovata* gene at 4,096 bp (Supplementary file 1: Table S4) indicating a much higher chance of generating complete gene models. This is supported by a BUSCO completeness value of 95.20%, where only 1% of the genes present in a selected eudicot cohort are fragmented in this assembly (24 out of 2,326 genes). In addition, the high mapping rate of reads from RNA-seq data to this assembly (up to 96.10%) will facilitate accurate data interpretation by preventing false positives in downstream analyses such as for transcriptomics[19].

We used the LAI score to evaluate the continuity of our assembly where the program requires at least 0.1% intact LTR-RTs and 5% LTR-RTs as a proportion of the total genome size[20]. Ou et al.[20] evaluated 103 genomes with contents of intact LTR-RTs ranging from 0.28% to 18.34% and total amounts of LTR-RTs from 5.49 to 69.38%. Our assembly meets these criteria with 8.38% intact and 52% total LTR-RTs. This assembly has an LAI score of 10.27 (Table 1, Supplementary file 3). Based on the classification of assembled repeat sequences using the LAI score[20] our assembly can be classified as a reference ($10 \leq LAI \leq 20$). Advances in technology to produce longer reads with higher accuracy could further improve the current assembly to gold or even platinum standard.

Of note is that this assembly has a lower LAI score (10.27) than the raw LAI (15.90) (Table 1, Supplementary file 3). About 25% (26/103) of genomes studied[20] show the same trend. All these genomes, including our assembly (96.34%), have a whole genome LTR identity higher

than 94%. It has been suggested that those species with recent LTR-RT amplifications provide more intact raw LTR elements that are thus represented by a higher LTR identity.

Three regions in this assembly were detected as originating from mitochondrial sequences based on contamination screening during genome submission to NCBI database. Two regions are in HiC_scaffold_1 (chromosome 1), with one in HiC_scaffold_2 (chromosome 2). The lengths are 250 bp, 149 bp, and 177 bp. However, PacBio long reads span these three regions with no breaks suggesting that they are genuinely part of the nuclear genome (Supplementary file 6). Michalovova et al.[21] reported insertions of nuclear mitochondrial DNA (NUMT) and nuclear plastid DNA (NUPT) in six plant species. They reported the insertions were localised near centromeres in rice and Arabidopsis. During manual curation of the *P. ovata* annotation file, genes from the chloroplast and mitochondria were found in the nuclear assembly, suggesting these three regions are most likely to be NUMT. Further research is needed to investigate gene transfer from organelles to the nuclear genome to characterise NUMT and NUPT in *P. ovata*.

The haploid genome size estimated from Illumina short reads using the *k*-mer frequency spectrum is between 398.59 Mb and 583.57 Mb (Table 2). The *P. ovata* genome size has been previously estimated using a flow cytometer and reported in three different studies. Badr et al.[22] reported diploid *P. ovata* from Cairo has a genome of between 484.11 Mb (C value: 0.495 pg) and 523.23 Mb (C value: 0.535 pg). Pramanik and Raychaudhuri[10] studied this species from India (Anand) and reported 537.9 Mb (C value: 0.55 pg), whilst Dhar et al.[12] showed that the *P. ovata* genome size is about 621 Mb (C value: 0.635 pg). Potentially the range in sizes could be due to using different methods[12] or could also represent intraspecific variation. Schmuths et al.[23] found significant differences of about a 1.1-fold range between the genome size of 21 Arabidopsis accessions. Our assembly size (500.94 Mb) sits within the range of estimated haploid genome size using the *k*-mer method and the published genome size based on flow cytometry (FCM) analysis.

The guanine (G) and cytosine (C) content of DNA has been reported to play an important role in gene regulation and can be associated with how organisms adapt to their environment[24,25]. Šmarda et al.[24] observed that plants with GC-rich DNA were more adaptative in extreme climates. Our results show that the GC content of the genome is about 38.4% (Table 1, Supplementary file 7). Dhar et al.[9] stated that the *P. ovata* genome had 55% GC content adjusting this four years later to an AT content of 59.7%[12] and dropping the GC content to 40.3%. The later study[12] was conducted using flow cytometry (FCM). Šmarda et al.[26] compared the GC content of 11 rice species using FCM versus sequence data. They found that GC contents from sequence data are consistently lower than those from the flow cytometer. The different methods could explain why the GC content reported by Dhar et al.[12] is slightly higher than our calculation of 38.4%, based on genomic sequences. However, Dhar et al.[12] and this study agree that the *P. ovata* genome is AT-rich. As AT base pairs have lower thermal stability than the GC base pairs[24], having low GC content could signify that the plant is potentially less adaptive to extreme climates.

Wang et al.[27] found that plant domestication contributed to higher A and T content in maize and soybean compared to their wild relatives. Commercial *P. ovata* accessions could display the same increase in AT content due to domestication but breeding efforts have not been as intense in this species as for other crops. To test this hypothesis, we could measure and compare the GC content of Australian native *Plantago* species described in Cowley et al.[28] to the commercial accessions.

The GC content of the CDS, at 44.4%, (Table 3, Supplementary file 1: Table S7) is higher by 6% compared to the genomic regions (Table 1, Supplementary file 7). Kotwal et al.[29] also found that the GC content from the *P. ovata* transcripts in their study was higher than the genomic GC content. However, as they only extracted and sequenced one tissue type (ovaries) this may not be a valid comparison. Kotwal et al.[29] also compared the GC content of *P. ovata* transcripts with *A. thaliana*, rice, tomato, and *Eucalyptus*. They classified *P. ovata* and *Eucalyptus*

(eudicot) in the same group as rice (monocot) with GC contents of 45-50% while *A. thaliana* and tomato (eudicot) have a lower GC content ranging from 40 to 45%[29]. However, *P. ovata* has a unimodal distribution (one peak) (Fig. 3 in Kotwal et al.[29], Supplementary file 1: Table S7). In contrast, rice has a bimodal distribution (two peaks) (Fig. 3 in Kotwal et al.[29]) so they should not be classified in the same group. Singh et al.[25] studied the GC content from 20 plant genomes and ranked the highest GC content from grass genomes (including rice), followed by a non-grass monocot and then finally from eudicots. Their results also showed that the eudicot genome has a unimodal distribution while grass monocots have a bimodal distribution[25]. Bimodal distribution is shaped by highly heterogenous GC content among genes in the grass genomes, giving one peak with GC-rich genes and another with GC-poor genes[25]. In contrast, eudicots show low variability or homogenous GC content among genes resulting in only one peak[25]. High GC content was found to be positively correlated with high recombination sites[30], which may be important for breeding strategies.

According to the centromere positions, *P. ovata* chromosome 1 is classified as metacentric, chromosome 2 as submetacentric while chromosomes 3 and 4 are subtelocentric[11]. However, for this assembly, the position of the centromeres is not accurately fixed but may be indicated using euchromatin and heterochromatin patterns. Euchromatin is active chromatin in the genome where more genes are transcribed, while heterochromatin is a less active and highly condensed region in the chromosome (Fig. 3). Dhar et al.[12] reported that euchromatic areas are located at the distal ends of all chromosomes and cover entirely one arm of chromosome 1. Our results agree with this but also provide additional information (Fig. 3). From Fig. 3, we can define heterochromatic regions from 60 to 120 Mb on chromosome 1, from 30-110 Mb on chromosome 2 and 40-110 Mb on chromosome 3 and 15-80 Mb on chromosome 4. The statistics on repeat content (61.90%, Supplementary file 1: Table S3) and proportion of total gene lengths (19.09%) that account for less than a quarter of chromosome lengths (Table 3) support the earlier finding using C binding and fluorescence in situ hybridization (FISH)

methods that most of the regions in the *P. ovata* genome are heterochromatin containing highly repetitive DNA[12].

We are confident in labelling HiC_scaffold_1 as chromosome 1 because of the presence of 5S rDNA cluster[11,31] and HiC_scaffold_2 as chromosome 2 as it does not have 45S rDNA. Only chromosomes 3 and 4 have 45S rDNA (Fig. 3)[11,31]. However, the location of 45S rDNA on chromosome 3 in our assembly, near the middle of the chromosome, is not the same as those proposed based on ribosomal physical mapping[11]. They found 45S rDNA signals at the ends of the short arms (beginning of chromosomes) of chromosomes 3 and 4. This could represent intraspecific variation or missed joins in the assembly. Better quality of long raw reads could address the problem of misjoined contigs. In addition, optical mapping technology could be used to validate the orientation of *de novo* assembly in the future[32].

We used this current genome assembly to identify candidate genes of the glycosyltransferase family 61 (GT61) family, which appear to encode key enzymes involved in mucilage biosynthesis[33,34] with a significant impact on final husk quantity and quality. Two out of eighteen genes (*PoGT61_4* and *PoGT61_4L*) mapped to the exact location of GeneID KN361_4g000369 on chromosome 4 (Supplementary file 1: Table S5) indicating that they are in fact the same gene, but with differentially spliced transcripts. Thus, analysis of the *P. ovata* contigs derived from the *de novo* transcriptome assembly[33] was not sufficient to fully resolve the single gene origin of alternative splice variants, which has only been clarified by use of this reference genome. In addition, two novel *PoGT61* genes (*PoGT61_18* and *PoGT61_19*), not represented in the transcriptomics resources, were found in this assembly. *PoGT61_18* is located on chromosome 2, near *PoXYLT* (Supplementary file 1: Table S5). Phylogenetic analysis of all GT61 protein sequences shows *PoGT61_18* to be closely related to *PoXYLT, AtXYLT*, and *OsXYLT* (Fig. 4), suggesting that the protein product of *PoGT61_18* could be a β-1,2-xylosyltransferase[35]. *PoGT61_19* and the other nine genes located on chromosome 4 (Fig. 4, Supplementary file 1: Table S5) are also clustered in the phylogenetic tree, but they lack

similarities to *GT61* genes from the model organisms Arabidopsis and rice, suggesting that they may be responsible for intrinsic differences in the *Plantago* xylan polysaccharides (Fig. 4). According to our annotation (JAHHQI010000000, Supplementary file 1: Table S5) these ten genes on chromosome 4 are predicted to be xylan arabinosyltransferases (α-1,3-arabinosyltransferase). Heterologous expression of these genes in other species could confirm their function. For example, the heterologous expression of rice and wheat *GT61* genes in Arabidopsis increased arabinose substitution and provided gain-of-function evidence for arabinosyltransferase activity[36]. The significantly higher number of *Plantago GT61* gene duplications has previously been suggested to be linked to the high density/complexity of backbone substitutions on the heteroxylan of *P. ovata* mucilage[33]. It is possible that different *GT61* enzymes add specific types of heteroxylan backbone decorations and the expression of multiple *Plantago GT61* genes in tandem may reveal such roles.

Overall, all parameters assessed indicate that we have generated a high quality assembled and annotated genome. The genome can be used as a reference, but we also provide Supplementary files that can benefit future research. Supplementary file 1: Table S6 contains information about lncRNA and mRNA candidates for future functional analysis to study how gene expression may be controlled by epigenetic mechanisms. Supplementary file 8 lists annotation for LTR Copia and Gypsy retrotransposons that may be helpful to study *Plantago* domestication. Identified location and sequences of genes linked to histone modifications and DNA methylation can be found in Supplementary file 1: Table S8, providing an additional epigenetic resource. The telomere sequences in Supplementary file 4 can be used for evolutionary analysis as suggested in the review by Peska and Garcia[37].

Conclusions

This study generated the first *P. ovata* genome assembly together with gene annotations. We achieved a chromosome-level assembly using *de novo* assembly of PacBio CLR data and contig scaffolding utilising Hi-C data. Our assembly is about 500 Mb in size and comprises four chromosomes. This resource will help accelerate *Plantago* breeding programs. Markers can be developed and candidate genes identified related to key phenotypes using Genome-Wide Selection (GWS) or RNA-seq strategies by comparing two distinct genotypes occurring in nature or generated by mutation. Specific regions in the genome can be targeted to improve the quantity and quality of psyllium using the latest technology, such as CRISPR/Cas9 or to select favourable traits in breeding programs.

Materials and methods

DNA extraction, library preparation and sequencing

P. ovata (Fig. 1) seeds were obtained from Accolent Dried Herbs, Queensland, Australia [33]. Plants were grown in the glasshouse as per Phan et al.[33]. Leaf tissues from mature plants were used for genomic DNA extraction for PacBio and Hi-C library construction. The study complies with local and national guidelines.

For PacBio sequencing, DNA extraction was achieved by combining protocols from Sikorskaite et al.[38] and QIAGEN® Genomic-tip Protocols. First, leaf tissues were washed with deionized water and blotted dry before freezing in liquid nitrogen. The tissues were ground into a fine powder using a pre-chilled mortar and pestle. The ground tissues (1 g) were resuspended in 25 mL cold Nuclei Isolation Buffer (NIB) and mixed until completely homogenized (15-30 min). The composition of NIB was as per Sikorskaite et al.[38]. The mixture was filtered through pre-wetted miracloth and left on ice for 20 min. The chlorophyll layer was separated by centrifugation at 18,000 rpm for 20 min at 4 °C. This layer was discarded, and only the pellet was kept. The pellet was resuspended in 25 mL NIB. The remaining chlorophyll layer was separated again by centrifugation. The pellet was resuspended in 2 mL lysis buffer (QIAGEN® Genomic-tip) before adding 4 µL DNAse-free RNase and incubating for 30 min at 37 °C. Proteinase K (0.8 mg/mL) was added to this mixed solution before incubating for one hour at 50 °C with gentle agitation. To remove insoluble debris, the solution was centrifuged for 30 min at 4,000 rpm. The supernatant was treated following QIAGEN® Genomic-tip Protocols. The genomic DNA in TE buffer (pH 7.6) was sent to the Australian Genome Research Facility Ltd (AGRF) for library preparation and PacBio Sequel I (PacBio Sequel System, RRID:SCR_017989).

Hi-C libraries were prepared using the Proximo Hi-C (Plant) Prep Kit (Phase Genomics, Seattle, WA, US). A *P. ovata* plant was incubated in the dark for 48 hours before the collection of leaf

material. Young leaves (0.2 g) were collected and chopped finely and immediately added to 10 mL of crosslinking solution to crosslink the chromatin. After 15 minutes of incubation, 100 µL of quenching solution was added, and the samples were incubated again for 20 min while rotating. The leaf material was pelleted by centrifugation, washed with 1 x CRB provided and patted dry before grinding into a fine powder. The ground leaf sample was suspended in cell lysis buffer to release the chromatin, followed by fragmentation of the chromatin, proximity ligation and library preparation which were carried out according to the Proximo Hi-C (Plant) Prep Kit protocol v.2. The final library concentration was determined using a Qubit fluorometer (Invitrogen, Carlsbad, CA, US), while the library quality and size was assessed with the LabChip GX Touch 24 using the HT DNA HiSens Dual Protocol Reagents (PerkinElmer, Hopkinton, MA, US). The final library had a median size of 570 bp. The libraries were sequenced on a HiSeq 2500 System (Illumina, San Diego, CA, US) by GENEWIZ (Suzhou, China) in paired-end mode, generating 150 bp reads (PE 150).

De novo genome assembly

Continuous long reads (CLRs) from the PacBio platform (~76X coverage) were used to assemble the *P. ovata* genome. Several steps were used to process raw reads, including removing unwanted reads, contig assembly, polishing, and purging haplotigs (alternative contigs) (Fig. 5). Firstly, seven unaligned subread BAM files from the PacBio Sequel I System were converted into FASTQ files using bam2fastx v1.3.0 (PacBio Sequel System, RRID:SCR_017989). Unwanted reads (mitochondria and chloroplasts) were removed by filtering out reads that mapped to either the chloroplast genome (GenBank: MH205737.1)[16] or the sole *P. ovata* mitochondrial gene available at the time (GenBank: EU069524.1). The read alignment was performed using Minimap2 v2.17 (Minimap2, RRID:SCR_018550) with "-ax map-pb" parameters and SAMtools v1.9 (SAMTOOLS, RRID:SCR_002105) used to extract unmapped reads. The total number of sequences and sequence lengths were checked before and after removing unwanted reads using FastQC v0.11.9 (FastQC, RRID:SCR_014583). Reads

after cleaning were assembled following a pipeline by Canu v2.0 (Canu, RRID:SCR_015880) with optimised parameters.

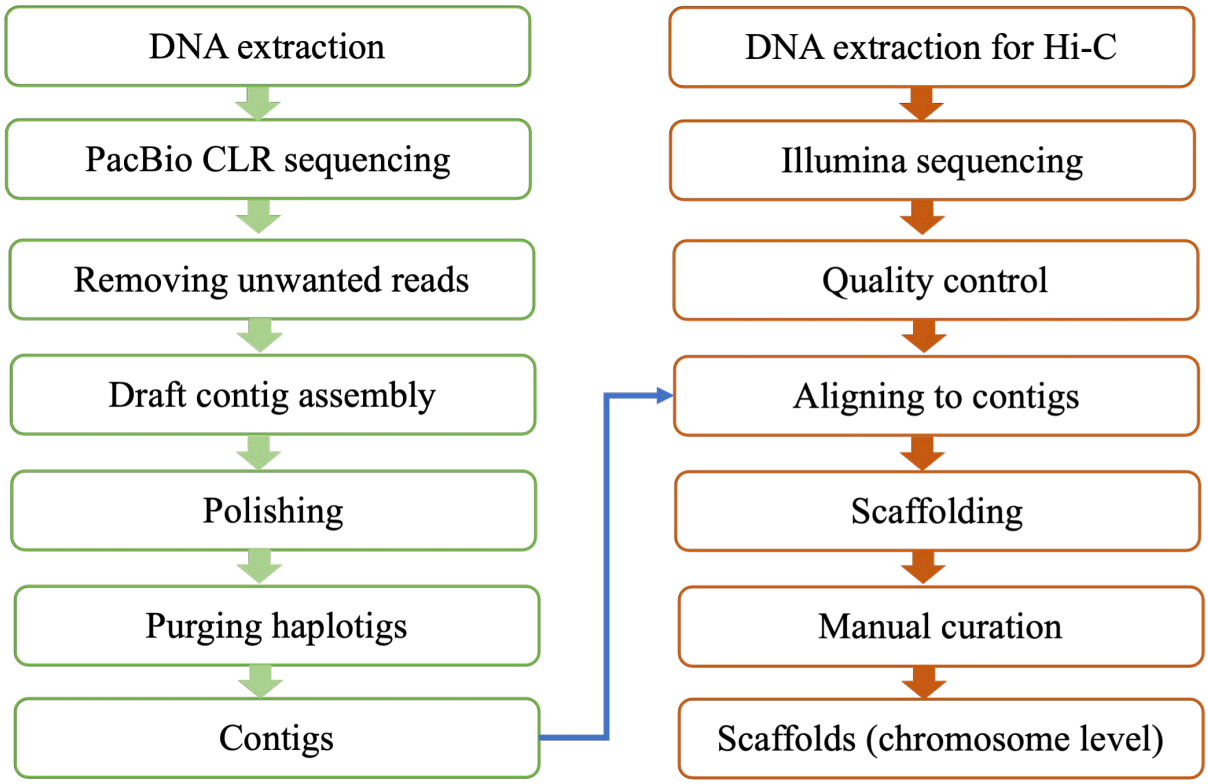


Figure 5: Illustration of the genome assembly strategy.

The draft contig assembly was polished with PacBio CLR reads. The subset of CLR reads used for polishing excluded the reads previously identified as being derived from the mitochondria or chloroplasts. Following the mapping of reads to the draft assembly, the polishing step was parallelised to decrease memory requirements and improve wall-time. This was achieved through a scatter-gather approach where each contig was independently processed. Filtered reads were mapped to the draft contig assembly using pbbam v1.6.0 before the polishing steps using pbgcpp v1.0.0 (PacBio Sequel System, RRID:SCR_017989). Circular and bubble contigs were removed from the polished contig assembly using seqtk v1.3 (Seqtk, RRID:SCR_018927).

After polishing, a purge was performed to remove haplotigs. First, the polished contig assembly was indexed, then the clean reads were mapped onto the improved assembly using Minimap2 v2.17 (Minimap2, RRID:SCR_018550) and sorted using SAMtools v1.9 (SAMTOOLS, RRID:SCR_002105). After mapping, purge_haplotigs v1.1.1 (Purge_haplotigs, RRID:SCR_017616) was used to detect and separate the primary and alternative contigs. To improve handling of repetitive regions, a list of contigs that were predicted as repeats from the Canu report [39] were parsed into the purge_haplotigs pipeline. Cut-offs were applied at "-l 5 -m 70 -h 190". The clipping option in purge_haplotigs was also used, to find and trim overlapping contigs that may prevent scaffolding.

Chromosome level assembly

Hi-C data was used to link contigs into a chromosome level assembly. First, the quality of the Hi-C reads was checked using FastQC v0.11.9 (FastQC, RRID:SCR_014583) and Trimmomatic v0.39 (Trimmomatic, RRID:SCR_011848) was used to remove primer sequences. To assess library preparation quality, the pipeline suggested by Phase Genomics (https://phasegenomics.github.io/2019/09/19/hic-alignment-and-qc.html) was followed. The contig assembly was indexed and clean Hi-C reads were aligned to the assembly using bwa v0.7.17 (BWA, RRID:SCR_010910). Reads derived from PCR duplicates were subsequently identified and flagged using samblaster v0.1.26 (SAMBLASTER, RRID:SCR_000468). Read alignments where the read is unmapped, the mate is unmapped, is not a primary alignment, or is a supplementary alignment (SAMtools parameter "-F 2316") were discarded. The mapped reads were also filtered using matlock v20181227 (https://github.com/phasegenomics/matlock) with default parameters and the QC of the reads were checked before and after filtering (Supplementary file 9). Although both QC reports showed that the Hi-C library was good quality, the filtering increased the numbers of high quality read pairs from 38.68% to 100%. Following the mapping of Hi-C reads, two different tools (SALSA2[17] and 3D-DNA[18]) were tested for scaffolding performance. Only 3D-DNA yielded chromosome-scale superscaffolds.

Firstly, aligning Hi-C clean reads to contig assembly was done using Juicer (Juicer, RRID:SCR_017226). After running the 3D-DNA pipeline, candidate assembly was visualised and reviewed using Juicebox Assembly Tools (JBAT)[18]. Then, a new assembly was generated by running a 3D-DNA post review pipeline. To meet NCBI submission requirements, we removed sequences with less than 200 nucleotides (nt) and reduced the unknown gap length (NNN) from 500 nt to 100 nt. Chromosomes were numbered from 1 to 4 from longest to shortest.

Genome size prediction and assembly quality

Publically available Illumina genomic data under SRA number SRR10076762 generated by the CSIR-Central Salt and Marine Chemicals Research Institute, India were used to estimate genome size by running genomescope2 (GenomeScope, RRID:SCR_017014). The genotype was GI-20. A test was performed to see if the reads from GI-20 could be mapped to our genome assembly by running Minimap2 v2.17 (Minimap2, RRID:SCR_018550), then sorting and counting mapped reads using SAMtools v1.9 (SAMTOOLS, RRID:SCR_002105). The frequency distribution of 21-, 31-, 39-, 45-, 50-, 70- and 100-mers was counted using jellyfish v2.2.10 (Jellyfish, RRID:SCR_005491). The histograms were further processed using genomescope2 to estimate the genome size, repeat content, and heterozygosity.

The quality of our assembly was assessed using the following parameters: assembly contiguity, gene set completeness, mapping rates of genomic and transcriptomic reads, and assembly continuity. The genome size and N50 value of each assembly stage were calculated using Perl script “n50.pl”[40]. Assembly completeness was measured by Benchmarking Universal Single-Copy Orthologs v4.1.4 (BUSCO, RRID:SCR_015008) against a eudicot database (2,326 genes)[41]. The assembly continuity was also evaluated by calculating the Long Terminal Repeat Assembly Index (LAI) using LTR_retriever[20].

Repeat content estimation and identification

Repeats were identified using RepeatModeler v2.6.1 (RepeatModeler, RRID:SCR_015027) and calculated by RepeatMasker v4.1.1 (RepeatMasker, RRID:SCR_012954). Firstly, a custom repeat library (CRL) was built by running RepeatModeler on the *P. ovata* genome against the Dfam transposable element family database. Repeats derived from protein-coding regions were removed from the library. Viridiplantae protein-encoding sequences were obtained from the UniProt Knowledgebase (UniProtKb) database (https://www.uniprot.org/uniprot/). The transposable element homolog sequences were detected by transposonPSI.pl (http://transposonpsi.sourceforge.net/) and removed by gaas_fasta_removeSeqFromIDlist.pl[42] from the collected proteome. The filtered proteome was segregated from the repeat library by searching the homologies using BLASTX (BLASTX, RRID:SCR_001653) and excluding them via ProtExcluder.pl[43]. The filtered CRL was then used to calculate the *P. ovata* genome by RepeatMasker following the tutorial by Dainat (https://www.biostars.org/p/411101/#411101). In the final annotated genome, remaining repeat annotations overlapping with protein-coding genes were removed manually based on NCBI's discrepancy report.

Generating a gene model or prediction

P. ovata gene models were predicted using evidence from a set of RNA-seq data obtained from a range of tissues (88 fastq files) (Supplementary file 1: Table S2). The data was grouped depending on how the library was prepared (Supplementary file 1: Table S2). The first category is paired stranded libraries with Illumina sequencing (56 files). The second contains paired un-stranded libraries with Illumina sequencing (2 files). The third group is a single-stranded library with Illumina (8 files). The fourth is a single un-stranded library with Illumina (12 files) while the fifth group is a single un-stranded library with Roche 454 sequencing (10 files).

The RNA-seq data was filtered by quality checking, trimming, cleaning, and aligning reads to the reference genome to generate accurate gene models. Quality checking was performed using FastQC v0.11.9 (FastQC, RRID:SCR_014583) and MultiQC v1.8 (MultiQC,

RRID:SCR_014982) with default parameters. Trimmomatic v0.39 (Trimmomatic, RRID:SCR_011848) was used to remove adapter and PCR primer fragments. Two read groups were treated in the paired-end mode and the other three groups with single-end mode. To remove contaminants, BBDuk, BBmap v38.87 (BBmap, RRID:SCR_016965) was used. Rates of contamination (1.5-94.5%) and mapping (54.9-96.1%) varied across samples (Supplementary file 1: Table S2). Contamination rates were higher for leaf, bract, stem, and capsule tissues (>20%) than for integument and ovaries (< 20 %). To generate a high-quality genome annotation, only RNA-seq data with a high mapping rate of greater than 85% was used, and to be consistent two samples with contamination rates of more than 95% were removed leaving 46 samples (71 fastq files).

Clean reads were aligned to the reference genome using STAR v2.7.6a (STAR, RRID:SCR_015899). After mapping reads to the reference genome, transcripts were generated separately for each group using Cufflinks v2.2.1 (Cufflinks, RRID:SCR_014597). Then, all transcripts were merged using gffcompare[44] to generate a gene model. All scripts were written in Snakemake v5.26.1 (Snakemake, RRID:SCR_003475).

Identification of coding and non-coding genes

To annotate protein-coding genes, we combined pipelines from TransDecoder (TransDecoder, RRID:SCR_017647) and Trinotate (Trinotate, RRID:SCR_018930) in Snakemake v5.26.1 (Snakemake, RRID:SCR_003475). Transdecoder was used to identify coding regions within transcripts, while Trinotate was employed for automatic functional annotation. BLASTP (BLASTP, RRID:SCR_001010) and BLASTX (BLASTX, RRID:SCR_001653) were used to search homologous genes against a local database created from the UniProtKB database (Viridiplantae). Another Gff/Gtf Analysis Toolkit (AGAT)[45] were used to fix the coding gene annotation. This was achieved by using agat_convert_sp_gxf2gxf.pl to standardise the annotation file, agat_sp_add_start_and_stop.pl to add start and stop codons, and agat_sp_filter_incomplete_gene_coding_models.pl to keep only complete coding genes[45].

Three non-coding RNA databases and three bioinformatics tools were used to search and annotate non-coding RNA using genomic and transcript sequences. A local database was built from RNAcentral[46], a plant non-coding RNA database PNRD[47], and CANTATAdb[48] and sequence homologies identified using BLASTN (BLASTN, RRID:SCR_001598). We also used tools RNAMMER (RNAmmer, RRID:SCR_017075) for ribosomal RNA (rRNA), tRNAscan-SE v2.0.7 (tRNAscan-SE, RRID:SCR_010835) for transfer RNA (tRNA), and FEELnc[49] for long non-coding RNA (lncRNA) detection. Finally, the validated annotation results for coding and non-coding genes and repeat regions were combined using GenomeTools v1.2.1 (GenomeTools, RRID:SCR_016120). The NCBI tool (tbl2asn, RRID:SCR_016636) was used to manually evaluate and fix errors in the sequences in preparation for genome submission.

After NCBI accepted the annotated genome, the GC content and numbers of CDS, mRNA, and exon features were calculated. To do this, the features were extracted using the script agat_sp_extract_sequences.pl from AGAT[45] followed by gaas_fasta_statistics.pl from Genome Assembly Annotation Service (GAAS)[42] to calculate the GC content. We also used EMBOSS infoseq (EMBOSS, RRID:SCR_008493) to calculate the GC content of each CDS.

Eighteen genes from the glycosyltransferase (GT) 61 family previously identified in different *P. ovata* tissues including mucilage-producing tissues[33,34] were mapped to the genome assembly using GMAP v2021.08.25 (GMAP, RRID:SCR_008992). Seven genes from *PoGT61_1* to *PoGT61_7* (KC894060 to KC894066) were obtained from Jensen et al.[34]. Eleven genes, namely *PoGT61_1L*, *PoGT61_4L*, *PoGT61_8* to *PoGT61_11*, *PoGT61_11L*, *PoGT61_12*, *PoGT61_13*, *PoGT61_17*, and *PoXYLT* (KY440071 to KY440081, respectively) were identified from Phan et al.[33]. Mapping coordinates of these genes to the assembly were compared to the *P. ovata* annotation generated from this study.

To build a phylogenetic tree, ten *Arabidopsis thaliana* and 26 *Oryza sativa* spp. *japonica* coding sequences were extracted from the EnsemblPlants database using PFAM ID PF04577[36]

(https://plants.ensembl.org/biomart/martview/). EMBOSS Transeq (EMBOSS, RRID:SCR_008493) was used to translate nucleic acid sequence before performing multiple sequence alignment using MUSCLE v3.8.1551 (MUSCLE, RRID:SCR_011812) on GT61 protein sequences. Phylogenetically informative regions contained in the alignment were selected using BMGE v1.12 (Block Mapping and Gathering with Entropy) with parameters "-t AA -m BLOSUM62"[50]. A phylogenetic tree was built from the aligned sequences using FastTree v2.1.10 (FastTree, RRID:SCR_015501) and visualised using FigTree v1.4.4 (FigTree, RRID:SCR_008515).

Data availability

The datasets generated during this study were deposited in the NCBI SRA (Sequence Read Archive) database under the BioProject ID: PRJNA732452. The genome sequence data (PacBio and Hi-C) are available under accession numbers SRR14643405 and SRR14643406. Transcriptome data are available under accession numbers SRR14643399-SRR14643404 and SRR14643407-SRR14643436. Metadata and permanent links of previously published datasets analysed during the current study are listed in Supplementary file1: Table S1. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JAHHQI000000000. The version described in this paper is version JAHHQI010000000, which includes the annotation. Source code and software used for the analyses are included within the article (see also Supplementary file 1: Table S9 and Supplementary file 10).

References

1 Gonçalves, S. & Romano, A. The medicinal potential of plants from the genus *Plantago* (Plantaginaceae). *Ind Crops Prod*. **83**, 213-226 (2016).

2 Phan, J. L. *et al.* The novel features of *Plantago ovata* seed mucilage accumulation, storage and release. *Sci. Rep.* **10**, 1-14 (2020).

3 Cowley, J. M. & Burton, R. A. The goo-d stuff: *Plantago* as a myxospermous model with modern utility. *New Phytol*. **229**, 1917-1923 (2021).

4 Cowley, J. M. *et al.* A small-scale fractionation pipeline for rapid analysis of seed mucilage characteristics. *Plant Methods* **16**, 1-12 (2020).

5 Patel, D., Patel, H., Patel, P., Patel, H. & Amin, A. Evaluation of stable and non shattering isabgol cultivar-Gujarat isabgol 4. *JOSAC* 88-90 https://doi.org/10.25081/josac.2018.v27.i1.1022 (2018).

6 McNeil D. A preliminary report on work conducted in 1985 to evaluate *Plantago ovata* as a potential crop in the Ord River irrigation area. https://researchlibrary.agric.wa.gov.au/pubns/24/ (1985).

7 Kumar, M. *et al.* Phenotypic and molecular characterization of selected species of *Plantago* with emphasis on *Plantago ovata*. *Aust. J. Crop Sci.* **8**, 1639 (2014).

8 Shahriari, Z., Heidari, B., Dadkhodaie, A. & Richards, C. M. Analysis of karyotype, chromosome characteristics, variation in mucilage content and grain yield traits in *Plantago ovata* and *P. psyllium* species. *Ind Crops Prod.* **123**, 676-686 (2018).

9 Dhar, M., Kaul, S., Sareen, S. & Koul, A. *Plantago ovata*: Genetic diversity, cultivation, utilization and chemistry. *Plant Genet. Resour.* **3**, 252-263 (2005).

10 Pramanik, S. & Raychaudhuri, S. S. DNA content, chromosome composition, and isozyme patterns in *Plantago* L. *Bot. Rev*. **63**, 124-139 (1997).

11 Dhar, M., Kaul, S., Friebe, B. & Gill, B. Chromosome identification in *Plantago ovata* Forsk. through C-banding and FISH. *Curr. Sci.* 150-152 (2002).

12 Dhar, M., Fuchs, J. & Houben, A. Distribution of eu-and heterochromatin in *Plantago ovata*. *Cytogenet. Genome Res*. **125**, 235-240 (2009).

13 Saha, P., Das, D., Roy, S., Chakrabarti, A. & Sen Raychaudhuri, S. Effect of gamma irradiation on metallothionein protein expression in *Plantago ovata* Forsk. *Int. J. Radiat. Biol*. **89**, 88-96 (2013).

14 Lal, R. K. *et al. Plantago ovata* plant named 'Mayuri'. Google Patents https://patents.google.com/patent/USPP17505P3/en (2017).

15 Tucker, M. *et al.* Dissecting the genetic basis for seed coat mucilage heteroxylan biosynthesis in *Plantago ovata* using gamma irradiation and infrared spectroscopy. *Front. Plant Sci.* **8** (2017).

16 Li, S., Sun, H. & Wang, K. The complete chloroplast genome sequence of *Plantago ovata*. *Mitochondrial DNA Part B.* **4**, 346-347 (2019).

17 Ghurye, J. *et al.* Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput. Biol.* **15**, e1007273; 10.1371/journal.pcbi.1007273 (2019).

18 Dudchenko, O. *et al*. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. Preprint at https://www.biorxiv.org/content/10.1101/254797v1 (2018).

19 Price, A. & Gibas, C. The quantitative impact of read mapping to non-native reference genomes in comparative RNA-Seq studies. *PloS One*. **12**, e0180904; 10.1371/journal.pone.0180904 (2017).

20 Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126-e126; 10.1093/nar/gky730 (2018).

21 Michalovova, M., Vyskot, B. & Kejnovsky, E. Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. *Heredity (Edinb)* **111**, 314-320; 10.1038/hdy.2013.51 (2013).

22 Badr, A., Labani, R. & Elkington, T. Nuclear DNA variation in relation to cytological features of some species in the genus *Plantago* L. *Cytologia* **52**, 733-737; 10.1508/cytologia.52.733 (1987).

23 Schmuths, H., Meister, A., Horres, R. & Bachmann, K. Genome size variation among accessions of *Arabidopsis thaliana*. *Ann. Bot.* **93**, 317-321 (2004).

24 Šmarda, P. *et al.* Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E4096 (2014).

25 Singh, R., Ming, R. & Yu, Q. Comparative analysis of GC content variations in plant genomes. *Trop. Plant Biol.* **9** (2016).

26 Šmarda, P., Bureš, P., Šmerda, J. & Horová, L. Measurements of genomic GC content in plant genomes with flow cytometry: a test for reliability. *New Phytol*. **193**, 513-521 (2012).

27 Wang, J. *et al.* Genome-wide nucleotide patterns and potential mechanisms of genome divergence following domestication in maize and soybean. *Genome Biol.* **20**, 74 (2019).

28 Cowley, J. M., O'Donovan, L. A. & Burton, R. A. The composition of Australian Plantago seeds highlights their potential as nutritionally-rich functional food ingredients. *Sci. Rep.* **11**, 12692 (2021).

29 Kotwal, S. *et al. De novo* transcriptome analysis of medicinally important *Plantago ovata* using RNA-Seq. *PloS One*. **11**, e0150273 (2016).

30 Sundararajan, A. *et al.* Gene evolutionary trajectories and GC patterns driven by recombination in *Zea mays*. *Front. Plant Sci*. **7**, 1433 (2016).

31 Dhar, M. K., Friebe, B., Kaul, S. & Gill, B. S. Characterization and physical mapping of ribosomal RNA gene families in *Plantago*. *Ann. Bot*. **97**, 541-548 (2006).

32 Udall, J. A. & Dawe, R. K. Is it ordered correctly? validating genome assemblies by optical mapping. *Plant Cell* **30**, 7-14 (2018).

33 Phan, J. L. *et al.* Differences in glycosyltransferase family 61 accompany variation in seed coat mucilage composition in *Plantago* spp. *J. Exp. Bot.* **67**, 6481-6495 (2016).

34 Jensen, J. K., Johnson, N. & Wilkerson, C. G. Discovery of diversity in xylan biosynthetic genes by transcriptional profiling of a heteroxylan containing mucilaginous tissue. *Front. Plant Sci.* **4**, 183-183 (2013).

35 Jensen, J. K., Johnson, N. R. & Wilkerson, C. G. *Arabidopsis thaliana* IRX10 and two related proteins from psyllium and *Physcomitrella patens* are xylan xylosyltransferases. *Plant J.* **80**, 207-215 (2014).

36 Anders, N. *et al.* Glycosyl transferases in family 61 mediate arabinofuranosyl transfer onto xylan in grasses. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 989-993 (2012).

37 Peska, V. & Garcia, S. Origin, diversity, and evolution of telomere sequences in plants. *Front. Plant Sci.* **11**, 117 (2020).

38 Sikorskaite, S., Rajamäki, M.-L., Baniulis, D., Stanys, V. & Valkonen, J. P. Protocol: optimised methodology for isolation of nuclei from leaves of species in the Solanaceae and Rosaceae families. *Plant Methods* **9**, 1-9 (2013).

39 Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. **27**, 722-736 (2017).

40 Telatin, A., Fariselli, P. & Birolo, G. SeqFu: a suite of utilities for the robust and reproducible manipulation of sequence files. *Bioengineering* **8**, 59 (2021).

41 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212 (2015).

42 Daniat, J., Binzer-Panchal, M., Olsen, R. A. et al. NBISweden/GAAS: GAAS-v1.2.0 (v1.2.0). Zenodo https://doi.org/10.5281/zenodo.3835504 (2020).

43 Campbell, M. S. *et al.* MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164**, 513-524 (2014).

44 Pertea, G. & Pertea, M. GFF utilities: GffRead and GffCompare [version 2; peer review: 3 approved]. *F1000research* **9**, 304 https://doi.org/10.12688/f1000research.23297.2 (2020).

45 Dainat J, Hereñú D, Pascal-git. NBISweden/AGAT: AGAT-v0.8.0 (v0.8.0). Zenodo https://doi.org/10.5281/zenodo.5336786 (2021).

46 The Rnacentral Consortium. RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Res.* **47**, D221-D229 (2019).

47 Yi, X., Zhang, Z., Ling, Y., Xu, W. & Su, Z. PNRD: a plant non-coding RNA database. *Nucleic Acids Res.* **43**, D982-D989 (2015).

48 Szcześniak, M. W., Rosikiewicz, W. & Makałowska, I. CANTATAdb: a collection of plant long non-coding RNAs. *Plant Cell Physiol.* **57**, e8-e8 (2016).

49 Wucher, V. *et al.* FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res.* **45**, e57-e57 (2017).

50 Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).

51 Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088-3090 (2017).

Acknowledgements

The authors thank Dr Fabien Voisin for technical support in utilising Phoenix-HPC. We recognise Pastor Julian for his work on *P. ovata* genomic short read data. We acknowledge the useful discussions provided by Aaron L. Phillips in contig assembly. This study was supported by the Australian Research Council (ARC) Centres of Excellence in Plant Cell Walls (CE110001007), Plant Energy Biology (CE140100008) and Linkage Grant (LP180100971). This work was also supported with supercomputing resources provided by the Phoenix HPC service at the University of Adelaide. LH is supported by the University of Adelaide's Adelaide Graduate Research Scholarship (AGRS) and The National Research and Innovation Agency (BRIN-Indonesia).

Authors' information

Affiliations

School of Agriculture, Food and Wine, Waite Research Institute, University of Adelaide, Waite Campus, Urrbrae, SA, Australia

Lina Herliana, Julian G. Schwerdt, Tycho R. Neumann, Jana L. Phan, James M. Cowley, Neil J. Shirley, Matthew R. Tucker, Tina Bianco-Miotto & Rachel A. Burton

South Australian Genomics Centre (SAGC), SA, Australia

Nathan S. Watson-Haigh

Australian Genome Research Facility, Victorian Comprehensive Cancer Centre, Melbourne, VIC 3000, Australia

Nathan S. Watson-Haigh

IP Australia, PO Box 200, Woden, ACT, 2606, Australia

Tycho R. Neumann

School of Biological Sciences, University of Western Australia, Crawley, WA 6009, Australia

Anita Severn-Ellis & Jacqueline Batley

Contributions

R.A.B. and J.G.S. conceived the project. R.A.B., N.S.W., and T.B. supervised the study and revised the manuscript. L.H. and N.S.W. developed workflows and analysed the data. L.H. performed genome assembly to annotation and wrote the draft manuscript. T.R.N. performed DNA extraction and library preparation for PacBio CLR sequencing. J.B. and A.S. performed DNA extraction and library preparation for Hi-C experiments. J.G.S., J.L.P., M.R.T., J.M.C., and N.J.S. provided RNA-seq data for this study. All authors read, edited, and approved the manuscript.

Corresponding authors

Correspondence to Rachel A. Burton and Nathan S. Watson-Haigh

Additional information

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare no competing interests

Supplementary File 1: Table S1. Summary of RNA-seq and genomic data on *P. ovata*.

Accession number	Link
SRR066373	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR066373
SRR066374	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR066374
SRR066375	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR066375
SRR066376	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR066376
SRR342350	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR342350
SRR342351	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR342351
SRR629688	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR62968
SRR1311174	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR1311174
SRR1311175	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR1311175
SRR1311176	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR1311176
SRR1311177	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR1311177
SRR3883622	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR3883618
SRR3883620	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR3883619
SRR3883621	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR3883620
SRR3883618	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR3883621
SRR3883619	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR3883622
SRR3885726	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR3885726
SRR3885727	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR3885727
SRR3885728	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR3885728
SRR5434206	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR5434206
SRR5434207	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR5434207
SRR5434208	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR5434208
SRR5434209	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR5434209
SRR5434211	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR5434210
SRR5434210	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR5434211
SRR5434213	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR5434212
SRR5434212	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR5434213
SRR10076762	https://trace ncbi nlm nih.gov/Traces/sra/?run=SRR10076762

Supplementary File 1: Table S2. Statistic of contamination and mapping rates of RNA-seq data.

No	Accession number	Tissues	Group	Contamination rate	Mapping rate	Note
1	SRR5434207 + SRR5434206	Leaves	1	35.50%	88.90%	I
2	SRR5434209 + SRR5434208	Leaves	1	94.50%	67.80%	E
3	SRR5434211 + SRR5434210	Leaves	1	83.00%	89.00%	I
4	SRR5434213 + SRR5434212	Leaves	1	41.90%	81.50%	E
5	SRR14643432	Capsules	1	27.20%	91.70%	I
6	SRR14643431	Capsules	1	29.30%	92.70%	I
7	SRR14643430	Capsules	1	33.10%	92.20%	I
8	SRR14643429	Capsules	1	26.80%	92.30%	I
9	SRR14643428	Capsules	1	96.60%	92.60%	E
10	SRR14643427	Capsules	1	30.00%	92.10%	I
11	SRR14643426	Capsules	1	32.30%	92.00%	I
12	SRR14643425	Capsules	1	25.60%	91.90%	I
13	SRR14643423	Capsules	1	31.70%	90.50%	I
14	SRR14643422	Capsules	1	26.70%	91.80%	I
15	SRR14643421	Capsules	1	25.80%	92.50%	I
16	SRR14643420	Capsules	1	29.50%	92.10%	I
17	SRR14643419	Capsules	1	27.10%	91.80%	I
18	SRR14643418	Capsules	1	26.20%	91.20%	I
19	SRR14643417	Capsules	1	95.50%	91.80%	E
20	SRR14643416	Capsules	1	25.50%	90.20%	I
21	SRR14643415	Capsules	1	27.80%	90.90%	I
22	SRR14643414	Capsules	1	25.40%	90.60%	I
23	SRR14643412	Capsules	1	26.10%	91.30%	I
24	SRR14643411	Capsules	1	34.30%	91.90%	I
25	SRR14643410	Capsules	1	26.00%	91.90%	I
26	SRR14643409	Capsules	1	27.50%	91.80%	I
27	SRR14643408	Capsules	1	24.70%	92.20%	I
28	SRR14643407	Capsules	1	29.50%	92.50%	I
29	SRR629688	Ovaries	2	9.00%	92.40%	I
30	SRR14643436	Integument	3	3.00%	95.50%	I
31	SRR14643424	Integument	3	2.70%	95.50%	I
32	SRR14643404	Integument	3	1.90%	96.00%	I
33	SRR14643402	Integument	3	2.70%	96.00%	I
34	SRR14643435	Integument	3	3.00%	95.50%	I
35	SRR14643413	Integument	3	2.70%	95.50%	I

36	SRR14643403	Integument	3	1.90%	96.00%	I
37	SRR14643401	Integument	3	2.70%	96.10%	I
38	SRR3883618	Integument	4	1.50%	94.50%	I
39	SRR3883619	Leaves	4	81.30%	93.40%	I
40	SRR3883620	Leaves	4	85.50%	90.80%	I
41	SRR3883621	Bractea	4	83.10%	91.70%	I
42	SRR3883622	Leaves, stems, flowers, roots	4	73.50%	92.70%	I
43	SRR3885726	Integument	4	1.50%	95.10%	I
44	SRR3885727	Integument	4	1.50%	95.00%	I
45	SRR3885728	Integument	4	1.50%	95.20%	I
46	SRR14643400	Integument	4	5.00%	94.90%	I
47	SRR14643399	Integument	4	5.40%	94.70%	I
48	SRR14643434	Integument	4	4.40%	95.00%	I
49	SRR14643433	Integument	4	2.60%	95.30%	I
50	SRR066373	Integument	5	17.40%	54.90%	E
51	SRR066374	Stems	5	46.90%	86.70%	I
52	SRR066375	Integument	5	18.80%	56.60%	E
53	SRR066376	Integument	5	9.90%	63.20%	E
54	SRR342350	Integument	5	15.80%	58.00%	E
55	SRR342351	Integument	5	6.30%	64.20%	E
56	SRR1311174	Leaves	5	87.60%	77.00%	E
57	SRR1311175	Leaves	5	87.10%	77.90%	E
58	SRR1311176	Leaves	5	83.70%	75.50%	E
59	SRR1311177	Leaves	5	85.50%	63.50%	E

Group 1 = paired stranded libraries with Illumina sequencing (No. 1 to 28, 56 files); Group 2 = paired un-stranded libraries with Illumina sequencing (No. 29, 2 files); Group 3 = a single-stranded library with Illumina (No. 30 to 37, 8 files); Group 4 = a single un-stranded library with Illumina (No. 38 to 49, 12 files); Group 5 = a single un-stranded library with Roche 454 sequencing (No. 50 to59, 10 files); E: Excluded; I: Included in the analysis. Data number 1-4, 29, 38-45, and 50-59 were collected from SRA NCBI while data number 5-28, 30-37, and 46-49 were submitted to NCBI during this study. Contamination refers to reads derived from chloroplast, mitochondria, and ribosomal RNA (rRNA).

Supplementary File 1: Table S3. Summary of repeat content.

Repeat types	Number of elements	Length	Percentage
Retroelements	109,391	209,201,223 bp	41.76%
LINEs	16,218	8,608,936 bp	1.72%
CRE/SLACS	2,116	1,726,768 bp	0.34%
L2/CR1/Rex	1,818	186,338 bp	0.04%
L1/CIN4	12,051	6,607,560 bp	1.32%
LTR elements	93,173	200,592,287 bp	40.04%
BEL/Pao	12	1,946 bp	0.00%
Ty1/Copia	41,829	98,644,380 bp	19.69%
Gypsy/DIRS1	50,802	101,643,016 bp	20.29%
DNA transposons	17,524	9,038,595 bp	1.80%
Hobo-Activator	2,162	559,504 bp	0.11%
Tc1-IS630-Pogo	697	165,153 bp	0.03%
Tourist/Harbinger	3,269	2,160,345 bp	0.43%
Rolling circles	2,994	1,357,620 bp	0.27%
Unclassified	259,941	85,448,836 bp	17.06%
Total interspersed repeats		303,688,654 bp	60.62%
Small RNA	2,373	854,404 bp	0.17%
Satellites	366	98,902 bp	0.02%
Simple repeats	78,701	3,488,223 bp	0.70%
Low complexity	13,014	613,727 bp	0.12%
Total all repeats		310,101,530 bp	61.90%

LINEs: long interspersed nuclear elements; LTR: long terminal repeat

Supplementary File 1: Table S4. Statistics of genome annotation.

Metric	Coding	tRNA	rRNA	lncRNA
Number of genes	23,346	1,295	108	213
Number of transcripts	63,228	1,295	108	328
Number of CDSs	63,228			
Mean transcripts per gene	2.7	1	1	1
Total gene length	95,628,788	94,317	51,752	633,634
Total transcript length	251,575,257	94,317	51,752	890,463
Total CDS length	61,502,350			
Total exon length	142,244,532			
Mean gene length	4,096	72	479	2,974
Mean transcript length	3,978	72	479	2,714
Mean CDS length	972			
Mean exon length	445			
Longest gene	54,497	100	6,820	19,444
Longest transcript	54,497	100	6,820	14,891
Longest CDS	15,276			
Longest exon	39,474			
Shortest gene	175	56	101	359
Shortest transcript	175	56	101	359
Shortest CDS	117			
Shortest exon	3			11

CDS: Coding sequence

Supplementary File 1: Table S5. Identified location of *GT61* genes on current *P. ovata* assembly.

Gene	Gene id	Chr 1	Scaffold	Start	End	Number of transcripts	Predicted protein product
PoGT61_12	KN361_1g003887	Chr 1	HiC_scaffold_1	25177363	25178827	1	xylan arabinosyltransferase 3
PoGT61_7	KN361_1g004725	Chr 1	HiC_scaffold_1	30917201	30919297	3	xylan arabinosyltransferase 2
PoGT61_2	KN361_1g004768	Chr 1	HiC_scaffold_1	31250497	31255465	4	xylan arabinosyltransferase 2
PoGT61_17	KN361_1g004772	Chr 1	HiC_scaffold_1	31270273	31272423	5	xylan arabinosyltransferase 2
PoGT61_9	KN361_1g006024	Chr 1	HiC_scaffold_1	40239202	40242594	3	xylan glycosyltransferase
PoXYLT	KN361_2g007734	Chr 2	HiC_scaffold_2	118259317	118264629	5	beta-1,2-xylosyltransferase
PoGT61_18	KN361_2g009231	Chr 2	HiC_scaffold_2	127439038	127440877	2	beta-1,2-xylosyltransferase
PoGT61_13	KN361_3g000067	Chr 3	HiC_scaffold_3	392963	396990	9	xylan arabinosyltransferase 3
PoGT61_10	KN361_3g006297	Chr 3	HiC_scaffold_3	58554890	58558304	3	xylan arabinosyltransferase 3
PoGT61_11L	KN361_4g000326	Chr 4	HiC_scaffold_4	2712573	2715212	3	xylan arabinosyltransferase 3
PoGT61_6	KN361_4g000365	Chr 4	HiC_scaffold_4	2971589	2974674	5	xylan arabinosyltransferase 3
PoGT61_1	KN361_4g000366	Chr 4	HiC_scaffold_4	2985315	2988043	2	xylan arabinosyltransferase 3
PoGT61_1L	KN361_4g000367	Chr 4	HiC_scaffold_4	2997105	3002929	2	xylan arabinosyltransferase 3
PoGT61_4 *PoGT61_4L*	KN361_4g000369	Chr 4	HiC_scaffold_4	3008109	3010915	2	xylan arabinosyltransferase 3
PoGT61_3	KN361_4g000371	Chr 4	HiC_scaffold_4	3040667	3042696	3	xylan arabinosyltransferase 3
PoGT61_5	KN361_4g000373	Chr 4	HiC_scaffold_4	3081109	3083405	2	xylan arabinosyltransferase 3
PoGT61_19	KN361_4g000374	Chr 4	HiC_scaffold_4	3089291	3091530	3	xylan arabinosyltransferase 3
PoGT61_8	KN361_4g000380	Chr 4	HiC_scaffold_4	3112369	3114668	2	xylan arabinosyltransferase 3
PoGT61_11	KN361_4g000381	Chr 4	HiC_scaffold_4	3129586	3132465	4	xylan arabinosyltransferase 3

Red texts indicate novel genes.

Supplementary File 1: Table S6. Identification of lncRNA/mRNA pairs.

lncRNA_gene	partnerRNA_gene	direction	type	distance	subtype	location	group	description
KN361_nc573	KN361_3g004237	antisense	genic	0	nested	exonic	Group 1	antisense genic exonic
KN361_nc303	KN361_2g005900	antisense	genic	0	overlapping	exonic	Group 1	antisense genic exonic
KN361_nc237	KN361_1g006366	antisense	genic	0	overlapping	exonic	Group 1	antisense genic exonic
KN361_nc492	KN361_3g004932	antisense	genic	0	overlapping	exonic	Group 1	antisense genic exonic
KN361_nc585	KN361_4g001072	antisense	genic	0	overlapping	exonic	Group 1	antisense genic exonic
KN361_nc574	KN361_3g000296	antisense	genic	0	nested	exonic	Group 1	antisense genic exonic
KN361_nc184	KN361_1g001276	antisense	genic	0	nested	exonic	Group 1	antisense genic exonic
KN361_nc598	KN361_4g003611	antisense	genic	0	overlapping	exonic	Group 1	antisense genic exonic
KN361_nc225	KN361_1g011394	antisense	genic	0	overlapping	exonic	Group 1	antisense genic exonic
KN361_nc289	KN361_2g005443	antisense	genic	0	containing	exonic	Group 1	antisense genic exonic
KN361_nc277	KN361_1g003018	antisense	genic	0	nested	exonic	Group 1	antisense genic exonic
KN361_nc353	KN361_2g001120	antisense	genic	0	overlapping	exonic	Group 1	antisense genic exonic
KN361_nc409	KN361_2g007371	antisense	genic	0	overlapping	exonic	Group 1	antisense genic exonic
KN361_nc397	KN361_2g009394	antisense	genic	0	nested	exonic	Group 1	antisense genic exonic
KN361_nc192	KN361_1g002684	antisense	genic	0	overlapping	exonic	Group 1	antisense genic exonic
KN361_nc429	KN361_2g006713	antisense	genic	0	overlapping	exonic	Group 1	antisense genic exonic
KN361_nc426	KN361_2g000169	antisense	genic	0	nested	exonic	Group 1	antisense genic exonic
KN361_nc585	KN361_4g001072	antisense	genic	0	containing	exonic	Group 1	antisense genic exonic
KN361_nc170	KN361_1g000669	antisense	genic	0	overlapping	exonic	Group 1	antisense genic exonic
KN361_nc707	KN361_4g000096	antisense	genic	0	nested	exonic	Group 1	antisense genic exonic
KN361_nc754	KN361_0g000203	antisense	genic	0	overlapping	exonic	Group 1	antisense genic exonic
KN361_nc585	KN361_4g001072	antisense	genic	0	overlapping	exonic	Group 1	antisense genic exonic
KN361_nc544	KN361_3g009002	antisense	genic	0	nested	intronic	Group 2	antisense genic intronic

KN361_nc544	KN361_3g009002	antisense	genic	0	nested	intronic	Group 2	antisense genic intronic
KN361_nc275	KN361_1g011560	antisense	genic	0	overlapping	intronic	Group 2	antisense genic intronic
KN361_nc544	KN361_3g009002	antisense	genic	0	nested	intronic	Group 2	antisense genic intronic
KN361_nc129	KN361_1g007127	antisense	intergenic	1381	divergent	upstream	Group 3	antisense intergenic
KN361_nc490	KN361_3g006250	antisense	intergenic	5810	divergent	upstream	Group 3	antisense intergenic
KN361_nc123	KN361_1g009056	antisense	intergenic	831	convergent	downstream	Group 3	antisense intergenic
KN361_nc115	KN361_1g009509	antisense	intergenic	12884	convergent	downstream	Group 3	antisense intergenic
KN361_nc115	KN361_1g009509	antisense	intergenic	12881	convergent	downstream	Group 3	antisense intergenic
KN361_nc476	KN361_3g006396	antisense	intergenic	28293	divergent	upstream	Group 3	antisense intergenic
KN361_nc115	KN361_1g009509	antisense	intergenic	12881	convergent	downstream	Group 3	antisense intergenic
KN361_nc110	KN361_1g009221	antisense	intergenic	29964	divergent	upstream	Group 3	antisense intergenic
KN361_nc679	KN361_4g001353	antisense	intergenic	15126	convergent	downstream	Group 3	antisense intergenic
KN361_nc500	KN361_3g005786	antisense	intergenic	7143	convergent	downstream	Group 3	antisense intergenic
KN361_nc370	KN361_2g005418	antisense	intergenic	4569	divergent	upstream	Group 3	antisense intergenic
KN361_nc229	KN361_1g000075	antisense	intergenic	17426	convergent	downstream	Group 3	antisense intergenic
KN361_nc457	KN361_3g005925	antisense	intergenic	67205	convergent	downstream	Group 3	antisense intergenic
KN361_nc652	KN361_4g004061	antisense	intergenic	10896	divergent	upstream	Group 3	antisense intergenic
KN361_nc008	KN361_1g005822	antisense	intergenic	893	convergent	downstream	Group 3	antisense intergenic
KN361_nc336	KN361_2g005417	antisense	intergenic	90524	divergent	upstream	Group 3	antisense intergenic
KN361_nc011	KN361_1g007802	antisense	intergenic	251	convergent	downstream	Group 3	antisense intergenic
KN361_nc295	KN361_2g004297	antisense	intergenic	2198	divergent	upstream	Group 3	antisense intergenic
KN361_nc464	KN361_3g007217	antisense	intergenic	332	divergent	upstream	Group 3	antisense intergenic
KN361_nc336	KN361_2g005417	antisense	intergenic	96925	divergent	upstream	Group 3	antisense intergenic
KN361_nc422	KN361_2g005689	antisense	intergenic	12224	divergent	upstream	Group 3	antisense intergenic
KN361_nc191	KN361_1g001084	antisense	intergenic	606	divergent	upstream	Group 3	antisense intergenic
KN361_nc295	KN361_2g004297	antisense	intergenic	2197	divergent	upstream	Group 3	antisense intergenic
KN361_nc324	KN361_2g006689	antisense	intergenic	4133	divergent	upstream	Group 3	antisense intergenic

KN361_nc556	KN361_3g005356	antisense	intergenic	9965	divergent	upstream	Group 3	antisense intergenic
KN361_nc612	KN361_4g002509	antisense	intergenic	31159	divergent	upstream	Group 3	antisense intergenic
KN361_nc336	KN361_2g005417	antisense	intergenic	90524	divergent	upstream	Group 3	antisense intergenic
KN361_nc419	KN361_2g005337	antisense	intergenic	1232	divergent	upstream	Group 3	antisense intergenic
KN361_nc196	KN361_1g008687	antisense	intergenic	6926	convergent	downstream	Group 3	antisense intergenic
KN361_nc109	KN361_1g009171	antisense	intergenic	143	divergent	upstream	Group 3	antisense intergenic
KN361_nc455	KN361_3g005255	antisense	intergenic	339	divergent	upstream	Group 3	antisense intergenic
KN361_nc679	KN361_4g001353	antisense	intergenic	15125	convergent	downstream	Group 3	antisense intergenic
KN361_nc464	KN361_3g007217	antisense	intergenic	378	divergent	upstream	Group 3	antisense intergenic
KN361_nc610	KN361_4g000427	antisense	intergenic	24360	convergent	downstream	Group 3	antisense intergenic
KN361_nc464	KN361_3g007217	antisense	intergenic	526	divergent	upstream	Group 3	antisense intergenic
KN361_nc345	KN361_2g007696	antisense	intergenic	143	convergent	downstream	Group 3	antisense intergenic
KN361_nc540	KN361_3g001121	antisense	intergenic	2619	convergent	downstream	Group 3	antisense intergenic
KN361_nc115	KN361_1g009509	antisense	intergenic	12884	convergent	downstream	Group 3	antisense intergenic
KN361_nc464	KN361_3g007217	antisense	intergenic	315	divergent	upstream	Group 3	antisense intergenic
KN361_nc464	KN361_3g007217	antisense	intergenic	443	divergent	upstream	Group 3	antisense intergenic
KN361_nc680	KN361_4g002770	antisense	intergenic	54124	convergent	downstream	Group 3	antisense intergenic
KN361_nc679	KN361_4g001353	antisense	intergenic	15125	convergent	downstream	Group 3	antisense intergenic
KN361_nc588	KN361_4g002261	antisense	intergenic	47805	divergent	upstream	Group 3	antisense intergenic
KN361_nc011	KN361_1g007802	antisense	intergenic	317	convergent	downstream	Group 3	antisense intergenic
KN361_nc428	KN361_2g005801	antisense	intergenic	37740	divergent	upstream	Group 3	antisense intergenic
KN361_nc340	KN361_2g003945	antisense	intergenic	18193	divergent	upstream	Group 3	antisense intergenic
KN361_nc511	KN361_3g009077	antisense	intergenic	7417	convergent	downstream	Group 3	antisense intergenic
KN361_nc115	KN361_1g009509	antisense	intergenic	12884	convergent	downstream	Group 3	antisense intergenic
KN361_nc540	KN361_3g001121	antisense	intergenic	2175	convergent	downstream	Group 3	antisense intergenic
KN361_nc464	KN361_3g007217	antisense	intergenic	514	divergent	upstream	Group 3	antisense intergenic

KN361_nc526	KN361_3g008653	antisense	intergenic	6293	divergent	upstream	Group 3	antisense intergenic
KN361_nc091	KN361_1g011271	antisense	intergenic	1349	divergent	upstream	Group 3	antisense intergenic
KN361_nc418	KN361_2g005059	antisense	intergenic	16042	convergent	downstream	Group 3	antisense intergenic
KN361_nc268	KN361_1g001840	antisense	intergenic	1764	divergent	upstream	Group 3	antisense intergenic
KN361_nc115	KN361_1g009509	antisense	intergenic	12885	convergent	downstream	Group 3	antisense intergenic
KN361_nc560	KN361_3g006781	antisense	intergenic	47671	divergent	upstream	Group 3	antisense intergenic
KN361_nc340	KN361_2g003945	antisense	intergenic	18193	divergent	upstream	Group 3	antisense intergenic
KN361_nc428	KN361_2g005801	antisense	intergenic	37770	divergent	upstream	Group 3	antisense intergenic
KN361_nc370	KN361_2g005418	antisense	intergenic	4557	divergent	upstream	Group 3	antisense intergenic
KN361_nc421	KN361_2g007497	antisense	intergenic	3231	convergent	downstream	Group 3	antisense intergenic
KN361_nc284	KN361_0g000284	antisense	intergenic	262	convergent	downstream	Group 3	antisense intergenic
KN361_nc216	KN361_1g000191	antisense	intergenic	2689	convergent	downstream	Group 3	antisense intergenic
KN361_nc110	KN361_1g009221	antisense	intergenic	22239	divergent	upstream	Group 3	antisense intergenic
KN361_nc540	KN361_3g001121	antisense	intergenic	2176	convergent	downstream	Group 3	antisense intergenic
KN361_nc471	KN361_3g006042	antisense	intergenic	22534	convergent	downstream	Group 3	antisense intergenic
KN361_nc487	KN361_3g007688	antisense	intergenic	27456	divergent	upstream	Group 3	antisense intergenic
KN361_nc464	KN361_3g007217	antisense	intergenic	511	divergent	upstream	Group 3	antisense intergenic
KN361_nc464	KN361_3g007217	antisense	intergenic	525	divergent	upstream	Group 3	antisense intergenic
KN361_nc306	KN361_2g004940	antisense	intergenic	20687	divergent	upstream	Group 3	antisense intergenic
KN361_nc659	KN361_4g006273	antisense	intergenic	105	convergent	downstream	Group 3	antisense intergenic
KN361_nc246	KN361_1g010227	antisense	intergenic	4668	convergent	downstream	Group 3	antisense intergenic
KN361_nc364	KN361_2g009078	antisense	intergenic	2499	convergent	downstream	Group 3	antisense intergenic
KN361_nc143	KN361_1g010323	antisense	intergenic	3377	divergent	upstream	Group 3	antisense intergenic
KN361_nc464	KN361_3g007217	antisense	intergenic	319	divergent	upstream	Group 3	antisense intergenic
KN361_nc371	KN361_2g004365	antisense	intergenic	40044	convergent	downstream	Group 3	antisense intergenic
KN361_nc679	KN361_4g001353	antisense	intergenic	15724	convergent	downstream	Group 3	antisense intergenic
KN361_nc115	KN361_1g009509	antisense	intergenic	12881	convergent	downstream	Group 3	antisense intergenic

KN361_nc635	KN361_4g001988	antisense	intergenic	491	divergent	upstream	Group 3	antisense intergenic
KN361_nc286	KN361_2g003488	antisense	intergenic	190	convergent	downstream	Group 3	antisense intergenic
KN361_nc330	KN361_2g004599	antisense	intergenic	28883	divergent	upstream	Group 3	antisense intergenic
KN361_nc377	KN361_2g006452	antisense	intergenic	8325	divergent	upstream	Group 3	antisense intergenic
KN361_nc659	KN361_4g006273	antisense	intergenic	105	convergent	downstream	Group 3	antisense intergenic
KN361_nc295	KN361_2g004297	antisense	intergenic	2200	divergent	upstream	Group 3	antisense intergenic
KN361_nc681	KN361_4g002961	antisense	intergenic	55085	convergent	downstream	Group 3	antisense intergenic
KN361_nc679	KN361_4g001353	antisense	intergenic	15384	convergent	downstream	Group 3	antisense intergenic
KN361_nc330	KN361_2g004599	antisense	intergenic	28900	divergent	upstream	Group 3	antisense intergenic
KN361_nc011	KN361_1g007802	antisense	intergenic	251	convergent	downstream	Group 3	antisense intergenic
KN361_nc329	KN361_2g003362	antisense	intergenic	1103	divergent	upstream	Group 3	antisense intergenic
KN361_nc612	KN361_4g002509	antisense	intergenic	31936	divergent	upstream	Group 3	antisense intergenic
KN361_nc158	KN361_1g010511	antisense	intergenic	14112	divergent	upstream	Group 3	antisense intergenic
KN361_nc641	KN361_4g001576	antisense	intergenic	70878	convergent	downstream	Group 3	antisense intergenic
KN361_nc679	KN361_4g001353	antisense	intergenic	15126	convergent	downstream	Group 3	antisense intergenic
KN361_nc230	KN361_1g000284	antisense	intergenic	1048	convergent	downstream	Group 3	antisense intergenic
KN361_nc567	KN361_3g003160	antisense	intergenic	8029	convergent	downstream	Group 3	antisense intergenic
KN361_nc221	KN361_1g005205	antisense	intergenic	3911	divergent	upstream	Group 3	antisense intergenic
KN361_nc168	KN361_1g009693	antisense	intergenic	43863	convergent	downstream	Group 3	antisense intergenic
KN361_nc670	KN361_4g002601	antisense	intergenic	17901	convergent	downstream	Group 3	antisense intergenic
KN361_nc374	KN361_2g004326	antisense	intergenic	54467	divergent	upstream	Group 3	antisense intergenic
KN361_nc115	KN361_1g009509	antisense	intergenic	12885	convergent	downstream	Group 3	antisense intergenic
KN361_nc491	KN361_3g006545	antisense	intergenic	302	divergent	upstream	Group 3	antisense intergenic
KN361_nc278	KN361_1g005666	antisense	intergenic	3957	convergent	downstream	Group 3	antisense intergenic
KN361_nc511	KN361_3g009077	antisense	intergenic	7422	convergent	downstream	Group 3	antisense intergenic
KN361_nc311	KN361_2g006182	antisense	intergenic	4787	divergent	upstream	Group 3	antisense intergenic

KN361_nc654	KN361_4g000270	antisense	intergenic	7420	convergent	downstream	Group 3	antisense intergenic
KN361_nc216	KN361_1g000191	antisense	intergenic	2640	convergent	downstream	Group 3	antisense intergenic
KN361_nc661	KN361_4g002249	antisense	intergenic	33564	divergent	upstream	Group 3	antisense intergenic
KN361_nc635	KN361_4g001988	antisense	intergenic	459	divergent	upstream	Group 3	antisense intergenic
KN361_nc090	KN361_1g010971	antisense	intergenic	7024	divergent	upstream	Group 3	antisense intergenic
KN361_nc248	KN361_1g009098	antisense	intergenic	33063	convergent	downstream	Group 3	antisense intergenic
KN361_nc216	KN361_1g000191	antisense	intergenic	2640	convergent	downstream	Group 3	antisense intergenic
KN361_nc241	KN361_1g000001	antisense	intergenic	5074	divergent	upstream	Group 3	antisense intergenic
KN361_nc468	KN361_3g008955	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc702	KN361_4g002393	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc016	KN361_1g009527	sense	genic	0	nested	exonic	Group 4	sense genic exonic
KN361_nc621	KN361_4g002087	sense	genic	0	containing	exonic	Group 4	sense genic exonic
KN361_nc468	KN361_3g008955	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc342	KN361_2g004905	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc363	KN361_2g004624	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc097	KN361_1g000735	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc506	KN361_3g006044	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc473	KN361_3g007179	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc698	KN361_4g000825	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc468	KN361_3g008955	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc016	KN361_1g009527	sense	genic	0	nested	exonic	Group 4	sense genic exonic
KN361_nc016	KN361_1g009527	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc499	KN361_3g007665	sense	genic	0	nested	exonic	Group 4	sense genic exonic
KN361_nc016	KN361_1g009527	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc587	KN361_4g001988	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc468	KN361_3g008955	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc195	KN361_1g008565	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic

KN361_nc638	KN361_4g002905	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc749	KN361_0g000711	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc461	KN361_3g006152	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc097	KN361_1g000735	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc097	KN361_1g000735	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc749	KN361_0g000711	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc468	KN361_3g008955	sense	genic	0	overlapping	exonic	Group 4	sense genic exonic
KN361_nc505	KN361_3g006236	sense	genic	0	containing	intronic	Group 5	sense genic intronic
KN361_nc502	KN361_3g006570	sense	genic	0	containing	intronic	Group 5	sense genic intronic
KN361_nc529	KN361_3g006906	sense	genic	0	containing	intronic	Group 5	sense genic intronic
KN361_nc017	KN361_1g009686	sense	genic	0	containing	intronic	Group 5	sense genic intronic
KN361_nc075	KN361_1g009527	sense	genic	0	overlapping	intronic	Group 5	sense genic intronic
KN361_nc727	KN361_0g000543	sense	genic	0	nested	intronic	Group 5	sense genic intronic
KN361_nc621	KN361_4g002087	sense	genic	0	containing	intronic	Group 5	sense genic intronic
KN361_nc662	KN361_4g004769	sense	genic	0	nested	intronic	Group 5	sense genic intronic
KN361_nc296	KN361_2g004406	sense	genic	0	containing	intronic	Group 5	sense genic intronic
KN361_nc463	KN361_3g006621	sense	intergenic	79586	same_strand	upstream	Group 6	sense intergenic
KN361_nc346	KN361_2g004852	sense	intergenic	23731	same_strand	downstream	Group 6	sense intergenic
KN361_nc071	KN361_1g008728	sense	intergenic	3602	same_strand	upstream	Group 6	sense intergenic
KN361_nc296	KN361_2g004406	sense	intergenic	1429	same_strand	upstream	Group 6	sense intergenic
KN361_nc119	KN361_1g003628	sense	intergenic	3767	same_strand	upstream	Group 6	sense intergenic
KN361_nc657	KN361_4g003395	sense	intergenic	4408	same_strand	downstream	Group 6	sense intergenic
KN361_nc197	KN361_1g008738	sense	intergenic	15427	same_strand	upstream	Group 6	sense intergenic
KN361_nc232	KN361_1g007620	sense	intergenic	6609	same_strand	downstream	Group 6	sense intergenic
KN361_nc474	KN361_3g007521	sense	intergenic	47359	same_strand	upstream	Group 6	sense intergenic
KN361_nc071	KN361_1g008728	sense	intergenic	3663	same_strand	upstream	Group 6	sense intergenic

KN361_nc682	KN361_4g002106	sense	intergenic	48060	same_strand	downstream	Group 6	sense intergenic
KN361_nc692	KN361_4g002980	sense	intergenic	10340	same_strand	downstream	Group 6	sense intergenic
KN361_nc465	KN361_3g007331	sense	intergenic	15276	same_strand	downstream	Group 6	sense intergenic
KN361_nc463	KN361_3g006621	sense	intergenic	79542	same_strand	upstream	Group 6	sense intergenic
KN361_nc515	KN361_3g008661	sense	intergenic	10516	same_strand	downstream	Group 6	sense intergenic
KN361_nc242	KN361_1g008545	sense	intergenic	59248	same_strand	upstream	Group 6	sense intergenic
KN361_nc463	KN361_3g006621	sense	intergenic	79520	same_strand	upstream	Group 6	sense intergenic
KN361_nc347	KN361_2g004940	sense	intergenic	10453	same_strand	upstream	Group 6	sense intergenic
KN361_nc296	KN361_2g004406	sense	intergenic	1429	same_strand	upstream	Group 6	sense intergenic
KN361_nc296	KN361_2g004406	sense	intergenic	1809	same_strand	upstream	Group 6	sense intergenic
KN361_nc269	KN361_1g010724	sense	intergenic	3485	same_strand	upstream	Group 6	sense intergenic
KN361_nc463	KN361_3g006621	sense	intergenic	79695	same_strand	upstream	Group 6	sense intergenic
KN361_nc339	KN361_2g001963	sense	intergenic	2696	same_strand	downstream	Group 6	sense intergenic
KN361_nc347	KN361_2g004940	sense	intergenic	7911	same_strand	upstream	Group 6	sense intergenic
KN361_nc252	KN361_1g010724	sense	intergenic	2554	same_strand	upstream	Group 6	sense intergenic
KN361_nc355	KN361_2g005471	sense	intergenic	7668	same_strand	upstream	Group 6	sense intergenic
KN361_nc071	KN361_1g008728	sense	intergenic	3732	same_strand	upstream	Group 6	sense intergenic
KN361_nc462	KN361_3g006450	sense	intergenic	17159	same_strand	downstream	Group 6	sense intergenic
KN361_nc586	KN361_4g001697	sense	intergenic	3929	same_strand	downstream	Group 6	sense intergenic
KN361_nc296	KN361_2g004406	sense	intergenic	2017	same_strand	upstream	Group 6	sense intergenic
KN361_nc296	KN361_2g004406	sense	intergenic	1594	same_strand	upstream	Group 6	sense intergenic
KN361_nc613	KN361_4g002261	sense	intergenic	47737	same_strand	upstream	Group 6	sense intergenic
KN361_nc693	KN361_4g006952	sense	intergenic	2166	same_strand	downstream	Group 6	sense intergenic
KN361_nc474	KN361_3g007521	sense	intergenic	46525	same_strand	upstream	Group 6	sense intergenic
KN361_nc269	KN361_1g010724	sense	intergenic	8791	same_strand	upstream	Group 6	sense intergenic
KN361_nc140	KN361_1g008392	sense	intergenic	99412	same_strand	upstream	Group 6	sense intergenic
KN361_nc102	KN361_1g008941	sense	intergenic	16153	same_strand	downstream	Group 6	sense intergenic

KN361_nc541	KN361_3g008464	sense	intergenic	4080	same_strand	downstream	Group 6	sense intergenic
KN361_nc140	KN361_1g008392	sense	intergenic	99419	same_strand	upstream	Group 6	sense intergenic
KN361_nc317	KN361_2g005146	sense	intergenic	4239	same_strand	downstream	Group 6	sense intergenic
KN361_nc463	KN361_3g006621	sense	intergenic	79512	same_strand	upstream	Group 6	sense intergenic
KN361_nc463	KN361_3g006621	sense	intergenic	79244	same_strand	upstream	Group 6	sense intergenic
KN361_nc655	KN361_4g002509	sense	intergenic	41286	same_strand	upstream	Group 6	sense intergenic
KN361_nc463	KN361_3g006621	sense	intergenic	79244	same_strand	upstream	Group 6	sense intergenic
KN361_nc156	KN361_1g009088	sense	intergenic	2580	same_strand	downstream	Group 6	sense intergenic
KN361_nc321	KN361_2g005500	sense	intergenic	5802	same_strand	upstream	Group 6	sense intergenic
KN361_nc516	KN361_3g002014	sense	intergenic	1490	same_strand	downstream	Group 6	sense intergenic
KN361_nc468	KN361_3g008955	sense	intergenic	45	same_strand	downstream	Group 6	sense intergenic
KN361_nc541	KN361_3g008464	sense	intergenic	4913	same_strand	downstream	Group 6	sense intergenic
KN361_nc475	KN361_3g008784	sense	intergenic	4357	same_strand	downstream	Group 6	sense intergenic
KN361_nc634	KN361_4g002626	sense	intergenic	26741	same_strand	downstream	Group 6	sense intergenic
KN361_nc462	KN361_3g006450	sense	intergenic	17174	same_strand	downstream	Group 6	sense intergenic
KN361_nc510	KN361_3g006508	sense	intergenic	5584	same_strand	upstream	Group 6	sense intergenic
KN361_nc568	KN361_3g005761	sense	intergenic	5575	same_strand	downstream	Group 6	sense intergenic
KN361_nc133	KN361_1g010725	sense	intergenic	6752	same_strand	downstream	Group 6	sense intergenic
KN361_nc586	KN361_4g001697	sense	intergenic	3922	same_strand	downstream	Group 6	sense intergenic
KN361_nc226	KN361_1g010906	sense	intergenic	3626	same_strand	downstream	Group 6	sense intergenic
KN361_nc296	KN361_2g004406	sense	intergenic	1594	same_strand	upstream	Group 6	sense intergenic
KN361_nc288	KN361_2g004779	sense	intergenic	583	same_strand	upstream	Group 6	sense intergenic
KN361_nc117	KN361_1g004896	sense	intergenic	1734	same_strand	downstream	Group 6	sense intergenic
KN361_nc677	KN361_4g006501	sense	intergenic	5960	same_strand	upstream	Group 6	sense intergenic
KN361_nc346	KN361_2g004852	sense	intergenic	23730	same_strand	downstream	Group 6	sense intergenic
KN361_nc163	KN361_1g008069	sense	intergenic	262	same_strand	upstream	Group 6	sense intergenic

KN361_nc463	KN361_3g006621	sense	intergenic	79695	same_strand	upstream	Group 6	sense intergenic
KN361_nc719	KN361_4g004060	sense	intergenic	4734	same_strand	upstream	Group 6	sense intergenic
KN361_nc050	KN361_1g008736	sense	intergenic	20060	same_strand	downstream	Group 6	sense intergenic
KN361_nc537	KN361_3g006403	sense	intergenic	27370	same_strand	upstream	Group 6	sense intergenic
KN361_nc672	KN361_4g003579	sense	intergenic	7966	same_strand	downstream	Group 6	sense intergenic
KN361_nc051	KN361_1g009098	sense	intergenic	83519	same_strand	downstream	Group 6	sense intergenic
KN361_nc396	KN361_2g006765	sense	intergenic	3736	same_strand	downstream	Group 6	sense intergenic
KN361_nc532	KN361_3g005880	sense	intergenic	3587	same_strand	downstream	Group 6	sense intergenic
KN361_nc475	KN361_3g008784	sense	intergenic	4390	same_strand	downstream	Group 6	sense intergenic
KN361_nc501	KN361_3g006249	sense	intergenic	305	same_strand	upstream	Group 6	sense intergenic
KN361_nc369	KN361_2g004209	sense	intergenic	47574	same_strand	downstream	Group 6	sense intergenic
KN361_nc516	KN361_3g002014	sense	intergenic	1490	same_strand	downstream	Group 6	sense intergenic
KN361_nc117	KN361_1g004896	sense	intergenic	1749	same_strand	downstream	Group 6	sense intergenic
KN361_nc697	KN361_4g001495	sense	intergenic	2297	same_strand	upstream	Group 6	sense intergenic
KN361_nc655	KN361_4g002509	sense	intergenic	41260	same_strand	upstream	Group 6	sense intergenic
KN361_nc015	KN361_1g009219	sense	intergenic	55145	same_strand	downstream	Group 6	sense intergenic
KN361_nc693	KN361_4g006952	sense	intergenic	2170	same_strand	downstream	Group 6	sense intergenic
KN361_nc463	KN361_3g006621	sense	intergenic	79522	same_strand	upstream	Group 6	sense intergenic
KN361_nc510	KN361_3g006508	sense	intergenic	5023	same_strand	upstream	Group 6	sense intergenic
KN361_nc632	KN361_4g002842	sense	intergenic	26798	same_strand	upstream	Group 6	sense intergenic
KN361_nc454	KN361_3g005166	sense	intergenic	241	same_strand	downstream	Group 6	sense intergenic
KN361_nc523	KN361_3g007206	sense	intergenic	25801	same_strand	downstream	Group 6	sense intergenic
KN361_nc288	KN361_2g004779	sense	intergenic	485	same_strand	upstream	Group 6	sense intergenic
KN361_nc347	KN361_2g004940	sense	intergenic	10864	same_strand	upstream	Group 6	sense intergenic
KN361_nc705	KN361_4g001955	sense	intergenic	18559	same_strand	upstream	Group 6	sense intergenic
KN361_nc325	KN361_2g007191	sense	intergenic	3329	same_strand	upstream	Group 6	sense intergenic
KN361_nc475	KN361_3g008784	sense	intergenic	4360	same_strand	downstream	Group 6	sense intergenic

KN361_nc539	KN361_3g007299	sense	intergenic	1188	same_strand	downstream	Group 6	sense intergenic
KN361_nc218	KN361_1g004564	sense	intergenic	2218	same_strand	downstream	Group 6	sense intergenic
KN361_nc462	KN361_3g006450	sense	intergenic	17174	same_strand	downstream	Group 6	sense intergenic
KN361_nc147	KN361_1g007781	sense	intergenic	13328	same_strand	downstream	Group 6	sense intergenic
KN361_nc554	KN361_3g005908	sense	intergenic	1596	same_strand	upstream	Group 6	sense intergenic
KN361_nc251	KN361_1g006886	sense	intergenic	7513	same_strand	downstream	Group 6	sense intergenic
KN361_nc474	KN361_3g007521	sense	intergenic	46327	same_strand	upstream	Group 6	sense intergenic
KN361_nc133	KN361_1g010725	sense	intergenic	2199	same_strand	downstream	Group 6	sense intergenic
KN361_nc576	KN361_3g005908	sense	intergenic	4152	same_strand	upstream	Group 6	sense intergenic
KN361_nc347	KN361_2g004940	sense	intergenic	7798	same_strand	upstream	Group 6	sense intergenic
KN361_nc749	KN361_0g000711	sense	intergenic	105	same_strand	upstream	Group 6	sense intergenic
KN361_nc296	KN361_2g004406	sense	intergenic	1809	same_strand	upstream	Group 6	sense intergenic
KN361_nc321	KN361_2g005500	sense	intergenic	5809	same_strand	upstream	Group 6	sense intergenic
KN361_nc420	KN361_2g005417	sense	intergenic	27618	same_strand	upstream	Group 6	sense intergenic
KN361_nc288	KN361_2g004779	sense	intergenic	167	same_strand	upstream	Group 6	sense intergenic
KN361_nc302	KN361_2g004596	sense	intergenic	3737	same_strand	upstream	Group 6	sense intergenic
KN361_nc071	KN361_1g008728	sense	intergenic	3602	same_strand	upstream	Group 6	sense intergenic
KN361_nc296	KN361_2g004406	sense	intergenic	1809	same_strand	upstream	Group 6	sense intergenic
KN361_nc020	KN361_1g010507	sense	intergenic	298	same_strand	downstream	Group 6	sense intergenic
KN361_nc325	KN361_2g007191	sense	intergenic	3179	same_strand	upstream	Group 6	sense intergenic
KN361_nc749	KN361_0g000711	sense	intergenic	611	same_strand	upstream	Group 6	sense intergenic
KN361_nc261	KN361_1g010891	sense	intergenic	860	same_strand	downstream	Group 6	sense intergenic

Supplementary File 1: Table S7. GC content of each *P. ovata* CDS (45 out of 63228 transcripts were shown).

Name	Length	%GC	Description
KN361_1g000001A	1077	45.31	HiC_scaffold_1
KN361_1g000001B	1077	45.31	HiC_scaffold_1
KN361_1g000001C	1077	45.31	HiC_scaffold_1
KN361_1g000001E	1077	45.31	HiC_scaffold_1
KN361_1g000001F	1077	45.31	HiC_scaffold_1
KN361_1g000002C	708	40.68	HiC_scaffold_1
KN361_1g000004A	873	47.42	HiC_scaffold_1
KN361_1g000004D	549	47.72	HiC_scaffold_1
KN361_1g000004H	549	47.72	HiC_scaffold_1
KN361_1g000005A	1146	44.68	HiC_scaffold_1
KN361_1g000005C	525	40.19	HiC_scaffold_1
KN361_1g000005G	1644	42.7	HiC_scaffold_1
KN361_1g000006C	2907	43.52	HiC_scaffold_1
KN361_1g000006D	2907	43.52	HiC_scaffold_1
KN361_1g000006E	2907	43.52	HiC_scaffold_1
KN361_1g000006G	2907	43.52	HiC_scaffold_1
KN361_1g000006H	2907	43.52	HiC_scaffold_1
KN361_1g000007A	4890	44.44	HiC_scaffold_1
KN361_1g000007C	4725	44.15	HiC_scaffold_1
KN361_1g000007E	1080	41.48	HiC_scaffold_1
KN361_1g000008A	453	35.32	HiC_scaffold_1
KN361_1g000008B	453	35.32	HiC_scaffold_1
KN361_1g000008D	399	38.6	HiC_scaffold_1
KN361_1g000008E	453	35.32	HiC_scaffold_1
KN361_1g000011A	171	45.61	HiC_scaffold_1
KN361_1g000013A	1305	40.92	HiC_scaffold_1
KN361_1g000013B	1617	41.37	HiC_scaffold_1
KN361_1g000013C	594	40.4	HiC_scaffold_1
KN361_1g000014A	510	46.47	HiC_scaffold_1
KN361_1g000015E	255	48.24	HiC_scaffold_1
KN361_1g000017A	231	52.81	HiC_scaffold_1
KN361_1g000019F	324	40.12	HiC_scaffold_1
KN361_1g000019G	324	40.12	HiC_scaffold_1
KN361_1g000020E	1461	39.15	HiC_scaffold_1
KN361_1g000020H	1461	39.15	HiC_scaffold_1
KN361_1g000020I	1461	39.15	HiC_scaffold_1
KN361_1g000021B	1296	45.83	HiC_scaffold_1

Supplementary File 1: Table S8. Identified location of genes linked to histone modifications and DNA methylation. (Only data from Chr 1 (Chromosome 1) was presented)

Chr	Start	End	Gene_ID	Product
Chr 1	822117	826183	KN361_1g000105A	H3K9 histone-lysine N-methyltransferase and H3 lysine-9 specific SUVH1
Chr 1	3298277	3303405	KN361_1g000482A	putative DNA (cytosine-5)-methyltransferase CMT1
Chr 1	5067793	5071372	KN361_1g000766A	methyl-CpG-binding domain-containing protein 2
Chr 1	7008224	7020868	KN361_1g001095B	histone acetyltransferase HAC1
Chr 1	7453920	7464709	KN361_1g001186F	rRNA (cytosine-C(5))-methyltransferase NOP2C
Chr 1	10878110	10881241	KN361_1g001729A	histone-binding protein MSI1
Chr 1	11221783	11228319	KN361_1g001787A	methyl-CpG-binding domain-containing protein 8
Chr 1	12216413	12226600	KN361_1g001951B	H3K9 histone-lysine N-methyltransferase SUVR4
Chr 1	13133674	13136254	KN361_1g002097A	histone deacetylase complex subunit SAP18
Chr 1	13315953	13320128	KN361_1g002123A	histone-lysine N-methyltransferase CLF
Chr 1	13366086	13376103	KN361_1g002135A	lysine-specific histone demethylase 1 3
Chr 1	14046700	14052422	KN361_1g002247C	histone-binding protein MSI1
Chr 1	14506731	14507430	KN361_1g002330B	histone H4
Chr 1	17781130	17786255	KN361_1g002800D	COMPASS-like H3K4 histone methylase component WDR5B
Chr 1	18244989	18246248	KN361_1g002870B	histone H4
Chr 1	18386812	18387774	KN361_1g002890A	putative histone H2B.1
Chr 1	18580463	18584974	KN361_1g002924B	COMPASS-like H3K4 histone methylase component WDR5B
Chr 1	19521412	19529381	KN361_1g003065A	methyl-CpG-binding domain-containing protein 2
Chr 1	23276618	23292485	KN361_1g003601C	histone acetyltransferase subunit 3
Chr 1	24781353	24787083	KN361_1g003828B	COMPASS-like H3K4 histone methylase component WDR5A
Chr 1	25642363	25646832	KN361_1g003968A	histone-lysine N-methyltransferase ATXR3
Chr 1	25645386	25656627	KN361_1g003969A	histone-lysine N-methyltransferase ATXR3
Chr 1	26504729	26512310	KN361_1g004097A	COMPASS-like H3K4 histone methylase component WDR5B
Chr 1	26781108	26794213	KN361_1g004131C	methyl-CpG-binding domain-containing protein 9
Chr 1	27924070	27935319	KN361_1g004283B	histone-lysine N-methyltransferase ATX4
Chr 1	28649224	28652084	KN361_1g004382E	histone deacetylase HDT1
Chr 1	30584875	30587942	KN361_1g004682A	Sin3 binding region of histone deacetylase complex subunit SAP30
Chr 1	32800486	32808913	KN361_1g004987C	DNA (cytosine-5)-methyltransferase 1B
Chr 1	36588995	36595318	KN361_1g005539F	COMPASS-like H3K4 histone methylase component WDR5A
Chr 1	39038972	39040810	KN361_1g005908A	histone H4
Chr 1	42625650	42626611	KN361_1g006363A	histone H3.2

Chr 1	43422119	43426590	KN361_1g006467A	methyl-CpG-binding domain-containing protein 9
Chr 1	43923723	43924451	KN361_1g006527A	histone H2A.1
Chr 1	44438607	44441966	KN361_1g006609A	lysine-specific histone demethylase 1 3
Chr 1	44728710	44731732	KN361_1g006653B	single myb histone 4
Chr 1	45249035	45256123	KN361_1g006749C	COMPASS-like H3K4 histone methylase component WDR5A
Chr 1	47807382	47808111	KN361_1g007095A	histone H3.2
Chr 1	48283236	48285145	KN361_1g007146A	histone H1.2
Chr 1	49284743	49287157	KN361_1g007260A	COMPASS-like H3K4 histone methylase component WDR5B
Chr 1	49286170	49287981	KN361_1g007261A	COMPASS-like H3K4 histone methylase component WDR5B
Chr 1	49851257	49853514	KN361_1g007322A	methyl-CpG-binding domain-containing protein 4
Chr 1	50119917	50120915	KN361_1g007345B	putative histone H2AXb
Chr 1	52525047	52525715	KN361_1g007600A	histone H4
Chr 1	53314956	53320795	KN361_1g007694A	25S rRNA (cytosine-C(5))-methyltransferase NSUN5
Chr 1	53468769	53469862	KN361_1g007710A	C-5 cytosine-specific DNA methylase
Chr 1	53578700	53580039	KN361_1g007727A	C-5 cytosine-specific DNA methylase
Chr 1	53599399	53604861	KN361_1g007728D	rRNA (cytosine-C(5))-methyltransferase NOP2C
Chr 1	55075065	55079436	KN361_1g007852A	COMPASS-like H3K4 histone methylase component WDR5A
Chr 1	55087141	55090298	KN361_1g007854A	histone-lysine N-methyltransferaseH3 lysine-9 specific SUVH5
Chr 1	55499608	55503569	KN361_1g007887C	histone acetyltransferase GCN5
Chr 1	57086723	57094845	KN361_1g008047A	COMPASS-like H3K4 histone methylase component WDR5B
Chr 1	58350266	58357014	KN361_1g008109B	COMPASS-like H3K4 histone methylase component WDR5B
Chr 1	68062142	68063365	KN361_1g008458A	histone-lysine N-methyltransferase TRX1
Chr 1	68062502	68064288	KN361_1g008459A	histone H3-lysine
Chr 1	82973709	82979656	KN361_1g008898B	histone-lysine N-methyltransferase ATXR2
Chr 1	96941714	96950994	KN361_1g009258A	methyl-CpG-binding domain-containing protein 13
Chr 1	105366520	105376675	KN361_1g009544A	COMPASS-like H3K4 histone methylase component WDR5A
Chr 1	114609865	114613626	KN361_1g009742A	methyl-CpG-binding domain-containing protein 4-like protein
Chr 1	117060403	117060998	KN361_1g009825A	histone H4
Chr 1	121942616	121944100	KN361_1g010011A	histone-lysine N-methyltransferase ATXR6
Chr 1	121987925	121988681	KN361_1g010013A	histone H3.3
Chr 1	122125035	122125824	KN361_1g010015A	histone H3.3
Chr 1	125054023	125056862	KN361_1g010141F	methyl-CpG-binding domain-containing protein 9
Chr 1	126958273	126966092	KN361_1g010253A	histone acetyltransferase MCC1
Chr 1	127220390	127230158	KN361_1g010277A	histone-lysine N-methyltransferase ATXR6
Chr 1	130337387	130338127	KN361_1g010619A	histone H4
Chr 1	130388551	130389210	KN361_1g010624A	histone H4

Chr 1	130900530	130902937	KN361_1g010676D	putative inactive histone-lysine N-methyltransferase SUVR1
Chr 1	131978773	131984350	KN361_1g010805C	histone-lysine N-methyltransferase ATXR6
Chr 1	134206385	134217453	KN361_1g011126F	histone-lysine N-methyltransferase ATXR5
Chr 1	137323636	137326530	KN361_1g011540A	histone-lysine N-methyltransferase SUVR4
Chr 1	137324933	137326660	KN361_1g011541A	histone-lysine N-methyltransferase SUVR4
Chr 1	137679295	137682135	KN361_1g011597E	histone-lysine N-methyltransferase ASHH2

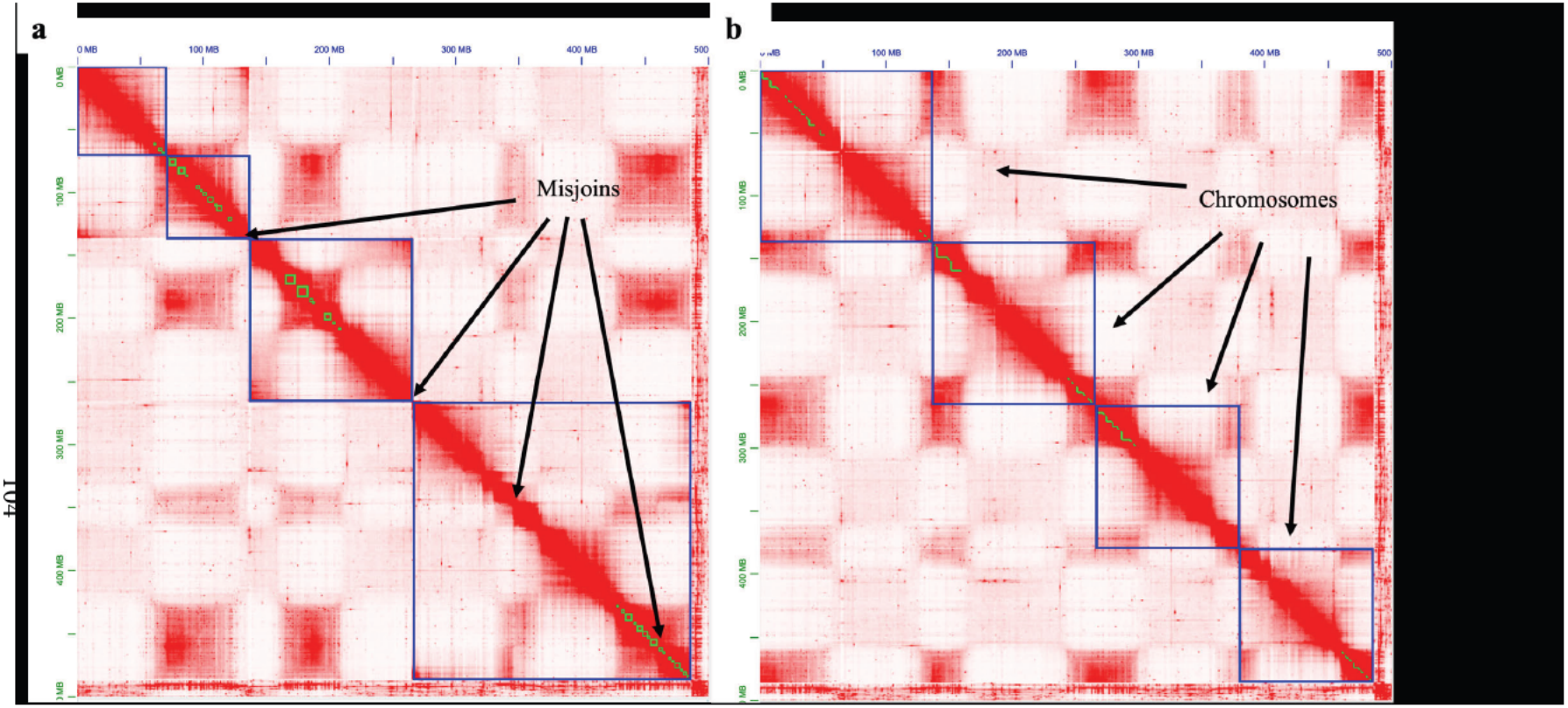
Supplementary File 1: Table S9: Software used.

Tools	References	Link
conda v4.6.11	(Conda, RRID:SCR_018317)	https://github.com/conda/conda
python v3.8.6	(Python Programming Language, RRID:SCR_008394)	https://www.python.org/downloads/release/python-386/
bamtofastx v1.3.0	(PacBio Sequel System, RRID:SCR_017989)	https://github.com/PacificBiosciences/bam2fastx
minimap2 v2.17	(Minimap2, RRID:SCR_018550)	https://github.com/lh3/minimap2
samtools v1.9	(SAMTOOLS, RRID:SCR_002105)	http://www.htslib.org/doc/1.9/samtools.html
fastqc v0.11.9	(FastQC, RRID:SCR_014583)	https://github.com/s-andrews/FastQC; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
canu v2.1.1	(Canu, RRID:SCR_015880)	https://canu.readthedocs.io/en/latest/pipeline.html/
pbgcpp v1.0.0	(PacBio Sequel System, RRID:SCR_017989)	https://github.com/PacificBiosciences/gcpp
pbcore v2.1.2 (pbcoretools v0.8.1)	(PacBio Sequel System, RRID:SCR_017989)	https://github.com/PacificBiosciences/pbcoretools
pbbam v1.6.0	(PacBio Sequel System, RRID:SCR_017989)	https://github.com/PacificBiosciences/pbbam
seqtk v1.3	(Seqtk, RRID:SCR_018927)	https://github.com/lh3/seqtk
purge_haplotigs v1.1.1	(Purge_haplotigs, RRID:SCR_017616)	https://bitbucket.org/mroachawri/purge_haplotigs/src/master/

trimmomatic v0.39	(Trimmomatic, RRID:SCR_011848)	http://www.usadellab.org/cms/?page=trimmomatic
Phase Genomics		https://github.com/phasegenomics/hic_qc
bwa v0.7.17	(BWA, RRID:SCR_010910)	https://github.com/lh3/bwa; http://bio-bwa.sourceforge.net/bwa.shtml
samblaster v0.1.26	(SAMBLASTER, RRID:SCR_000468)	https://github.com/GregoryFaust/samblaster
matlock v20181227		https://github.com/phasegenomics/matlock
salsa2	Ghurye et al.[17]	https://github.com/marbl/SALSA
3d-dna	Dudchenko et al.[18]	https://github.com/aidenlab/3d-dna
Juicer	(Juicer, RRID:SCR_017226)	https://github.com/aidenlab/juicer
Juicebox (JBAT)	Dudchenko et al.[18]	https://github.com/aidenlab/Juicebox
genomescope2 v2.0	(GenomeScope, RRID:SCR_017014)	https://github.com/schatzlab/genomescope
jellyfish v2.2.10	(Jellyfish, RRID:SCR_005491)	https://github.com/gmarcais/Jellyfish
busco v4.1.4	(BUSCO, RRID:SCR_015008)	https://busco.ezlab.org/
LAI and LTR_retriever	Ou et al.[20]	https://github.com/oushujun/LTR_retriever
RepeatModeler v2.6.1	(RepeatModeler, RRID:SCR_015027)	http://www.repeatmasker.org/RepeatModeler/
RepeatMasker v4.1.1	(RepeatMasker, RRID:SCR_012954)	http://www.repeatmasker.org/RepeatMasker/

n50 v1.3.0	Telatin et al.[40]	https://github.com/quadram-institute-bioscience/seqfu/tree/master/n50
transposonPSI.pl		http://transposonpsi.sourceforge.net/
GAAS	Daniat et al.[42]	https://github.com/NBISweden/GAAS/tree/master/bin
ProtExcluder.pl	Campbell et al.[43]	http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction-Advanced
multiqc v1.8	(MultiQC, RRID:SCR_014982)	https://github.com/ewels/MultiQC/releases; https://multiqc.info/docs/
bbmap v38.87 or v38.79	(Bbmap, RRID:SCR_016965)	https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/
star v2.7.6a	(STAR, RRID:SCR_015899)	https://github.com/alexdobin/STAR
cufflinks v2.2.1	(Cufflinks, RRID:SCR_014597)	http://cole-trapnell-lab.github.io/cufflinks/manual/
gffcompare	Pertea and Pertea[44]	https://ccb.jhu.edu/software/stringtie/gffcompare.shtml
snakemake v5.26.1	(Snakemake, RRID:SCR_003475)	https://snakemake.readthedocs.io/en/stable/
transdecoder v5.5.0	(TransDecoder, RRID:SCR_017647)	https://github.com/TransDecoder/TransDecoder/wiki
trinotate v3.2.1	(Trinotate, RRID:SCR_018930)	https://github.com/Trinotate/Trinotate.github.io/wiki
BLASTP	(BLASTP, RRID:SCR_001010)	
BLASTX	(BLASTX, RRID:SCR_001653)	
SRA-Toolkit/2.10.8	(NCBI Sequence Read Archive (SRA), RRID:SCR_004891)	https://github.com/ncbi/sra-tools

bedtools v2.29.2	(BEDTools, RRID:SCR_006646)	https://bedtools.readthedocs.io/en/latest/content/installation.html
AGAT	Dainat et al.[45]	https://github.com/NBISweden/AGAT
RNAMMER	(RNAmmer, RRID:SCR_017075)	
tRNAscan-SE v2.0.7	(tRNAscan-SE, RRID:SCR_010835)	https://github.com/UCSC-LoweLab/tRNAscan-SE
FEELnc	Wucher et al.[49]	https://github.com/tderrien/FEELnc
GenomeTools v1.2.1	(GenomeTools, RRID:SCR_016120)	http://genometools.org/
tbl2asn	(tbl2asn, RRID:SCR_016636)	https://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/
EMBOSS	(EMBOSS, RRID:SCR_008493)	
GMAP v2021.08.25	(GMAP, RRID:SCR_008992)	
MUSCLE v3.8.1551	(MUSCLE, RRID:SCR_011812)	
BMGE v1.12	Criscuolo and Gribaldo[50]	
FastTree v2.1.10	(FastTree, RRID:SCR_015501)	http://www.microbesonline.org/fasttree/
FigTree v1.4.4	(FigTree, RRID:SCR_008515)	http://tree.bio.ed.ac.uk/software/figtree/
KaryoploteR	Gel and Serra[51]	https://github.com/bernatgel/karyoploteR



Supplementary file 2: Hi-C interaction heat map generated by 3D-DNA pipeline, visualised, and corrected by JBAT (Juicebox Assembly Tools). **a**. Before curation, we found misjoins in the assembly. **b.** After curation, we obtained four superscaffolds or chromosomes. Green boxes represent scaffolds while blue boxes refer to chromosomes.

Supplementary file 3. Long Terminal Repeat Assembly Index (LAI) values for the whole genome and each window size 3 Mb and sliding step 300 Kb. (A screenshot part of the file)

Chr	From	To	Intact	Total	raw_LAI	LAI		
whole_genome		1	500939359		0.0838	0.5273	15.90	10.27
1	1	3000000	0.0031	0.0703	4.40	0		
1	300001	3300000	0.0020	0.0629	3.14	0		
1	600001	3600000	0.0036	0.0642	5.64	0.01		
1	900001	3900000	0.0036	0.0651	5.56	0		
1	1200001	4200000	0.0036	0.0615	5.89	0.26		
1	1500001	4500000	0.0056	0.0636	8.73	3.10		
1	1800001	4800000	0.0072	0.0645	11.22	5.59		
1	2100001	5100000	0.0072	0.0666	10.88	5.25		
1	2400001	5400000	0.0053	0.0649	8.12	2.49		
1	2700001	5700000	0.0073	0.0690	10.65	5.02		
1	3000001	6000000	0.0073	0.0638	11.52	5.89		
1	3300001	6300000	0.0073	0.0653	11.26	5.63		
1	3600001	6600000	0.0057	0.0648	8.80	3.17		
1	3900001	6900000	0.0070	0.0645	10.83	5.20		
1	4200001	7200000	0.0070	0.0651	10.73	5.10		
1	4500001	7500000	0.0050	0.0631	8.01	2.38		
1	4800001	7800000	0.0034	0.0646	5.21	0		
1	5100001	8100000	0.0034	0.0658	5.11	0		
1	5400001	8400000	0.0034	0.0667	5.04	0		
1	5700001	8700000	0.0013	0.0665	1.93	0		
1	6000001	9000000	0.0013	0.0691	1.85	0		
1	6300001	9300000	0.0029	0.0731	3.90	0		
1	6600001	9600000	0.0048	0.0778	6.17	0.54		
1	6900001	9900000	0.0035	0.0796	4.42	0		
1	7200001	10200000		0.0035	0.0799	4.40	0	
1	7500001	10500000		0.0035	0.0820	4.29	0	
1	7800001	10800000		0.0035	0.0829	4.24	0	
1	8100001	11100000		0.0043	0.0821	5.28	0	
1	8400001	11400000		0.0067	0.0838	8.01	2.38	
1	8700001	11700000		0.0077	0.0828	9.31	3.68	
1	9000001	12000000		0.0077	0.0824	9.36	3.73	
1	9300001	12300000		0.0061	0.0755	8.13	2.50	
1	9600001	12600000		0.0042	0.0713	5.88	0.25	

Supplementary file 4. Variant telomeric repeats in the *P. ovata* genome.

>Chromosome 1 [1-377] 377bp

5’AACCCCAAACCCTGAACCCGGAACCCTGAACCCTGAACCCTAAACCCTCAACCCCTCAACCCGG
AACCCTCAACCCCTCACCCGGAACCCCTCAACCCGAACCGGAACCCGGAACCCGGAACCCGGAAC
CGGAACCCGGAACCTGAACCCTGAACCCTAAACCCTCAACCCCTCAACCGGAACCCTCAACCCGA
ACCCGGAACCCGGAACCCCTCAACCCGAACCCGGAACCCGGAACCCGGAACACGGAACCCGGAAC
CCGGAACCCCGAACCCCGAAACCCCGAAACCGAAAACCCCGAACACCTGAACCCCGAAACCCCGA
AACCCGACCCCGAACCCGAACCCCGAAACCCTAAACCCTAAACCCTAAACCCT3’

>Chromosome 1 [137725100-137725283] 184bp

5’TTAGGGTTTGGGGTTGGGGTTGGGGGTTGGGGTTTTGGTTTGGGGTTTTAGGGTTTAGGGTTTAG
GGGTTAGGGTTTAGGGTTTAGGGTTTGGGGTTTGGGGTTAGGGTTTAGGGTTTTAGGGTTTAGGGTT
AGGGTTGGGGTTTGGGGTTTGGGGTTTGGGGTTTTGGGGTTGGGGTTTTGGG3’

>Chromosome 2 [4-362] 359bp

5’AAACCCTAAACCCTAAACCCTATAACCCTAAACCCTAAACCCTAAACCCAAAACCCTATAACCCA
AAACCCTATAACCCAAAACCCTATAACCCAAAACCCTATAACCCAAAACCCCAAACCCAAAAACC
CTATAACCCTAAACCCTATAACCCCAAACCCTATAACCCCAAACCCTATAACCCAAAACCCTATAA
CCCAAACCCTATAACCCAAAACCCTAAACCCTAAACCCTAAACCCCAAACCCAAAACCCTATAACC
CCAAACCCTAAAACCCCAAACCCTATAACCCCAAACCCTATAACCCAAAACCCTATAACCCAAACC
CTATAACCCTACACCCCAAACCCTAAACCCT3’

>Chromosome 2 [128866084-128866842] 759bp

5’TTTAGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTA
GGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGT
TTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAG
GGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTT
TAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGGTTTCGGTTTCGGGTTTAGGTTTCGGG
TTTAGGGTTTAGGGTTCGGGTTCGGGTTTAGGGTTTCGGGTTAGGGTTTAGGTTAGGGTTTAGGGTA

GGGTTAGGGGTTCGGGTTTAGGTTTAGGGTTGGGTTTTCGGTTGGGGTTTAGGGTTTAGGGTTTTAG
GTTTAGGGTTGGGGTTTAGGTTTAGGGTTTTCGGTTTAGGGTTTAGGGTTTAGGGTTTAGGTTTTAG
GGTTTTCGGTTTAGGGTTTAGGGTTTCGGGTTTATGGTTTAGGGTTTCGGGTTTCGGTTTCGGGTTTC
GGGTTTCGGGTTTTCGGGTTTCGGGTTAGGGTTTACGGGTTTCGGGTTAGGGTTTCGGGTTATAGGT
TTCGGGTTTCGGGTTTCGGTTCGGTTTAGGGTTTCCGGGTTTCGGTTTTCGGGTTTCGGGTTTGGGTT
TCGGGTTTCGGGGTTTCGGGG3’

>Chromosome 2 [128867197-128867510] 313bp

5’AAACCCTACACCCTAAACCCACTAAACCTAAACCCTAAACCCAAACCCTAACCCCTAAACCCTAA
ACCCGAACCCTAAACCCTAAACCTAAACCCTAACCCCTCCCCCCTAACCCCTACCCTAAACCCTAA
ACCCTAAACCCTCAACCCTAAACCCTAAACCCTAAACGCCTAAACCCCTAAACCCTAAACCCTAAC
CCTAAACCCTAAACCCTAAACCCTAACCCCTAACCCCTAAACCCTAAAAACCCTAAACCCTCAAAC
CCTAAACCCTAAACCCTAAACCCTCAAACCCTAAACCCGAAACCCTAACCC3’

>Chromosome 3 [114444289-114444890] 602bp

5’TTTAGGGTTTCGGGGTTTGGGGTTCAGGGGTTTGGGGTTTTCGGGTTTTCGGGTTCGGGGTTGAG
GGTTTAGGGTTTAGGGGTTCTGGGTTTGGGGTTCAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGT
TTAGGGTTCAGGGTTCAGGGTTCAGGGTTCAGGGTTCAGGGTTCAGGTTCAGGGTTCAGGGTTCAG
GGTTCAGGGTTGAGGGTTGAGGGTTCAGGGTTCAGGGTTCAGGGTTCAGGGTTCAGGGTTCAGGTT
CAGGGTTTAGGGTTTAGGGTTTAGGGTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGG
GTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTT
AGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGGTTAGGGTTTAGGGTTTAGGGTTTAGGGTTTAGGG
TTTAGGGTTTAGGGTTTAGGGTTAAGGGTTTAGGTTAGGGTTTAGGGTTTGGGTTTAGGGTTTTAG
GGGTTAGGGTTTAGGGTTTAGGGTTTAGGTTAGGGTTTAGGGTTTAGGGTTTAGGTTTAGGGTTTAG
GGTTT3’

>Chromosome 4 [4-518] 515bp

5’AAACCCTAAACCCTAAACCCTAAACCCTAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAA
CCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCC

TAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAA

CCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCC

TAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAA

ACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACC

CTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTAAACCCTA

AACCCTAAACCCTAAACCCTAAACCCTAACCCTAAACCCTAAACCCTAAACCCT3’

>Chromosome 4 [106345914-106,346,329] 416bp

5’TTTAGGGTTTGGGGTTTTAGGGTTTTGGGGTTTTGGGGTTTTGGGGTTTTGGGTTTAGGGTTTGGG

TTTTGGGGTTTTGGGTTTCGGGGTTTTAGGGTTTCGGGGTTTTAGGGTTTCGGGGTTTGGGTTTGGG

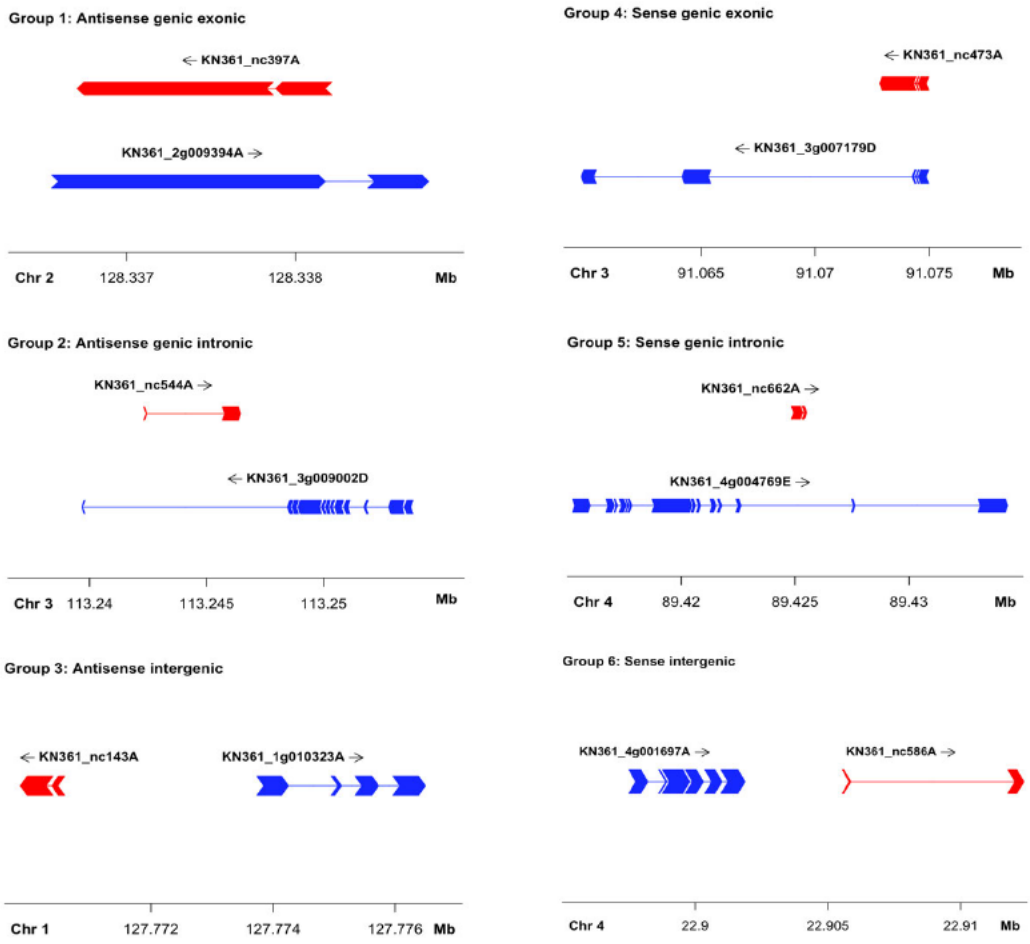
GTTTCGGGTTTGGGGTTTCGGGGTTTGGGTTTGGGTTTTGGGGTTCTGGGTTTCGGGGTTTTAGGGT

TTCGGGGTTTTAGGGTTTCGGGGTTTGGGTTTGGGTTTGGGGTTTAGGGTTTGGGGTTTCGGGGTTT

GGGTTTGGGGTTTAGGGTTTTGGGGTTCTGGGTTTCGGGGTTTGGGTTTAGGGTTTGGGGTTTAGGG

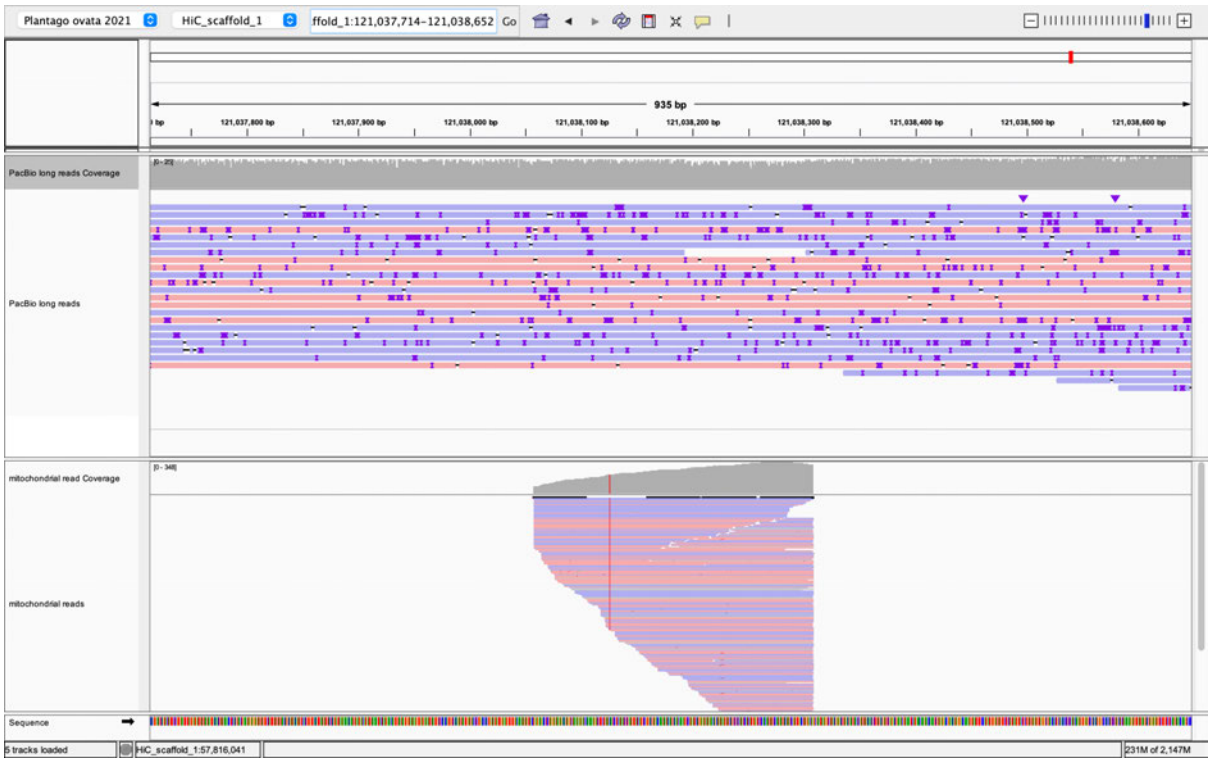
TTTAGGGTTTAGGGTTTAGGGTTTTAGGGTTTTAGGGTTTTAGGGTTTGAGGGTTTTAGGGTTTTAG

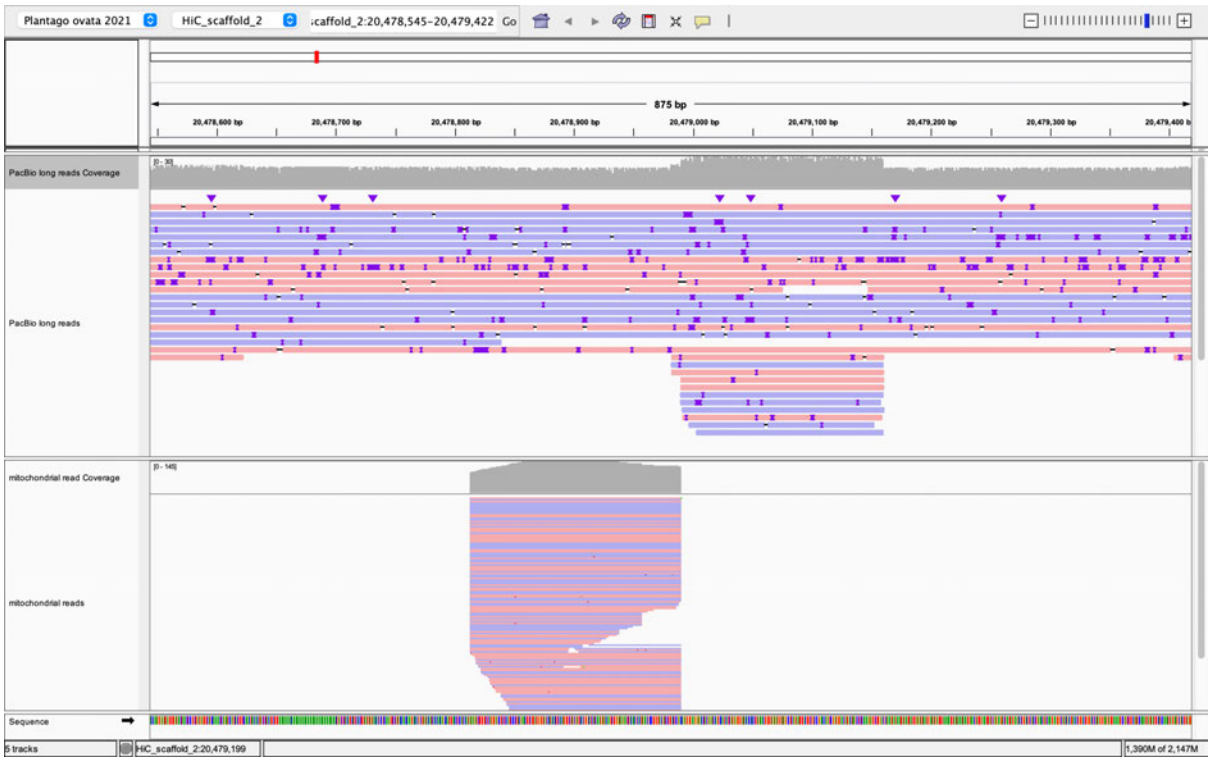
GGTTTAGGGTTTTGG3’



Supplementary file 5. Representatives from the six groups of lncRNA/mRNA pairs.

Supplementary file 6: Three locations of nuclear mitochondrial DNA (NUMT).



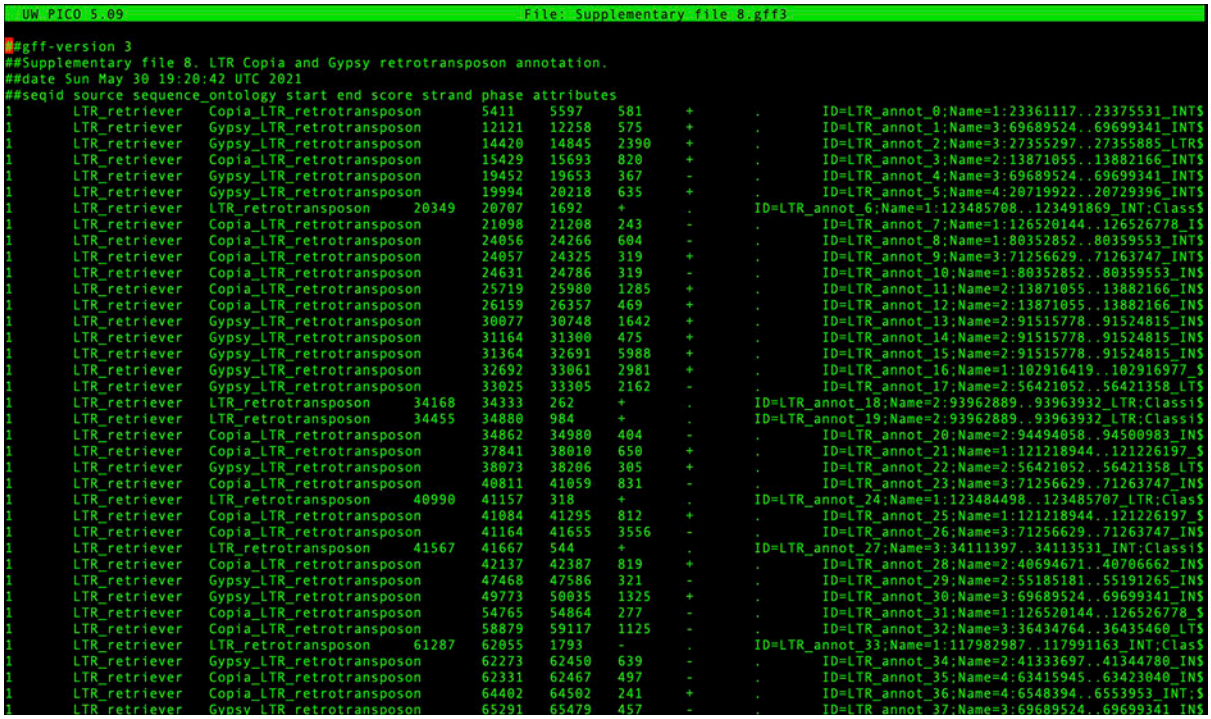

Plantago ovata 2021
HiC_scaffold_2
caffold_2:20,478,545-20,479,422
Go
875 bp
20,478,600 bp
20,478,700 bp
20,478,800 bp
20,478,900 bp
20,479,000 bp
20,479,100 bp
20,479,200 bp
20,479,300 bp
PacBio long reads Coverage
PacBio long reads
mitochondrial read Coverage
mitochondrial reads
Sequence
5 tracks
HiC_scaffold_2:20,479,199
1,390M of 2,147M

Supplementary file 7. GC content of *P. ovata* genome features (A screenshot).

Plantago_ovata_genome.fasta	
Analysis launched the 08/05/2021 at 18h39m55s	
Nb of sequences	876
Nb of sequences >1kb	858
Nb of sequences >10kb	610
Nb of nucleotides (counting Ns)	500939359
Nb of nucleotides U	0
Nb of sequences with U nucleotides	0
Nb of IUPAC nucleotides	0
Nb of sequences with IUPAC nucleotides	0
Nb of Ns	340400
Nb of internal N-regions (possibly links between contigs)	3404
Nb of long internal N-regions >10000 /!\ This is problematic for Genemark	0
Nb of pure (only) N sequences	0
Nb of sequences that begin or end with Ns	0
GC-content (%)	38.4
GC-content not counting Ns(%)	38.4
Nb of sequences with lowercase nucleotides	0
Nb of lowercase nucleotides	0
N50	128867510
L50	2
N90	106346329
L90	4

Note: GC contents on the gene, five prime UTR, CDS, exon, and three prime UTR were not shown.

Supplementary file 8. LTR Copia and Gypsy retrotransposon annotation (A screenshot)



```
 UW PICO 5.09                                File: Supplementary file 8.gff3

##gff-version 3
##Supplementary file 8. LTR Copia and Gypsy retrotransposon annotation.
##date Sun May 30 19:20:42 UTC 2021
##seqid source sequence_ontology start end score strand phase attributes
1       LTR_retriever   Copia_LTR_retrotransposon       5411    5597    581     +       .       ID=LTR_annot_0;Name=1:23361117..23375531_INT$
1       LTR_retriever   Gypsy_LTR_retrotransposon       12121   12258   575     +       .       ID=LTR_annot_1;Name=3:69689524..69699341_INT$
1       LTR_retriever   Gypsy_LTR_retrotransposon       14420   14845   2390    +       .       ID=LTR_annot_2;Name=3:27355297..27355885_LTR$
1       LTR_retriever   Copia_LTR_retrotransposon       15429   15693   820     +       .       ID=LTR_annot_3;Name=2:13871055..13882166_INT$
1       LTR_retriever   Gypsy_LTR_retrotransposon       19452   19653   367     -       .       ID=LTR_annot_4;Name=3:69689524..69699341_INT$
1       LTR_retriever   Gypsy_LTR_retrotransposon       19994   20218   635     +       .       ID=LTR_annot_5;Name=4:20719922..20729396_INT$
1       LTR_retriever   LTR_retrotransposon     20349   20707   1692    +       .       ID=LTR_annot_6;Name=1:123485708..123491869_INT;Class$
1       LTR_retriever   Copia_LTR_retrotransposon       21098   21208   243     -       .       ID=LTR_annot_7;Name=1:126520144..126526778_I$
1       LTR_retriever   Copia_LTR_retrotransposon       24056   24266   604     -       .       ID=LTR_annot_8;Name=1:80352852..80359553_INT$
1       LTR_retriever   Copia_LTR_retrotransposon       24057   24325   319     +       .       ID=LTR_annot_9;Name=3:71256629..71263747_INT$
1       LTR_retriever   Copia_LTR_retrotransposon       24631   24786   319     -       .       ID=LTR_annot_10;Name=1:80352852..80359553_IN$
1       LTR_retriever   Copia_LTR_retrotransposon       25719   25980   1285    +       .       ID=LTR_annot_11;Name=2:13871055..13882166_IN$
1       LTR_retriever   Copia_LTR_retrotransposon       26159   26357   469     +       .       ID=LTR_annot_12;Name=2:13871055..13882166_IN$
1       LTR_retriever   Gypsy_LTR_retrotransposon       30077   30748   1642    +       .       ID=LTR_annot_13;Name=2:91515778..91524815_IN$
1       LTR_retriever   Gypsy_LTR_retrotransposon       31164   31300   475     +       .       ID=LTR_annot_14;Name=2:91515778..91524815_IN$
1       LTR_retriever   Gypsy_LTR_retrotransposon       31364   32691   5988    +       .       ID=LTR_annot_15;Name=2:91515778..91524815_IN$
1       LTR_retriever   Gypsy_LTR_retrotransposon       32692   33061   2981    +       .       ID=LTR_annot_16;Name=1:102916419..102916977_$
1       LTR_retriever   Gypsy_LTR_retrotransposon       33025   33305   2162    -       .       ID=LTR_annot_17;Name=2:56421052..56421358_LT$
1       LTR_retriever   LTR_retrotransposon     34168   34333   262     +       .       ID=LTR_annot_18;Name=2:93962889..93963932_LTR;Classi$
1       LTR_retriever   LTR_retrotransposon     34455   34880   984     +       .       ID=LTR_annot_19;Name=2:93962889..93963932_LTR;Classi$
1       LTR_retriever   Copia_LTR_retrotransposon       34862   34980   404     -       .       ID=LTR_annot_20;Name=2:94494058..94500983_IN$
1       LTR_retriever   Copia_LTR_retrotransposon       37841   38010   650     +       .       ID=LTR_annot_21;Name=1:121218944..121226197_$
1       LTR_retriever   Gypsy_LTR_retrotransposon       38073   38206   305     +       .       ID=LTR_annot_22;Name=2:56421052..56421358_LT$
1       LTR_retriever   Copia_LTR_retrotransposon       40811   41059   831     -       .       ID=LTR_annot_23;Name=3:71256629..71263747_IN$
1       LTR_retriever   LTR_retrotransposon     40990   41157   318     +       .       ID=LTR_annot_24;Name=1:123484498..123485707_LTR;Clas$
1       LTR_retriever   Copia_LTR_retrotransposon       41084   41295   812     +       .       ID=LTR_annot_25;Name=1:121218944..121226197_$
1       LTR_retriever   Copia_LTR_retrotransposon       41164   41655   3556    -       .       ID=LTR_annot_26;Name=3:71256629..71263747_IN$
1       LTR_retriever   LTR_retrotransposon     41567   41667   544     +       .       ID=LTR_annot_27;Name=3:34111397..34113531_INT;Classi$
1       LTR_retriever   Copia_LTR_retrotransposon       42137   42387   819     +       .       ID=LTR_annot_28;Name=2:40694671..40706662_IN$
1       LTR_retriever   Gypsy_LTR_retrotransposon       47468   47586   321     -       .       ID=LTR_annot_29;Name=2:55185181..55191265_IN$
1       LTR_retriever   Gypsy_LTR_retrotransposon       49773   50035   1325    +       .       ID=LTR_annot_30;Name=3:69689524..69699341_IN$
1       LTR_retriever   Copia_LTR_retrotransposon       54765   54864   277     -       .       ID=LTR_annot_31;Name=1:126520144..126526778_$
1       LTR_retriever   Copia_LTR_retrotransposon       58879   59117   1125    -       .       ID=LTR_annot_32;Name=3:36434764..36435460_LT$
1       LTR_retriever   LTR_retrotransposon     61287   62055   1793    -       .       ID=LTR_annot_33;Name=1:117982987..117991163_INT;Clas$
1       LTR_retriever   Gypsy_LTR_retrotransposon       62273   62450   639     -       .       ID=LTR_annot_34;Name=2:41333697..41344780_IN$
1       LTR_retriever   Copia_LTR_retrotransposon       62331   62467   497     -       .       ID=LTR_annot_35;Name=4:63415945..63423040_IN$
1       LTR_retriever   Copia_LTR_retrotransposon       64402   64502   241     +       .       ID=LTR_annot_36;Name=4:6548394..6553953_INT;$
1       LTR_retriever   Gypsy_LTR_retrotransposon       65291   65479   457     -       .       ID=LTR_annot_37;Name=3:69689524..69699341_IN$
```

Supplementary file 9. Hi-C library quality control report (A screenshot).

PHASE GENOMICS

Hi-C Library QC Report

Genome Scaffolding Sufficiency

Label	Library statistics	Expected values
Subjective Hi-C library judgment	**SUFFICIENT**	See Judgment
Same strand high-quality* (HQ) read pairs (RPs)	**25.57%**	> 1.5%
Informative RPs**	**50.93%**	> 5.0%

*High quality (HQ) read pairs have minimum mapping quality >= 20, maximum edit distance <= 5, and are not duplicates.
**Informative read pairs are read pairs which have MAPQ > 0, are not PCR duplicates, and map to different contigs or >10 Kbp apart.

Metrics Demonstrating Strong Proximity Signal

Label	Library statistics	Expected values
Fraction of HQ RPs >10KB apart (CTGs >10KB)*	**26.01%**	> 3.0%
Fraction of HQ RPs Intercontig on CTGs >10KB**	**33.51%**	> 2.5%
Clustering usable HQ reads per contig (CTGs >5KB)***	**87.09**	> 600.0

*The proportion of *read pairs that span at least 10kbp*, out of all read pairs that map (a) with high-quality, (b) to the same contig, (c) where that contig is at least 10kbp long.
**The proportion of *read pairs mapping to two different contigs each greater than 10kbp*, out of all read pairs that map with high-quality.
***The average number of usable high-quality read pairs per contig, for contigs greater than 5kbp. Read pairs are "usable" if they map (a) with high-quality, (b) to different contigs, (c) where each of those contigs are greater than 5kbp and (c) both mappings are high-quality.

See below for information on differences between Phase Genomics Hi-C libraries and traditional Hi-C libraries.

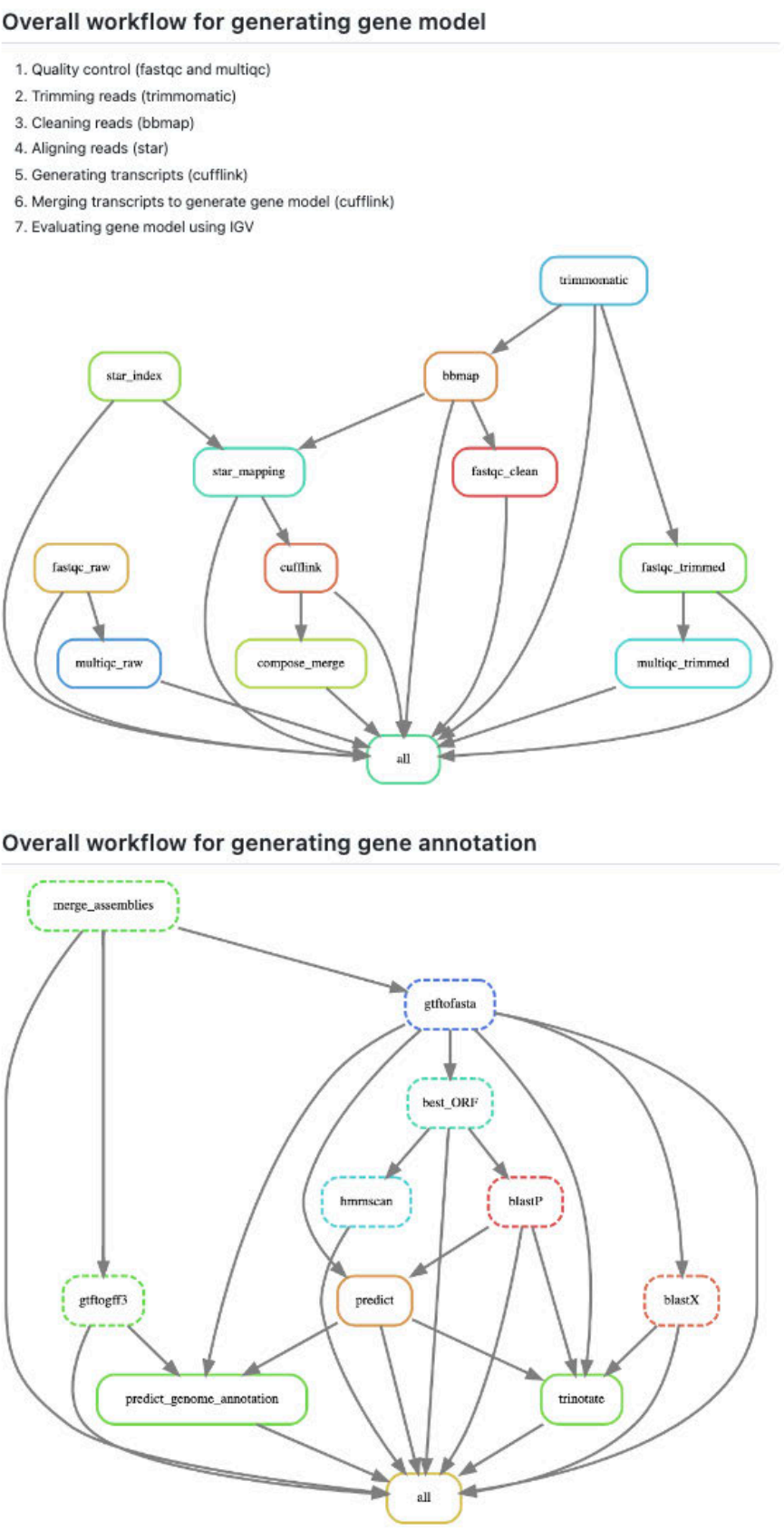
Noninformative Read Pair Breakdown

Label	Library statistics	Expected values
Noninformative RPs*	**3.73%**	<= 50.0%
Duplicate reads	**0.00%**	< 20.0%
Zero map distance read pairs	**3.73%**	<= 20.0%
Zero MAPQ reads	**0.00%**	<= 20.0%
Unmapped reads	**0.00%**	<= 10.0%

*Note that the sum of informative and noninformative read pairs is not 100% because read pairs with mapping distance between 1 and 10 Kbp are not classified as either informative or noninformative.
Because noninformative reads can belong to more than one category, these numbers may sum to a value larger than the overall noninformative read pair amount at the top of the report.

See below for information on differences between Phase Genomics Hi-C libraries and traditional Hi-C libraries.

Supplementary file 10. A screenshot of workflows from GitHub where the codes from this study are stored.



Chapter 4

Morphological and transcriptomic analysis of developing *Plantago ovata* seed capsules reveal structural features and key gene networks

ꕥ•ꕥ

Statement of Authorship

Title of Paper	Morphological and transcriptome analysis of *Plantago ovata* capsules
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Lina Herliana, James M. Cowley, Lisa A. O'Donovan, Tycho R. Neumann, Shi Fang Khor, Nathan S. Watson-Haigh, Tina Bianco-Miotto, Rachel A. Burton

Principal Author

Name of Principal Author (Candidate)	Lina Herliana
Contribution to the Paper	Collected samples and performed observation, processed, analysed, and interpreted morphological and transcriptomic data, and wrote the manuscript.
Overall percentage (%)	75%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 8/6/22

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);
ii. permission is granted for the candidate in include the publication in the thesis; and
iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	James M. Cowley
Contribution to the Paper	Provided measurement of fruit sizes and contributed to the preparation of the manuscript.
Signature	Date 8/6/22

Name of Co-Author	Lisa A. O'Donovan
Contribution to the Paper	Performed sectioning and developmental microscopy and contributed to the preparation of the manuscript.
Signature	Date 8/6/22

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Tycho R. Neumann		
Contribution to the Paper	Collected samples and performed RNA extraction.		
Signature		Date	8/6/22

Name of Co-Author	Shi Fang Khor		
Contribution to the Paper	Collected samples and performed RNA extraction.		
Signature		Date	8/6/22

Name of Co-Author	Nathan S. Watson-Haigh		
Contribution to the Paper	Assisted with bioinformatic analysis.		
Signature		Date	8/6/22

Name of Co-Author	Tina Bianco-Miotto		
Contribution to the Paper	Helped to design experiments and edited the manuscript.		
Signature		Date	8/6/22

Name of Co-Author	Rachel A. Burton		
Contribution to the Paper	Conceived the project, designed experiments, revised the manuscript, and the corresponding author.		
Signature		Date	8/6/22

Title Morphological and transcriptomic analysis of developing *Plantago ovata* seed capsules reveal structural features and key gene networks

Authors Lina Herliana[1], James M. Cowley[1], Lisa A. O'Donovan[1], Tycho R. Neumann[4], Shi Fang Khor[1], Nathan S. Watson-Haigh[2,3], Tina Bianco-Miotto[1], Rachel A. Burton[1]

[1]School of Agriculture, Food and Wine, Waite Research Institute, University of Adelaide, Waite Campus, Urrbrae, SA, Australia

[2]South Australian Genomics Centre (SAGC), SA, Australia

[3]Australian Genome Research Facility, Victorian Comprehensive Cancer Centre, Melbourne, VIC 3000, Australia

[4]IP Australia, PO Box 200, Woden, ACT, 2606, Australia

Keywords *Plantago ovata*, capsules, development, shattering, microscopy, transcriptome, gene

Abbreviations

ace	accelerato
BP	Biological process
Chr	Chromosome
CPM	Count per million
DEG	Differentially expressed gene
DPA	Days post anthesis
DZ	Dehiscence zone
FC	Fold change
FDR	False discovery rate
GO	Gene ontology
GSEA	Gene set enrichment analysis
KEGG	Kyoto encyclopaedia of genes and genomes
KO	KEGG ortholog
OXPHOS	Oxidative phosphorylation
PCA	Principal component analysis
RNAseq	RNA sequencing
WT	Wild type

Abstract

Seed shattering is a natural phenomenon displayed by dehiscent-type fruits that break open at maturity. This undesirable trait in domesticated plants, including the commercially important *Plantago ovata*, causes high yield losses, especially when triggered by extreme weather. Unlike the model plant *Arabidopsis* and economically important crops rice, barley, wheat and legumes, physical and molecular mechanisms controlling *P. ovata* fruit (capsule) development and shattering remain unknown. Here we identified morphological and transcriptional changes through capsule development using microscopy and RNA-sequencing, respectively. Microscopy revealed unique structural capsule features, including a cell layer present only in the lid with thickened xylan-rich walls that are also lignified, that joins the lid to the base via a novel operculum hook. Dehiscence occurs via a series of abscission and separation events. For transcriptomic analysis, data were generated from capsule tissues at two different developmental stages (younger (8-10 DPA) and older (12-14 DPA)) and from a parental wild type and a mutant line *ace*, generated by gamma irradiation, that displays accelerated capsule development. A total of 18,175 genes were expressed in capsule tissues with 1,722 differentially expressed genes (DEGs) in older compared to younger capsules. GSEA enrichment analysis showed that cell wall genes were primarily upregulated in older capsules. Four putative master regulators were identified, namely *PoMYB26*, *PoVND7*, *PoNST1*, and *PoMYB83* and *PoSTK*, known to control fruit size and seed abscission, showed altered expression in *ace* samples but not WT. In *ace* samples, downregulation of the oxidative phosphorylation (OXPHOS) cluster is associated with genes carrying mutations that might putatively be involved in capsule development, and that may also feedback to *PoSTK*, *PoADPG1* and photosynthesis-related genes. This work has greatly increased our understanding of *P. ovata* capsule biology and is likely to be invaluable in helping decipher the shattering mechanism/s, relevant for agronomic improvement of not only *P. ovata* but also other crops.

Introduction

Plantago ovata, psyllium or Isabgol has been cultivated for its husk, which is milled off the outer seed surface and turns into a gel-based mucilage upon wetting. Psyllium husk has been shown to have health benefits including lowering blood cholesterol levels and is useful in food applications such as gluten-free products (Cowley et al., 2020; Phan et al., 2020; Cowley and Burton, 2021), making it economically valuable. Currently, the demand for psyllium is increasing and thus prices are rising. Data from India, the largest global exporter, shows that in the last decade (2011-2021), the export of psyllium husk increased by 25% (40,512.16 to 50,442.71 tons) and the price rose by 162%, from $99.72 million to $261.44 million (Govt. India, Dept. of Commerce 2021), doubling the cost of psyllium husk per kg from $2.50 to $5.20. This crop is susceptible to environmental changes and seasonal availability and fluctuations of husk production combined with commodity stock from the previous year control the price (RACP, 2016). For example, from April 2016 to March 2017, the Isabgol price fluctuated between 8,000 and 11,000 INR in Rajasthan, India (RACP, 2016). Seed shattering or capsule dehiscence contributes to high yield loss, and reduction of seed quality, during unfavourable weather events (Cowley et al., 2022). Patel et al. (2018) reported that water dripping onto seed heads could induce seed shattering up to 70% in the variety Gujarat Isabgol (GI) 3. If we cannot find varieties resistant to shattering then increasing climate change with more unpredictable weather events will further threaten the psyllium industry. Understanding the mechanism(s) underlying shattering could provide benefit in guiding the selection of plants in breeding programs with reduced or non-shattering (indehiscent) traits that are critical to minimise harvest loss.

Elimination of seed shattering traits is crucial in crop domestication before improving other aspects such as increasing starch or oil contents (Lin et al., 2007; Olsen, 2012). Rice is a prime example where shattering has been investigated at the genetic level and several studies show

that changes in one or multiple genes can underlie the acquisition of shattering resistance by disrupting the abscission layer (AL) or dehiscence zone (DZ) in seed head tissues.

Lamba and Gupta (1981) provided a general description of the development of the dehiscence zone in *P. ovata* capsules that has not been much refined in 40 years. Certainly, the molecular mechanisms underlying shattering have not been investigated. Candidate genes from *A. thaliana* or other species can be targeted in *P. ovata* to help breed non-shattering cultivars but these will be of limited value if the shattering mechanisms are fundamentally different at the physical or genetic levels. Here, a detailed physical update of capsule development and structure is presented combined with a forward genetic approach via RNA sequencing to define candidate genes and pathways important across capsule development. Events in wild-type capsules are complemented by data from a new mutant, *accelerato* (*ace*), which displays more rapid capsule development.

Material and Methods

Plant materials and sample collections

Two *P. ovata* genotypes used in this study were WT and *accelerato* (*ace*); a 5th generation selfed gamma-irradiated mutant isolated from the population described by Tucker et al. (2017) (Figure 1a). This study complies with local and national guidelines. Plants were grown in an environmentally controlled plant growth room with lights on a 16/8 hrs light/cycle at 22-23ºC at the University of Adelaide. Flowers with freshly emerged anthers (erect and bright yellow) were marked with coloured fine-tip permanent markers and tagged with the date in order to harvest at the relevant day post-anthesis (DPA) following Phan et al. (2020). Inflorescences were harvested, and fruits were removed and dissected using a scalpel and fine-tip tweezers under a dissection microscope to remove the florets and bracts from either end.

For capsule and seed morphological observations from anthesis to 25 DPA, 5-6 biological replicates were imaged using a Zeiss Stemi 2000-C dissecting microscope with an attached AxioCam ERC 5s camera. Fruit areas were measured using the freehand selection tool in Fiji ImageJ v1.51. For sectioning material, capsules were nicked using a razor blade at one end and placed immediately in 80% (v/v) ethanol containing 0.25% (w/v) glutaraldehyde and 4% (w/v) paraformaldehyde and stored at 4°C. Samples were then dehydrated in a graded ethanol series and infiltrated with LR White resin (Phan *et al* 2016) before being polymerised in gelatine capsules for 48h at 60°C. Longitudinal sections (700nm) were cut on an ultramicrotome (Leica UC6) using a diamond knife (Diatome) and dried onto poly-L-lysine coated glass slides. Survey sections for morphology were stained with 1% (w/v) Toluidine Blue containing 1% (w/v) Borate in Distilled Water. To detect lignin deposition in the capsule, sections of *P. ovata* wild type and capsule mutant (*ace*) were covered with a saturated phloroglucinol solution in 20% HCL and imaged on a Nikon Ni-E optical microscope.

For RNA isolation, the fresh capsules at 8, 10, 12, and 14 DPA were separated from flower and developing seed parts, then immediately frozen in liquid nitrogen and stored at -80ºC until required. These four-time points were selected for the study of capsule development due to the contrast in dehiscence zone (DZ) position (Figure 1d).

Immunolabelling

Sections on glass slides were re-hydrated with PBS, incubated with 0.05m glycine to inactivate residual aldehyde groups (20mins) followed by blocking with 1% (w/v) bovine serum albumin (BSA) in PBS for 20mins. Sections were incubated with monoclonal antibodies raised against pectin (LM19; methylesterified homogalacturonan and LM20, un-esterified homogalacturonan), heteroxylan (M139), and branched chain xylan (M109) diluted 1 in 10 (Kerafast, USA) for 2h at room temperature. Sequential labelling was carried out where possible with washing three times in blocking buffer between antibody addition i.e. when antibodies had been raised in different species (i.e., anti-rat and anti-mouse). After washing with blocking buffer, sections were incubated with the appropriate Alexa Fluor® 488 goat anti-mouse IgG (H+L) or Alexa Fluor® 555 goat anti-rat IgG (H+L) (diluted 1:100, Invitrogen, Australia) for 1h at room temperature. The His-tagged carbohydrate-binding module CBM3a (crystalline cellulose, PlantProbes, Leeds, UK) was used with a triple indirect immunofluorescence labelling procedure as described previously (Phan *et al* 2016). Sections were washed with blocking buffer and counterstained with 0.1% (w/v) Calcofluor White for 90s, washed with water and mounted using Fluroshield (Sigma-Aldrich). Images were obtained using a Zeiss Axio Scan.Z1 slide scanner with Quad-band filter for DAPI, GFP, Cy3 and Cy5. Post processing was identical for all images and included removal of out of focus light and linear unmixing of overlapping spectra.

RNA extraction, cDNA synthesis and sequencing

Total RNA was extracted using the SpectrumTM Plant Total RNA kit (Sigma-Aldrich, USA) following the manufacturer's instructions. The Superscript® III Reverse Transcriptase kit (Invitrogen, USA) was used to synthesise cDNA according to Burton et al., (2008). Three biological replicates were used to generate cDNAs for each capsule developmental stage. The RNA concentration was determined using Qubit® RNA Assay Kit and a Qubit®2.0 Fluorometer (Life Technologies, CA, USA). Total RNA samples were submitted to the Flinders Genomics Facility, Bedford Park, SA 5042. A total of 24 cDNA libraries were sequenced with a read length of 150 bp and paired-end sequencing on the Illumina NovaSeq platform. The raw sequences are available at the SRA NCBI database under accession numbers SRR14643407 – SRR14643423 and SRR14643425 – SRR14643432. A reviewer link can be found at https://dataview.ncbi.nlm.nih.gov/object/PRJNA732452?reviewer=pnqfc729ma1fet4tcg5pbia040.

Pre-processing RNA-seq data and mutation detection

The RNA-seq data was processed by quality checking, trimming, cleaning, aligning reads to the reference genome, and variant calling. The workflow was adapted from Coudray et al. (2018). Quality checking was performed using FastQC v0.11.9 (Andrews, 2017) and MultiQC v1.8 (Ewels et al., 2016) with default parameters. Trimmomatic v0.39 (Bolger et al., 2014) removed adapter and PCR primer fragments. BBDuk, BBmap v38.87 (Bushnell, 2014) was used to remove unwanted reads. Unwanted reads originated from the *P. ovata* chloroplast genome, one mitochondrial gene, and ribosomal and transfer RNA (rRNA and tRNA) genes. Significant ribosomal RNA and chloroplast contents (>95%) were detected in two out of 24 samples. Due to the contamination in different sample groups, we removed one replicate from each group, so only 16 samples were used for final analysis. Clean reads were aligned to the reference genome using STAR v2.7.6a with a 2-pass procedure (Dobin et al., 2013). Before

variant calling, reads were sorted and indexed using SAMtools (Li et al., 2009), and duplicated reads were marked and removed using Picard MarkDuplicates (RRID: SCR_006525). Variant calling was done using GATK v3.8 MuTect2 (McKenna et al., 2010) with multi-sample mode then filtered using FilterMutectCalls (McKenna et al., 2010). Finally, the genetic variants were annotated and predicted using SnpEff v5.0e (Cingolani et al., 2012).

RNA-seq Analysis

After counting reads with gene-level summarisation using featureCounts (Liao et al., 2014), the first step in differential expression analysis was performed using edgeR (Robinson et al., 2010). Count per million (CPM) and The Trimmed Mean of the M-values (TMM) methods were used to normalise count data. For gene presence and absence analysis, genes from WT and *ace* samples were filtered separately where they had to meet two criteria; an expression level as CPM > 0.5 and expression in more than three samples, then an interaction function in R was applied. We kept only genes with CPM > 1 and expressed in more than four samples for any combination of genotypes and time-course points for clustering analysis with factoextra R package (PCA = Principal Component Analysis) (Kassambara and Mundt, 2017) and Clust v1.12.0 (Abu-Jamous and Kelly, 2018). The same parameters were applied for finding differentially expressed genes (DEGs) with limma (Ritchie et al., 2015) and for identifying enriched gene sets using GSEA (Gene Set Enrichment Analysis) (Subramanian et al., 2005).

Morphological observation and clustering analysis with PCA showed capsule pairs 8 and 10 DPA and between 12 and 14 DPA were phenotypically similar to each other and so clustered into two groups (young and old capsules). Therefore, we compared only young vs old instead of combinations between all four time-points (8, 10, 12, and 14 DPA). Six comparisons were performed (Table 1). The first, “wtvsmt” is a high-level universal comparison, where the differences in ages were ignored. Labelled as “wt_youngvsold” is a comparison of young to old capsules from only WT, while the comparison “mt_youngvsold” is the same comparison but in

the mutant background. For “youngvsold”, the comparison was made based on age only and genotype was ignored. Comparison “wtvsmt_young” refers to young capsules from WT compared to those from the mutant, while “wtvsmt_old” refers to old capsules from WT compared to those from the mutant.

Table 1. List of six pairwise comparisons between sample groups.

Comparison number	Symbols	Comparison		Number of samples
		Group 1	Group 2	
Comparison I	youngvsold	WT 8-10 DPA and mutant 8-10 DPA	WT 12-14 DPA and mutant 12-14 DPA	16
Comparison II	wt_youngvsold	WT 8-10 DPA	WT 12 -14 DPA	8
Comparison III	mt_youngvsold	mutant 8-10 DPA	mutant 12-14 DPA	8
Comparison IV	wtvsmt	WT 8-10 DPA and WT 12-14 DPA	mutant 8-10 DPA and mutant 12-14 DPA	16
Comparison V	wtvsmt_young	WT 8-10 DPA	mutant 8-10 DPA	8
Comparison VI	wtvsmt_old	WT 12-14 DPA	mutant 12-14 DPA	8

Using the limma package (Ritchie et al., 2015), the list of DEGs was obtained by applying the Benjamini and Hochberg (BH) correction of the p-values to reduce false positives and sorting by p-values. The significant DEGs were determined according to the adjusted p-values (FDR < 0.05) and an absolute value of the log-fold-change (logFC) > 1. VennDiagram (Chen and Boutros, 2011) and ComplexUpset (Lex et al., 2014) were used to visualise the number of DEGs.

Using the modified GOseq and combining two methods from Young et al. (2012) and the Trinity software package (https://github.com/trinityrnaseq/trinityrnaseq/wiki/Running-GOSeq), we performed gene ontology (GO) enrichment on DEGs (sorted by p-value, FDR < 0.05 and logFC!=0). Using the GSEA v4.1.0 (Subramanian et al., 2005) desktop application, we performed enrichment analysis on the normalised expression data, not the DEG list. We used a customised gene sets database (*P. ovata* GO term database), selected 1000 as the number of permutations without collapsing the dataset and used the gene set as permutation type and

selected a paired phenotype. For example, we selected MT vs WT as a pair to compare WT (WT) and mutant (MT). In this case, MT will have a positive enrichment score, and WT will have a negative score. Cytoscape v3.8.0 application (Shannon et al., 2003) with Enrichment Map installed was used to visualise the GSEA results.

P. ovata sequences and annotations are available at DDBJ/ENA/GenBank under JAHHQI010000000. Gene Ontology (GO) term and KEGG ortholog (KO) identifiers were obtained by submitting *P. ovata* protein sequences to the eggnog-mapper website (Cantalapiedra et al., 2021).

Results

Development of *Plantago ovata* capsules

P. ovata flowers were arranged in indeterminate inflorescences displaying asynchronous flowering, opening from the bottom to the top of the spike (Figure 1a-b) across several days. After pollination, the fruits, referred to here as capsules, and seeds inside them, underwent a series of changes (Figure 1) that can be divided into four stages: fruit set (1-7 days post anthesis, expansion (8-17 DPA), ripening (18-23 DPA), and maturation (23 DPA onwards). The *P. ovata* capsule was an ovoid pyxidium that was divided into two valves by a circumscissile dehiscence zone (Figure 1d). The upper valve was the operculum, or lid, that detaches at maturity whilst the bottom valve, or base, remained attached to the flowering spike (Figure 1h). During the first week of fruit set, capsule size increased threefold (Figure 1d). The fertilised seeds developed rapidly; initially they were mainly comprised of integument layers which were then compressed by the developing endosperm and embryonic cotyledon and radicle (Figures 1e and f). The future DZ was located towards the bottom of the young capsule at this early stage (Figure 2, Supplementary Figure S1). During the second stage of expansion, the capsule continued to increase in size and pigmentation started appearing across the upper valve (Figure 1d, Figure 2 and Supplementary Figure S1). Rapid changes in size were noticed in two main phases between 8-10 DPA and 11-14 DPA. During the next stage the capsule matured and seed growth rates slowed until they reached their maximum size, at around 17-18 DPA (Figure 1d and e, Supplementary Figure S2). During this stage, the dehiscence zone became more prominent (Figure 2) at the middle of the capsule at around 14 DPA, henceforth dividing it into equally-sized upper and lower valves (Figure 2). Both capsules and seeds turned brown as they ripened (Figure 1d and e, Supplementary Figures S1 and S2) and the first of the WT capsules shattered at 25 DPA.

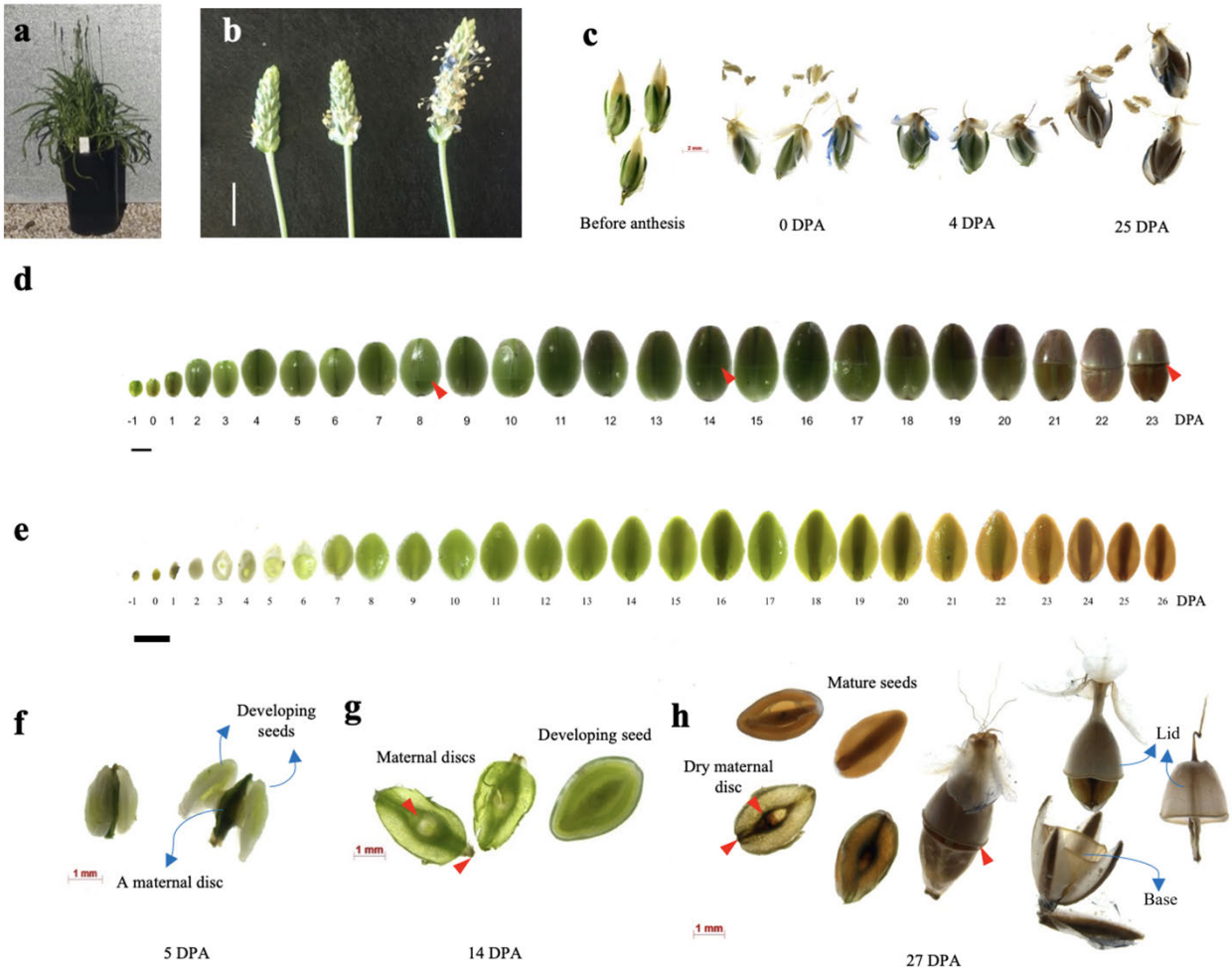


Figure 1. *P. ovata* organ development. (a) A WT 3-month-old plant, 30 cm tall and flowering; (b) Flowers are arranged in an indeterminate inflorescence; (c) Floral organs pre- and post-anthesis; (d) Developing capsules from -1 to 23 days post anthesis (DPA); (e) Developing ovules/seeds from -1 to 26 DPA; Developing seeds and the maternal disc at (f) 5 DPA (g) 14 DPA; (h) mature seeds, the dry maternal disc and mature capsules; complete, dehisced lid containing seeds, empty lid and empty base. Scale bars: 1 cm (b); 2 mm (c, e); 1 mm (d, f-h); Red arrows indicate abscission sites in (g) and (h) and DZ in (d) and (h). Images acquired used dissecting microscope (c-h).

Figure 2. *Plantago ovata* fruit development between 8 and 20 days post anthesis (DPA) showing the change in the position of the dehiscent zone on the fruit (arrowhead). The zone itself is not migrating, but instead, increasing seed size causes stretching of the base (B) which increases significantly in size during development while the lid (L) stays similarly sized. Scale = 1 mm.

Internal capsule morphology through development to dehiscence

Longitudinal sections through the capsule were stained with toluidine blue to reveal internal capsule morphology (Figure 3). Both base and lid were composed of at least five cell layers. The outermost epidermal layer on both valves had heavily thickened cell walls but the cells on the lid were square whereas on the base they were elongated (Figure 3a). On the base there were four layers of inner parenchyma cells which appeared to get somewhat narrower or compressed as the capsule aged. Inside the epidermal layer on the lid was a layer of thin walled parenchyma cells followed by a layer of smaller, blocky cells with the walls closest to the capsule exterior displaying heavy thickening. The final cell layer on the very inside of the lid was long, thin-walled parenchyma cells (Figure 3a.).

The anatomy of the dehiscence zone was clearly visible even at the early stages of capsule development, at 8 DPA, somewhat resembling a human knee joint (Figure 3a). Where the lid met the base in the DZ, additional parenchyma cells between the epidermal layer and the inner

thickened layer widen the rim of the lid, making a terminal bulge. In the DZ on the outer surface of the capsule where the epidermal and parenchyma layers met, there were two layers of very small cells backed by an internal air space. This air space constituted an inner circumscissile groove that ran around the entire capsule. Behind the circumscissile groove was the inner layer of small blocky sclerenchyma cells with the thick outer walls, that extended past the other cell layers of the lid to attach to the terminus of the base, in a structure we have called the operculum hook (Figure 3b). This hook anchored the two valves of the capsule together.

As the capsule reached the later stages of maturity the layers of very small cells external to the circumscissile groove separated (Figure 3c and 3d) - potentially this was a natural abscission event, deepening the gap between the two valves. However, the capsule did not fully dehisce until the operculum hook detached from the base, and the lid, with the seeds inside, came off (Figure 3e).

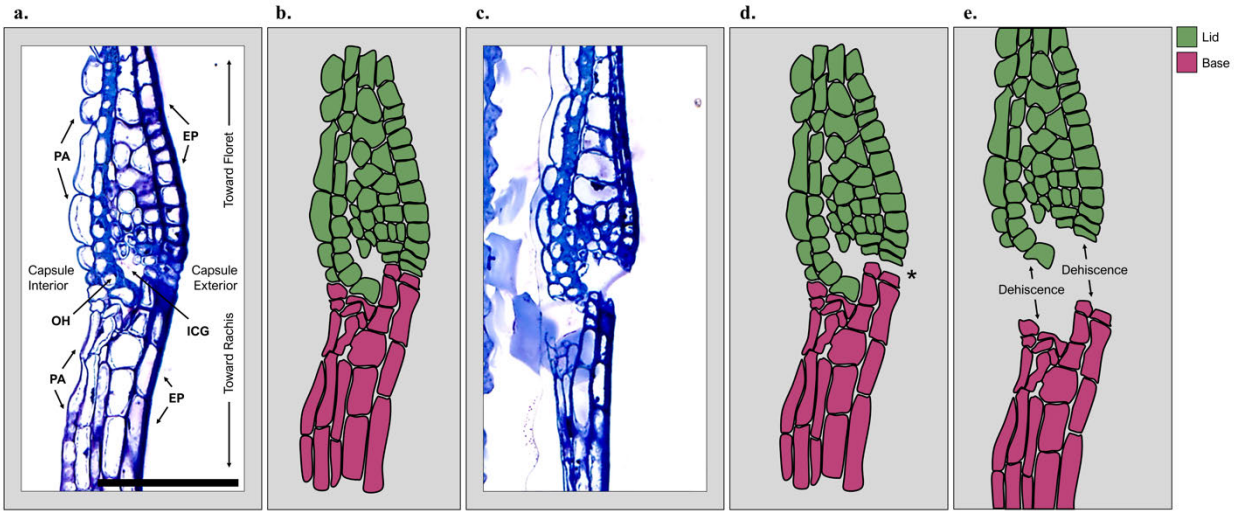


Figure 3. a. Prior to maturity and dehiscence, key structures of the dehiscent zone are visible in thin sections stained with toluidine blue. **b.** A colour-coded schematic shows which cells in a. belong to the lid or base. **c.** Thin sections of mature dehiscent zones stained with toluidine blue show the distinct interlocking structures of the lid and base of the capsule. **d.** at maturity, a gap is visible at the dehiscent zone on the outer surface of the capsule indicating two distinct dehiscence events occur, with the exterior side dehiscing first (asterisk) before **e.** the main capsule dehiscence occurs. Abbreviations: EP, epidermal cells; ICG, inner circumscissile groove; OH, operculum hook; PA, parenchyma cells. Scale in a is 50 µm and also applies to c.

External capsule changes through development to dehiscence

The DZ was clearly visible on the capsule surface from 8 DPA onwards (Figure 2). At these early stages the capsule base was smaller than the lid and expanded rapidly, being stretched as the seeds developing inside pushed the base downwards, making the tissue thinner and highly wrinkled (Figure 4a). This had the effect of pushing the DZ upwards until it became equatorial at 14 DPA (Figure 2). From this stage onwards the surface of the base was much more wrinkly than the lid, which stayed smooth and started to become pigmented (Figure 4a). During ripening the edges of the valves at the DZ, particularly on the lid, became thicker and protuberant (Figure

4b) and the lower valve of the capsule shrunk inwards (Figure 4a and d). At this stage the rows of small cells on the outer layers where the valves met, through to the circumscissile groove, separated to form a clear gap (Figure 4a, b and d) but the capsule had not yet dehisced. This did not occur until the operculum hook finally detached from the base and the lid and seeds came free, revealing the uneven edge on the base where the two outer cell layers sit in front of the cellular niche where the operculum hook was attached (Figure 4c, e and f).

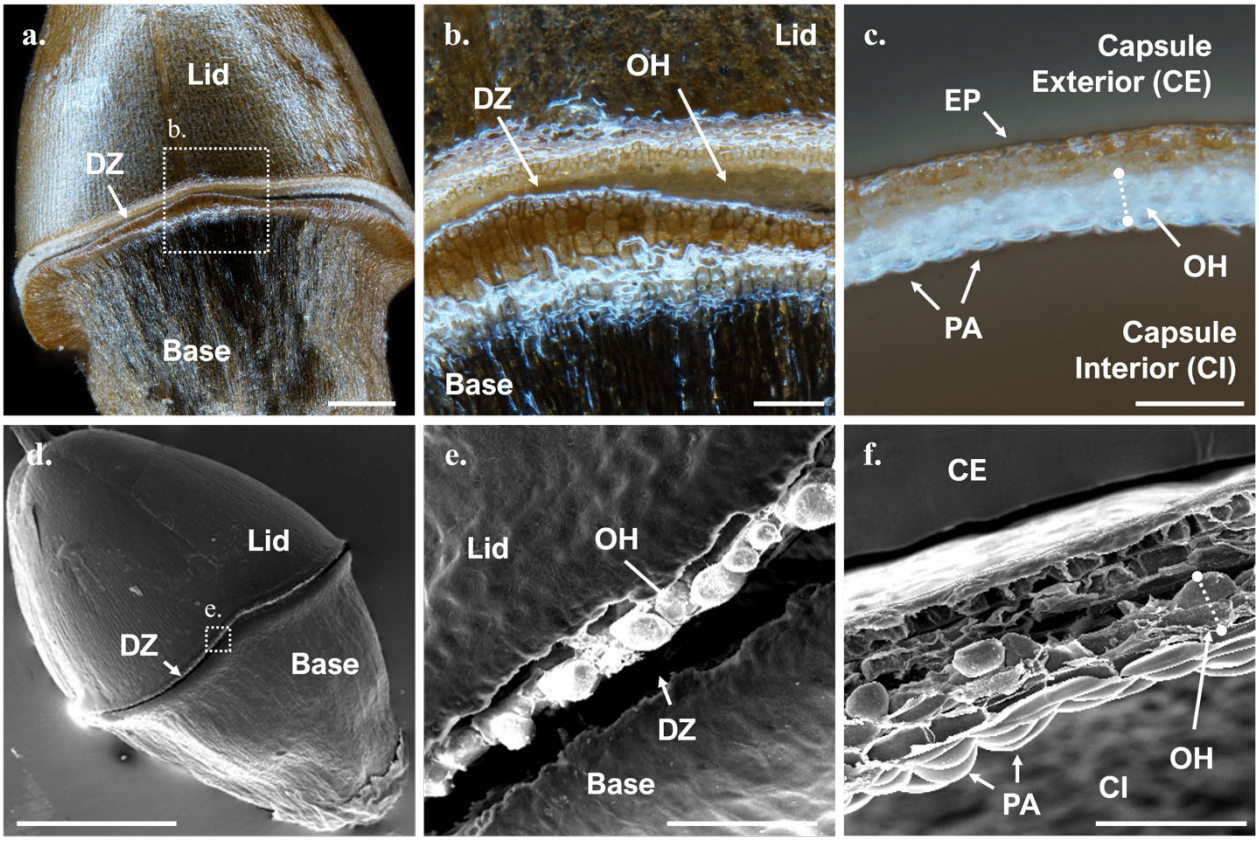


Figure 4. Dissection (**a-c**) and scanning electron (**d-f**) microscopy of mature *Plantago ovata* capsules. **a,d.** at maturity, a gap is visible at the dehiscent zone (DZ) between the lid and the base. **b,e.** within the gap of the dehiscent zone, the front face of the operculum hook (OH) is visible. **c,f.** when the transverse surface of the dehiscent zone is viewed after dehiscence, the cells of the operculum hook are visible and distinct from other capsule cells, particularly the bulbous nature of the parenchyma (PA). Scale = 200 µm (a); 100 µm (b); 50 µm (c, e, f); 1 mm (d).

Capsule cell wall composition

The cell walls of the capsule through development from 8 to 20 DPA, were labelled with calcofluor white which binds to cellulose and M139 (Ruprecht et al., 2017) which binds to xylan. The calcofluor white delineated the entire capsule, showing an ovoid shape (Figure 5, top panel) and bound to numerous internal seed tissues. The M139 antibody selectively bound more strongly to the lid of the capsule and some of the internal seed tissues from 16 DPA (Figure 5 top and bottom panels). A more detailed labelling of the DZ from 8 to 20 DPA is shown in Figure 6, using a range of stains, antibodies and binding modules to reveal polysaccharide composition. Staining with calcofluor white and binding of CBM3a (Ruel et al., 2012), which detects crystalline cellulose, indicated that cellulose was present in all cell layers, particularly from 12 DPA onwards. Antibodies LM19 and LM20 (Verhertbruggen et al., 2009), which bind to homogalacturonan in its unesterified and esterified forms respectively, both strongly labelled the walls of the outer epidermal cell layer of lid and base, indicating the presence of pectic polysaccharides. Two antibodies were used to detect xylans and showed differential labelling. M109 (Ruprecht et al., 2017) consistently labelled the outer surface of the capsule, that was the outer wall of the epidermal cell layer, of both the lid and base, from 10 DAP onwards. Labelling of the internal cell layers was minimal. In contrast, M139 strongly labelled walls in the cell layer that forms the operculum hook and labelling was absent from any cell layers in the base, and it also bound to the outer surface of both lid and base (Figure 6). Labelling of whole longitudinal capsule cross sections with M139 shows strong signals extending up this cell layer to the capsule apex (Figure 5).

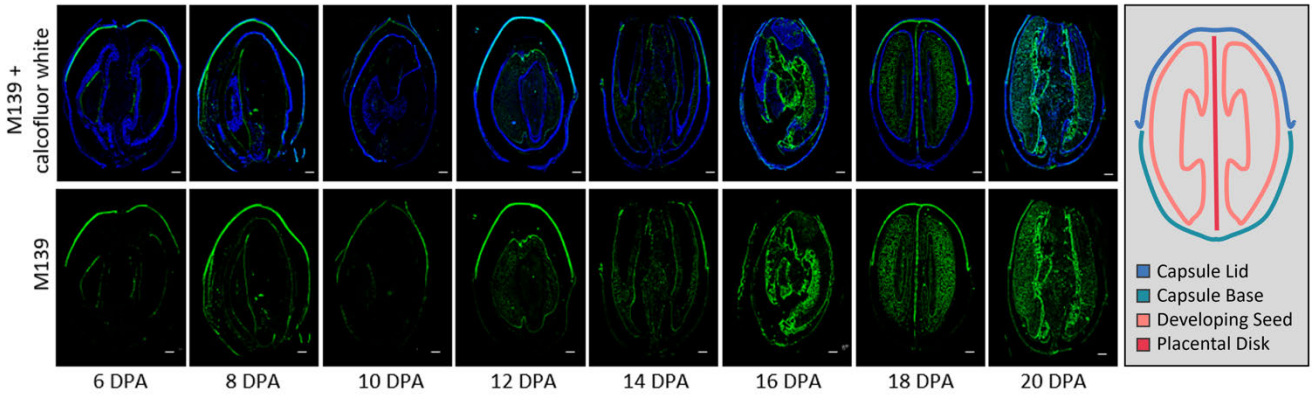


Figure 5. Immunodetection of xylan in longitudinal sections of developing *Plantago ovata* fruit. Fruit were immunolabelled with CCRC M139 antibody which recognises low arabinose-substituted xylan backbone (**a**; merged with calcofluor white), and counterstained with calcofluor white (**b.**) which broadly recognises cell wall glucans. Scale = 200 µm. Included is a schematic (**c.**) showing the different tissues depicted in microscopy images.

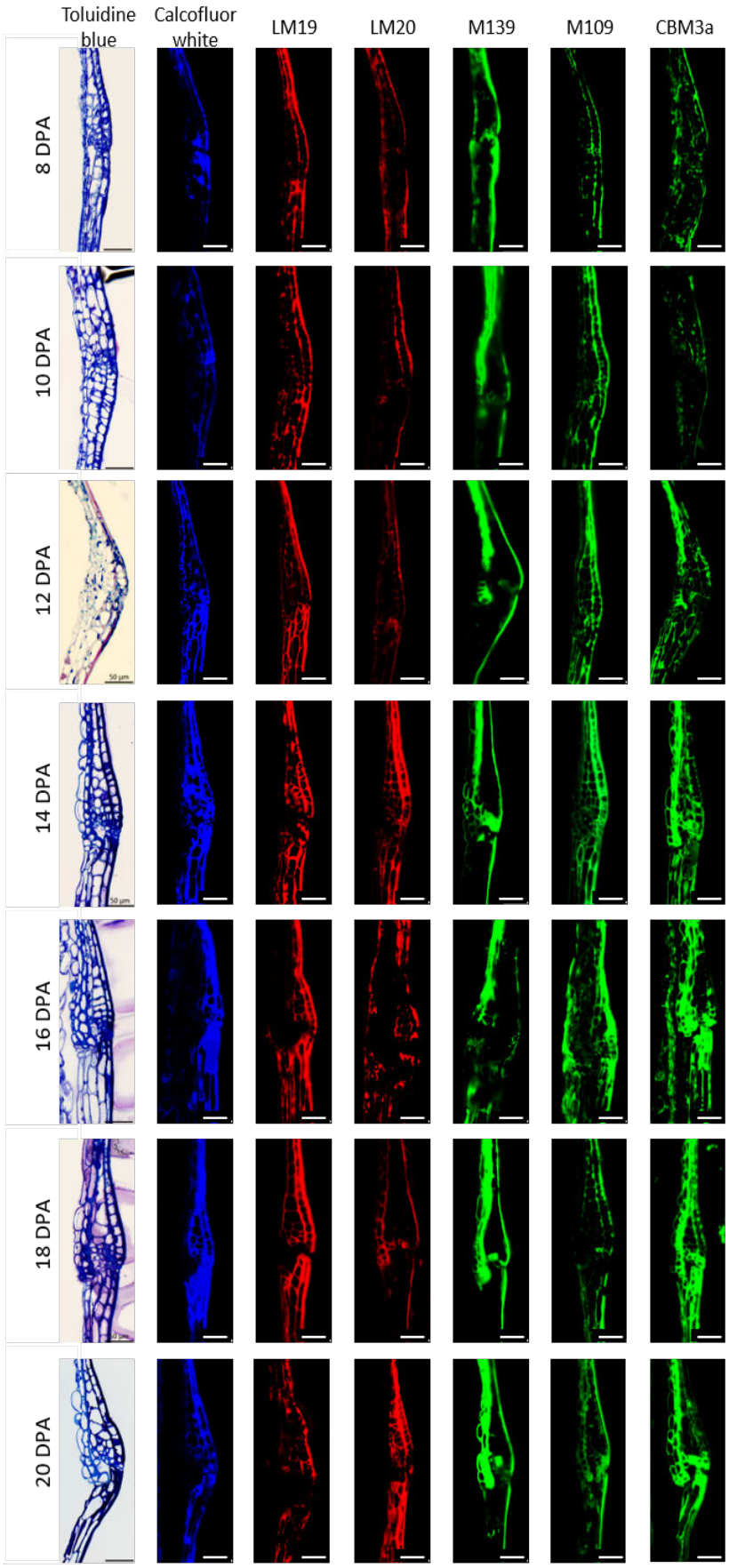

Toluidine blue
Calcofluor white
LM19
LM20
M139
M109
CBM3a
8 DPA
10 DPA
12 DPA
14 DPA
16 DPA
18 DPA
20 DPA

Figure 6. Staining and immunolabelling of the dehiscence zone (DZ) on transverse sections of *P. ovata* capsules. The sections were stained with toluidine blue and calcofluor white and labelled with cell wall antibodies to detect pectic polysaccharides using LM19 (unesterified HG) and LM20 (methyl esterified HG), xylan using M139 (Heteroxylan) and M109 (branched xylan), and cellulose using the carbohydrate-binding molecule CBM3a. All images have the same scale bar = 50 µm.

Based on the published report by Lamba and Gupta, capsule tissues from 12 to 20 were stained for lignin using phloroglucinol. Strong signals were detected only in the thickened walls of the operculum hook (Figure 7).

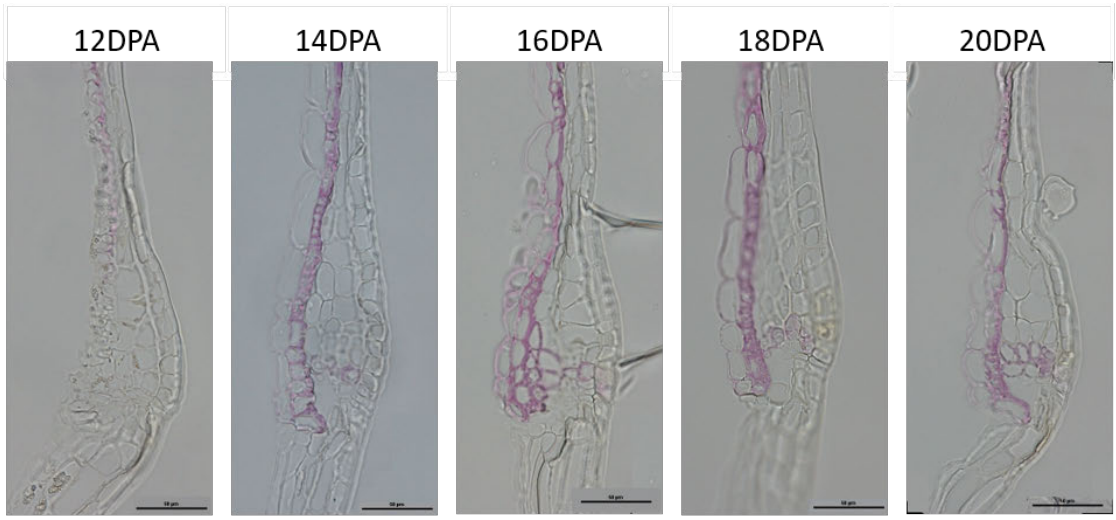


Figure 7. Sections were covered with a saturated solution of phloroglucinol in 20% (v/v) HCl and imaged on a Nikon Ni-E optical microscope. Scale bar = 50 um

The *accelerato* (*ace*) mutant

A gamma-irradiated mutant line of *P. ovata* was identified by screening from a large collection (Tucker et al., 2017). This mutant consistently showed more rapid capsule development and was named *accelerato* (*ace*). The overall developmental pattern of *ace* was similar to WT except the processes occurred more rapidly. The capsules were larger from 2 DPA onwards and

pigment appeared earlier compared to WT (Figure 8a and b). Maximum fruit size was observed at 15.9 DPA for *ace,* compared to 17.3 DPA for WT (Figure 8c) and the capsules started to shatter earlier at around 22/23 DPA (data not shown).

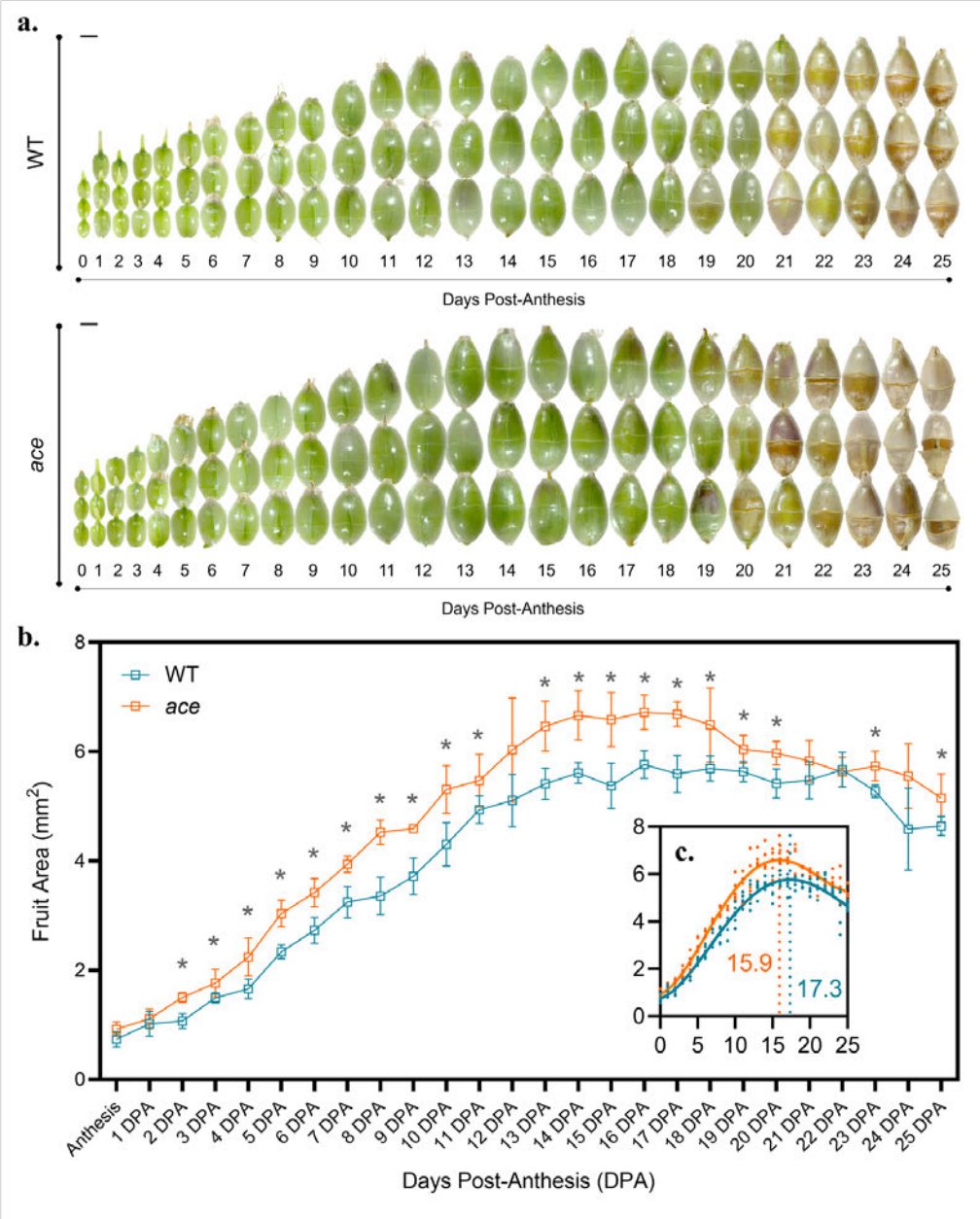


Figure 8. Temporal differences in *P. ovata* capsule development between WT and *ace* from anthesis to 25 DPA.

Transcriptomic analysis of wildtype and *ace* capsules

Transcriptomic analysis was performed to investigate the molecular mechanism underlying capsule development and identify genes associated with more rapid capsule development in a mutant, *ace*. Total RNA was collected from capsule tissues at two different developmental stages, younger (8-10 DPA) and older (12-14 DPA) and from a parental wild type and a mutant line *ace*.

Overall, about 18,175 genes representing 74 % of all known *P. ovata* genes were expressed in WT and *ace* capsules (Figure 9a). About 5,659 genes (26%) were not expressed in capsule tissues at all (Figure 9a). There were 606 genes (approximately 3%) transcribed in WT capsules but not in the equivalent *ace* tissues, including 22 long non-coding RNAs (lncRNAs), one small nuclear RNA, and two small nucleolar RNAs (Supplementary Table S1). Conversely, there were 605 genes expressed in *ace*, but not WT capsules, including seven lncRNAs, three small nuclear RNAs, and one small nucleolar RNA (Supplementary Table S2).

About 81 KO (KEGG ortholog) identifiers for the transcripts were assigned in WT with 121 KO identifiers for *ace*. No entire metabolic pathway was represented in the list of 81 transcripts specific to WT, but one complete metabolic pathway was identified in the list of 121 transcripts that were switched on in *ace* capsules. This pathway was pectin degradation (M00081) that consists of three KO identifiers (K01051, K01184, and K01213) represented by two genes for K01051 and one gene each for K01184 and K01213. These genes encode putative pectinesterase 68 (KN361_2g009302), putative pectinesterase inhibitor 12 (KN361_1g0034472), and two genes encode polygalacturonases (KN361_3g001758 and KN361_3g008374).

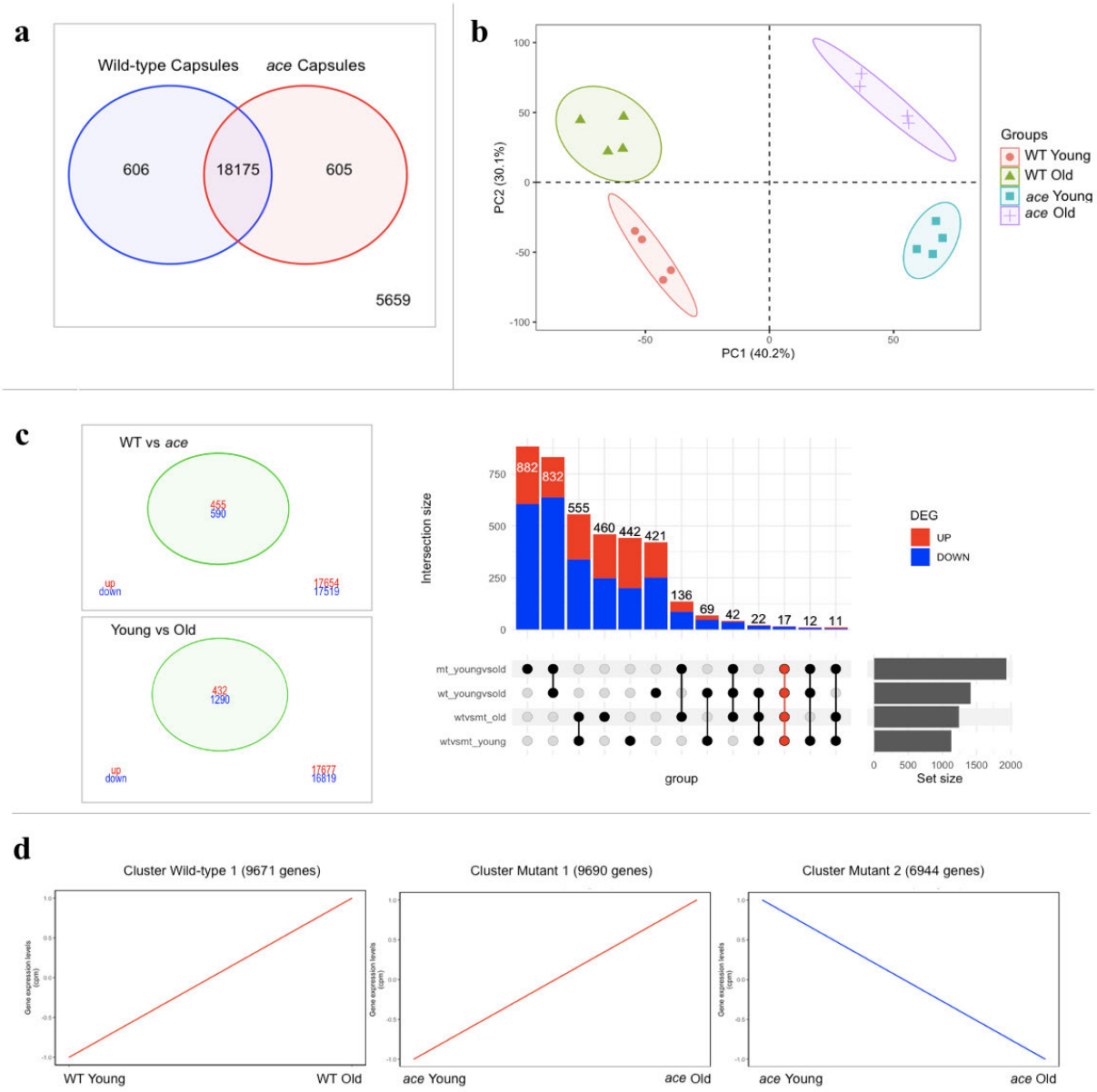


Figure 9. Expressed genes in capsule tissues from WT and *ace* at 8 and 10 DPA (young) and 12 and 14 DPA (old). (a) Venn diagram for normalised RNAseq data from eight samples for each genotype (CPM > 0.5 and expressed in more than three samples) shows specific genes expressed in WT (606 genes), *ace* (605), both genotypes (18,175 genes) and not expressed in capsule tissues of both genotypes (5,659 genes); (b) Principal Component Analysis (PCA) on normalised and filtered RNAseq data (CPM > 1 in more than four samples, n=16 samples); (c) Differentially expressed genes (DEGs) between samples defined using the Venn diagram and ComplexUpset; (d) Gene expression patterns of clusters of co-expressed genes from normalised and filtered gene expression data defined using Clust.

Table 2. Leading edge subset of genes in the plant cell wall cluster.

Gene ID	Protein product
KN361_2G006239	putative beta-1,4-xylosyltransferase IRX10/GUT2
KN361_4G000140	putative beta-1,4-xylosyltransferase IRX10/GUT2
KN361_1G002269	protein trichome birefringence-like 34 TBL34
KN361_3G004788	transcription factor MYB26
KN361_3G004104	UDP-glucuronate:xylan alpha-glucuronosyltransferase 2 GUX2
KN361_1G003640	UDP-glucuronate:xylan alpha-glucuronosyltransferase 1 GUX1
KN361_4G006888	laccase-22 LAC22/IRX12
KN361_1G008021	putative galacturonosyltransferase 12 GAUT12
KN361_3G007382	transcription factor MYB83
KN361_1G011248	NAC domain-containing protein 7 NAC007
KN361_1G004983	cellulose synthase A catalytic subunit 8 UDP-forming CESA8
KN361_1G000637	cellulose synthase A catalytic subunit 7 UDP-forming CESA7
KN361_1G001941	cellulose synthase A catalytic subunit 4 UDP-forming CESA4
KN361_2G009106	putative UDP-glucuronate:xylan alpha-glucuronosyltransferase 3 GUX3
KN361_3G005346	NAC domain-containing protein 12 NAC012
KN361_1G000107	alpha-L-arabinofuranosidase 1 ASD1
KN361_3G003100	putative glucuronoxylan glucuronosyltransferase IRX7

Capsule ages and genotypes explain variation in gene expression among samples

The principal component analysis (PCA) plot (PC1 and PC2) explains variation in sample groups by 70.3% (Figure 9b). PC1 (40.2%) differentiated sample groups based on genotype (WT versus *ace*), while PC2 (30.1%) separated older samples (12 and 14 DPA) from the younger ones (8 and 10 DPA). This gives rise to four separate capsule groups, namely young WT, old WT, young *ace*, and old *ace* (Figure 9b).

Gene expression comparison between two capsule developmental stages

The number of differentially expressed genes (DEGs) between young and older capsules (Comparison I, 'youngvsold') was 1,722, with 25% of these upregulated (432 genes) and 75% downregulated (1,290 genes) in older capsules compared to younger capsules (Figure 9c). The number of DEGs seen for the early versus later stages of capsule development in *ace* (1,932) (Comparison II, 'wt_youngvsold') was almost double that for WT capsules (1,045) (Comparison III, 'mt_youngvsold') (Figure 9c). There were 882 genes differentially expressed during capsule development (young to old) specific to *ace* (Comparison III), while only 421 DEGs appeared across the same time period for WT capsules (Comparison II, 'wt_youngvsold'). About 832 DEGs were linked to common developmental processes occurring as the capsules age in both WT (Comparison II) and *ace* (Comparison III), with only 17 genes differentially expressed across the same time span comparing WT and *ace*. Of note, three of these genes were involved in flavonoid biosynthesis (K00475, K05277, and K13082) and one gene in anthocyanin biosynthesis (K12930).

By performing a Gene Set Enrichment Analysis (GSEA), seven clusters of enriched GO terms were revealed as the capsule ages (Comparison I - III) corresponding to plant cell walls, plant organ senescence, cell cycle, plasma membrane, photosynthesis and two lipid metabolism clusters (Figure 10a). Only two clusters, the upregulated plant cell wall and downregulated photosynthesis-related genes, were significantly altered in older WT capsules (Comparison II,

'wt_youngvsold'). In the mutant (Comparison III, 'mt_youngvsold'), upregulated genes were associated with lipid metabolism and plant organ senescence, while downregulated genes were related to the cell cycle and the plasma membrane (Figure 10a). In addition, GO analysis using Goseq (Figure 10c) showed that genes encoding plant cell wall proteins were also enriched in older mutant capsules (Comparison III).

The plant cell wall cluster consists of GO terms GO:0009834, GO:0045491, and GO:0010410 (Figure 8a). There were 23 genes in GO:0009834 (Biological Process (BP) plant-type secondary cell wall biogenesis), 22 genes in GO:0010410 (BP hemicellulose metabolic process), and 19 genes in GO:0045491 (BP xylan metabolic process) but some of these genes were the same, appearing under multiple GO terms, so removing redundancy leaves 35 genes in total for the plant cell wall cluster. Seventeen (Figure 11c) out of 35 (Figure 11b) of these genes were in the leading-edge subset based on the GSEA enrichment score (Table 2).

Gene expression differences between WT and *ace* across two capsule developmental stages

The number of DEGs between WT and *ace* capsules (Comparison IV, 'wtvsmt') irrespective of time point, was 1,045 (Figure 9c), with 44% (455 genes) upregulated and 56% (590 genes) downregulated. Taking the age of the capsules into account, DEG set size was about 1,128 for young capsules (Comparison V, 'wtvsmt_young') and 1,243 (Comparison VI, 'wtvsmt_old) for old capsules between WT and *ace*.

Comparing WT and mutant all GO terms (Comparison IV, 'wtvsmt') were enriched for downregulated genes in the oxidative phosphorylation (OXPHOS) pathway (Figure 10b). The OXPHOS cluster consists of two GO terms, namely GO:0098798 and GO:0098800. GO terms GO:0098798 (Cellular Component (CC) mitochondrial protein complex) and GO:0098800 (CC inner mitochondrial membrane protein complex) contain 19 and 18 genes respectively. About ten genes were in the leading-edge subset with four genes predicted to encode ATP synthases

(KN361_1G011499, KN361_3G000010, KN361_4G000226, and KN361_4G004188), five genes encode NADH dehydrogenases (KN361_2G008702, KN361_4G000770, KN361_1G006789, KN361_2G001686, and KN361_4G007594), and one gene encoding a serine hydroxymethyltransferase (KN361_4G003985).

According to comparisons based on age (Comparison V and VI), photosynthetic genes in GO:0009535, GO:0042651, GO:0034357, and GO:0044436 were downregulated in young *ace* capsules (Comparison V, 'wtvsmt_young') and upregulated in the older capsules (Comparison VI 'wtvsmt_old'). Only one GO term (GO:0022900, Biological Process (BP) electron transport chain) was consistently upregulated regardless of capsule age (Comparison IV, 'wtvsmt'). Three genes encodes ATP synthase (KN361_4G004110, KN361_4G004108, and KN361_1G005338), three genes for photosynthetic NDH subunit of lumenal location (KN361_3G003680, KN361_1G008017, and KN361_2G003872), two genes for ferredoxin (KN361_3G008734 and KN361_2G007002), one gene for NADPH:adrenodoxin oxidoreductase (KN361_2G007345), one gene for electron transfer flavoprotein-ubiquinone oxidoreductase (KN361_3G001867), protein NDH-dependent cyclic electron flow (KN361_2G007345), and rhodanese-like domain-containing protein 4 (KN361_2G003313). However, the photosynthesis cluster was not significantly different between WT and *ace* for young capsules (Comparison V, 'wtvsmt_young').

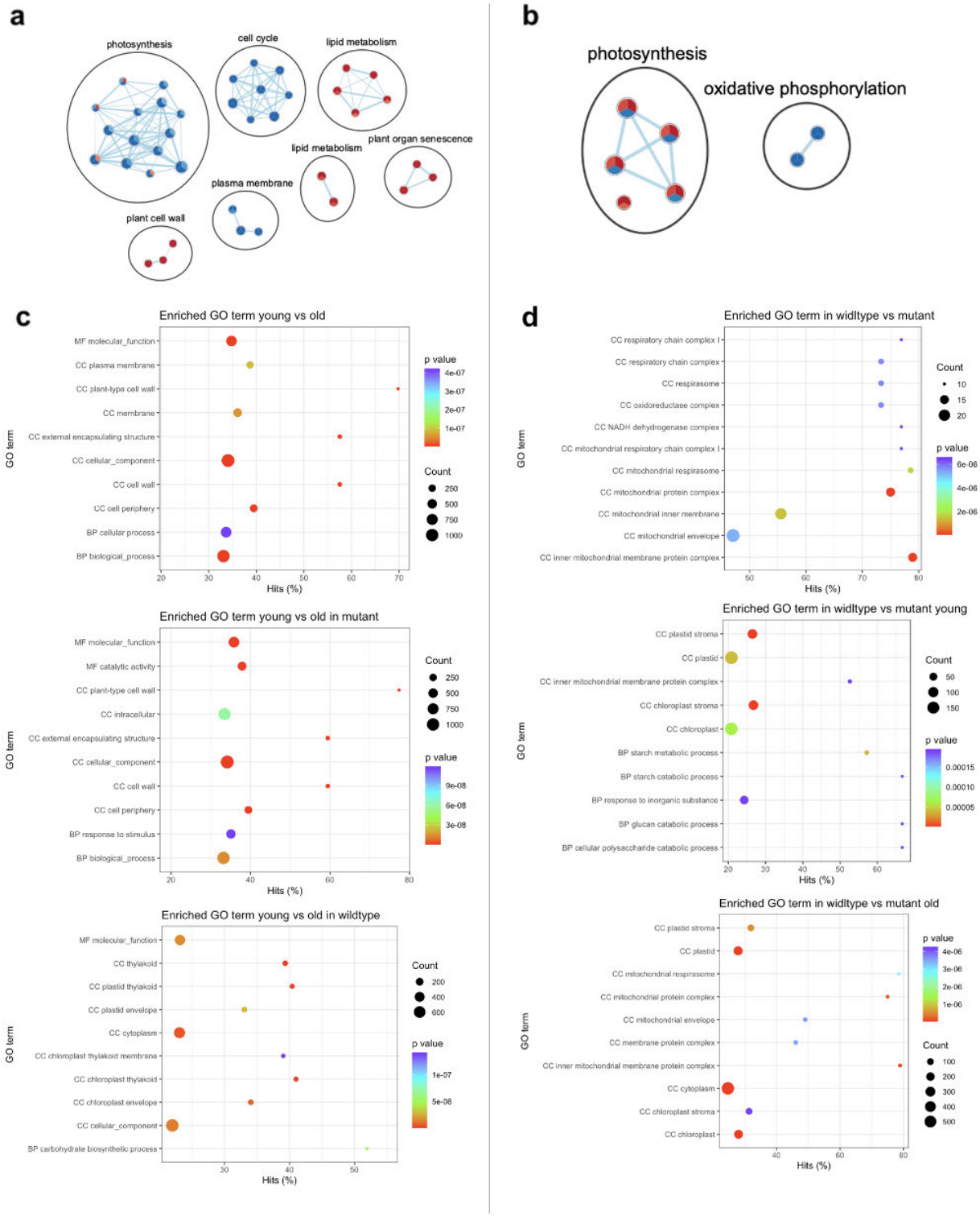

a
photosynthesis
cell cycle
lipid metabolism
lipid metabolism
plant organ senescence
plasma membrane
plant cell wall
b
photosynthesis
oxidative phosphorylation
c
Enriched GO term young vs old
MF molecular_function
CC plasma membrane
CC plant-type cell wall
CC membrane
CC external encapsulating structure
CC cellular_component
CC cell wall
CC cell periphery
BP cellular process
BP biological_process
GO term
Hits (%)
p value
Count
Enriched GO term young vs old in mutant
MF molecular_function
MF catalytic activity
CC plant-type cell wall
CC intracellular
CC external encapsulating structure
CC cellular_component
CC cell wall
CC cell periphery
BP response to stimulus
BP biological_process
Enriched GO term young vs old in wildtype
MF molecular_function
CC thylakoid
CC plastid thylakoid
CC plastid envelope
CC cytoplasm
CC chloroplast thylakoid membrane
CC chloroplast thylakoid
CC chloroplast envelope
CC cellular_component
BP carbohydrate biosynthetic process
d
Enriched GO term in widltype vs mutant
CC respiratory chain complex I
CC respiratory chain complex
CC respirasome
CC oxidoreductase complex
CC NADH dehydrogenase complex
CC mitochondrial respiratory chain complex I
CC mitochondrial respirasome
CC mitochondrial protein complex
CC mitochondrial inner membrane
CC mitochondrial envelope
CC inner mitochondrial membrane protein complex
Enriched GO term in widltype vs mutant young
CC plastid stroma
CC plastid
CC inner mitochondrial membrane protein complex
CC chloroplast stroma
CC chloroplast
BP starch metabolic process
BP starch catabolic process
BP response to inorganic substance
BP glucan catabolic process
BP cellular polysaccharide catabolic process
Enriched GO term in widltype vs mutant old
CC plastid stroma
CC plastid
CC mitochondrial respirasome
CC mitochondrial protein complex
CC mitochondrial envelope
CC membrane protein complex
CC inner mitochondrial membrane protein complex
CC cytoplasm
CC chloroplast stroma
CC chloroplast

Figure 10. Gene ontology (GO) term enrichment analysis on gene sets. Network enrichment analysis (a & b) was built using all expressed genes in all tissues processed using GSEA then visualised using Cytoscape v3.8.0 (FDR q value > 0.1). (a) Overlap of three separate enrichment analyses between combined young and old capsules (Comparison I, 'youngvsold'), WT young and old capsules (Comparison II, 'wt_youngvsold'), and *ace* young and old capsules (Comparison III 'mt_youngvsold'). Each overlapping enriched gene is represented in a node slice (circle). Small circles represent nodes (GO terms) where red indicates upregulated genes and blue shows downregulated genes; the circle size represents gene number, nodes with shared genes are connected by pale blue lines (edges), and all nodes are connected, annotated, and clustered using the MCL algorithm. (b) Overlap of three separate enrichment analyses between WT and mutant capsules for all ages combined (Comparison IV, 'wtvsmt'), all young (Comparison V, 'wtvsmt_young'), and all old (Comparison VI, 'wtvsmt_old'). (c-d) Enriched GO terms on the DEG list using Goseq for different comparisons.

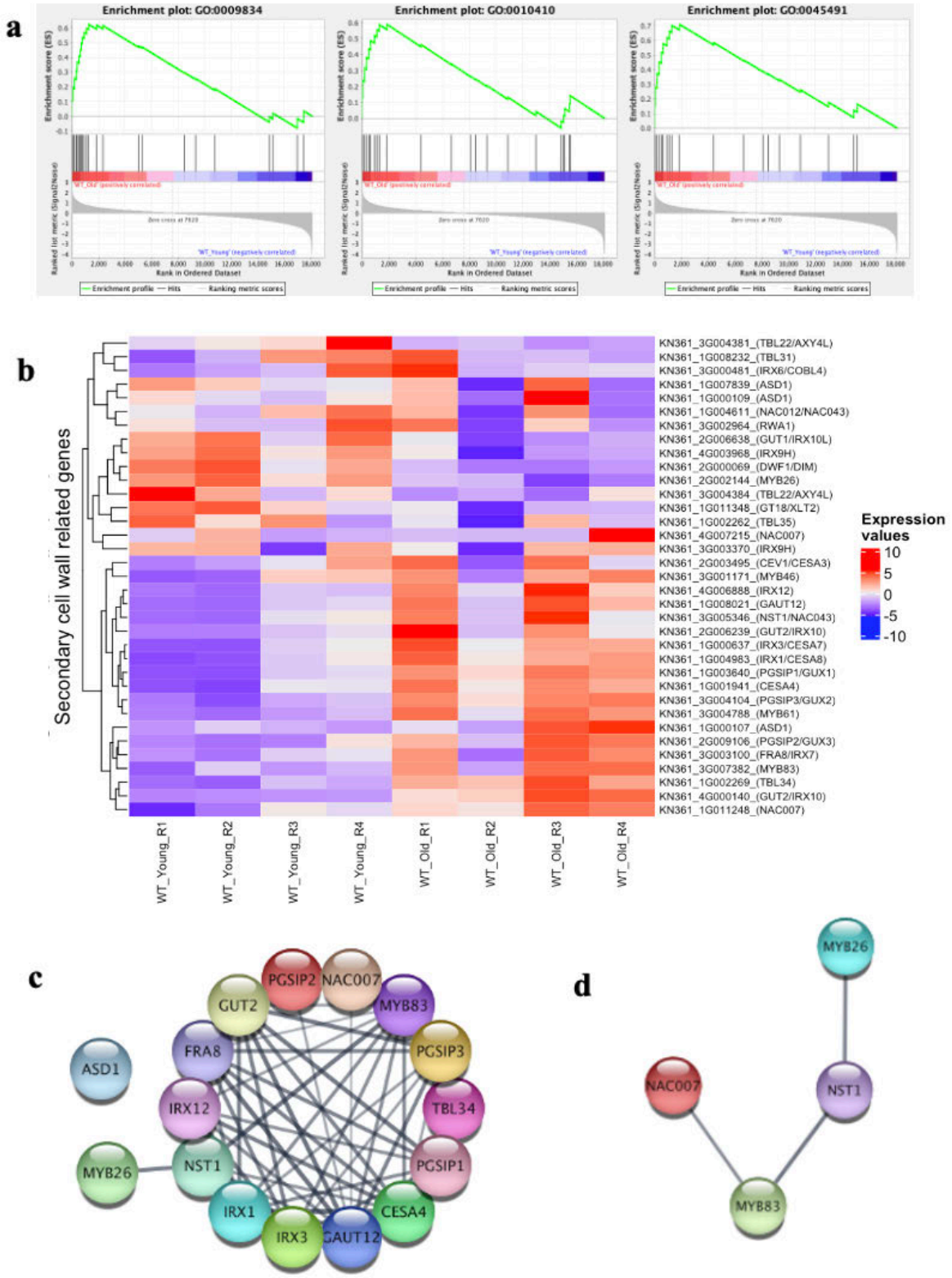


Figure 11. Enriched genes related to secondary cell walls in older capsules of WT (Comparison I, 'wt_youngvsold'). a. Three enrichment plots for GO:0009834 (BP plant-type secondary cell wall biogenesis), GO:0045491 (Biological Process (BP) xylan metabolic process), and GO:0010410 (BP hemicellulose metabolic process) were generated using GSEA v4.1.0; b. The

heat map shows the 35 non-redundant enriched genes in the plant cell wall cluster using ComplexHeatmap across eight samples. c. STRING analysis using protein names to retrieve protein-protein interaction networks from *A. thaliana*. d. Identification of four master regulators putatively controlling *P. ovata* capsule development.

Investigating lists of differentially expressed genes, present and absent genes, mutation location and co-expression in clusters compared between two genotypes

This study was carried out to identify genes or pathways associated with accelerated fruit ripening in *ace*. Figure 12a shows overlapping gene densities from three different analyses. Locations of present and absent, upregulated and downregulated DEGs, and putative highly mutated genes were distributed across all four chromosomes but were more concentrated at the ends than in the middle of each chromosome. Only one gene known to be associated with shattering, *SEEDSTICK* (*STK* or *AGL11*), was present or induced in the mutant samples (Figure 12b). Only two WT samples (8 DPA) in WT show *STK* expression (Figure 12b). However, no mutation was detected in the coding region of this gene.

As genes associated with oxidative phosphorylation pathways were enriched from the DEG list (Figures 10b & d), we investigated co-expressed gene lists from clustering analysis (Figure 9d). The WT has only one cluster of consistently co-expressed genes, while *ace* has two clusters. Upregulated gene clusters belonging to WT contain 9,671 genes whilst *ace* has 9,690 genes. Another *ace* cluster with 6,944 genes was downregulated (Figure 9d). All genes in the oxidative phosphorylation pathway were grouped in cluster 1 for WT, while for *ace* they were grouped into two clusters. The numbers of KEGG orthologs (KO) associated with the oxidative phosphorylation pathway (map00190) were 38 for WT cluster 1, 26 for *ace* cluster 1, and 33 for *ace* cluster 2. About 32 KO numbers were identified corresponding to putative mutated genes related to OXPHOS for *ace*. Overlapping gene changes in this pathway can be seen in Figure 13, and mutation types of the OXPHOS genes are listed in Table 3. Three putative

mutations were identified as highly possible: one mutation leads to gain of a premature stop codon (KN361_1g001751) and there were two frameshift variants (KN361_1g005227 and KN361_2g002464) (Table 3).

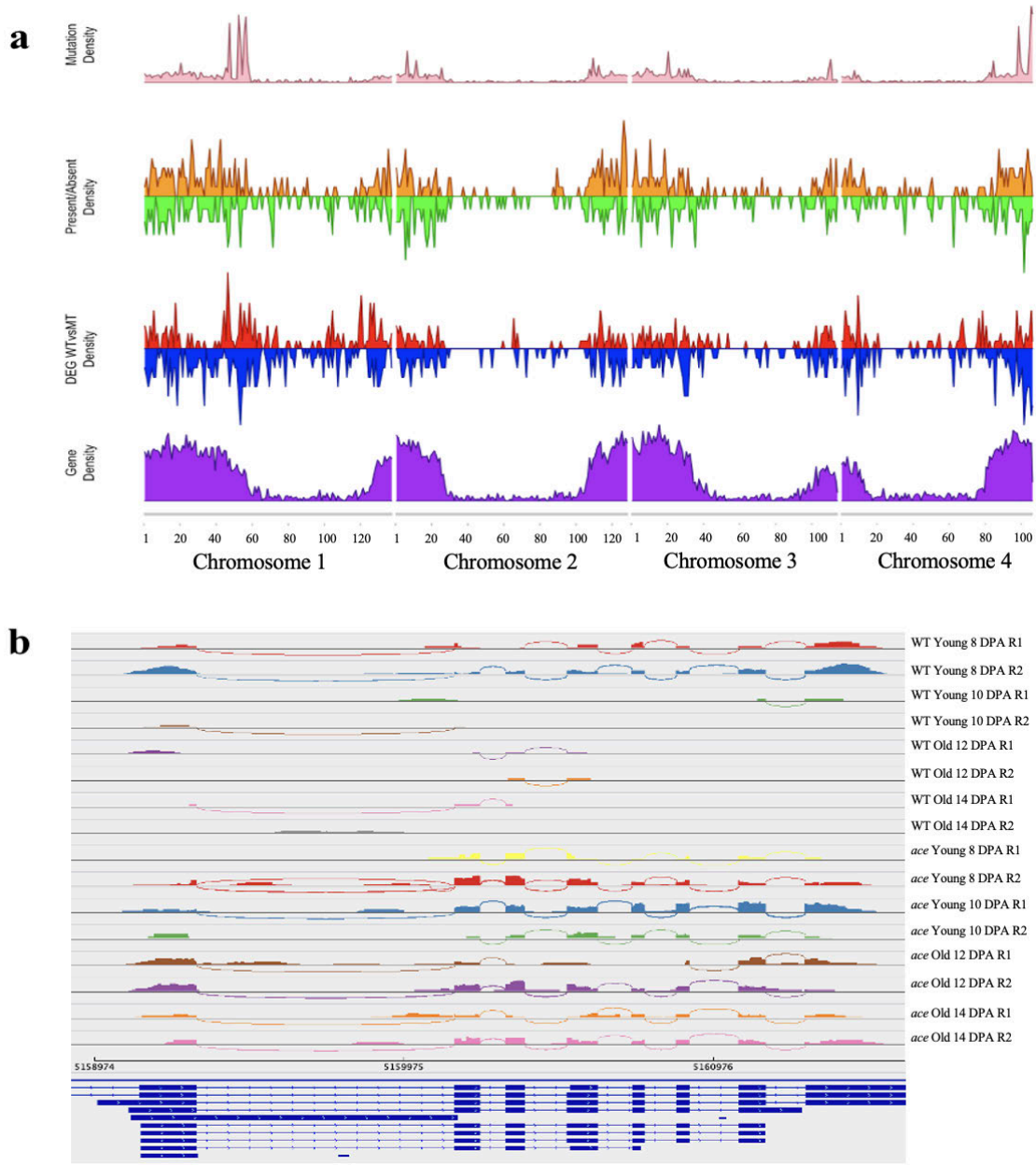


Figure 12. *P. ovata* gene densities (a) with coding genes (purple), downregulated DEGs (blue) and upregulated DEGs (red), transcripts absent in *ace* (green) and present only in *ace* but not in WT capsules (orange), and filtered mutated genes were visualised using KaryoploteR. (b). Expression of one candidate gene (*PoSTK*) associated with shattering. R1 = replicate 1 and R2 = replicate 2.

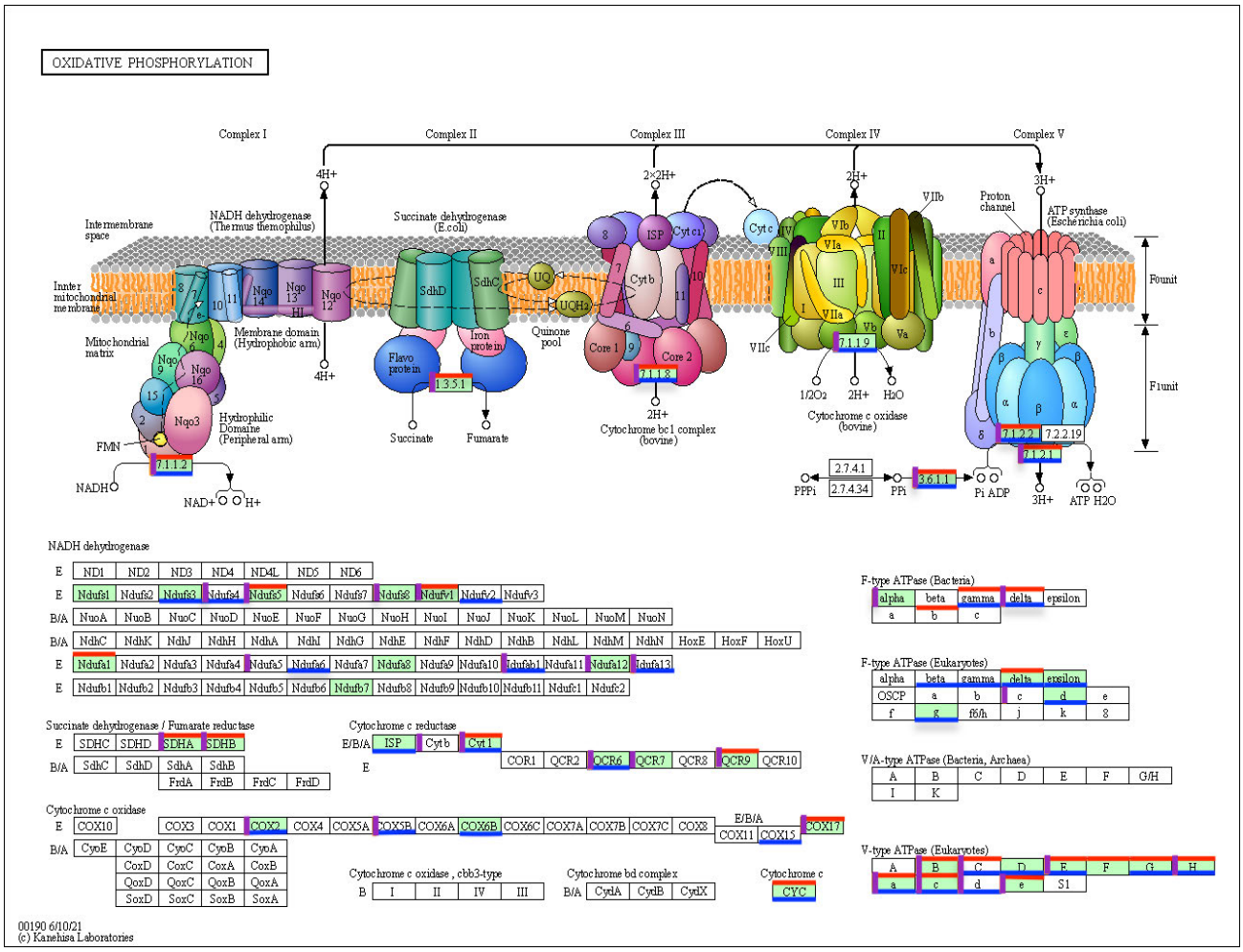


Figure 13. Oxidative phosphorylation (OXPHOS) pathway comparison between WT and *ace*. Background green on the pathway number and protein names indicate upregulated trend in older WT capsules while a red line on the top of the boxes is an upregulated trend in older *ace* capsules. A blue line under the name indicates downregulated genes in older *ace* capsules, while there were no downregulated genes in WT. Vertical purple lines before the names represents mutated genes.

Table 3. Information related to mutated genes in the OXPHOS pathway.

GeneID	Chr	Position (bp)	Ref	Alt	Type	Level	CW1 (Up)	CM 1 (Up)	CM2 (Down)
KN361_1g000774	Chr 1	5,109,371	A	G	missense variant	MR	-	Yes	-
KN361_1g001273	Chr 1	7,972,016	A	G	5' UTR variant	MF	Yes	Yes	-
KN361_1g001751	Chr 1	11,034,254	G	T	stop gained	HG	Yes	-	Yes
KN361_1g005227	Chr 1	34,345,747	AAC	A	frameshift variant	HG	-	-	Yes
KN361_1g005493	Chr 1	36,317,258	C	T	downstream gene variant	MF	-	-	Yes
KN361_1g006608	Chr 1	44,437,661	A	T	5' UTR variant	MF	Yes	Yes	-
KN361_1g006789	Chr 1	45,597,218	C	T	intron variant	MF	Yes	-	-
KN361_1g007824	Chr 1	54,674,642	T	C	intron variant	MF	Yes	Yes	-
KN361_1g010877	Chr 1	132,535,614	C	T	5' UTR variant	MF	-	-	Yes
KN361_2g000814	Chr 2	4,998,744	G	T	missense variant	MR	-	-	Yes
KN361_2g001055	Chr 2	6,246,276	C	T	synonymous variant	LW	Yes	Yes	-
KN361_2g001483	Chr 2	8,778,206	TG	T	upstream gene variant	MF	Yes	Yes	-
KN361_2g002253	Chr 2	14,012,680	T	C	intron variant	MF	-	-	Yes
KN361_2g002464	Chr 2	15,437,047	TG	T	frameshift variant	HG	-	-	Yes
KN361_2g002477	Chr 2	15,493,725	TATC	T	disruptive inframe deletion	MR	Yes	Yes	-
KN361_2g002851	Chr 2	18,198,845	A	G	3' UTR variant	MF	-	-	Yes
KN361_2g006538	Chr 2	110,289,430	C	T	3' UTR variant	MF	-	Yes	-
KN361_2g007976	Chr 2	119,763,744	A	G	3' UTR variant	MF	-	-	Yes
KN361_3g000304	Chr 3	1,910,854	C	G	upstream gene variant	MF	-	-	Yes
KN361_3g001569	Chr 3	9,632,988	A	G	upstream gene variant	MF	Yes	Yes	-
KN361_3g003199	Chr 3	19,337,209	T	A	3' UTR variant	MF	Yes	Yes	-
KN361_3g003361	Chr 3	20,304,407	A	G	3' UTR variant	MF	Yes	Yes	-
KN361_3g004267	Chr 3	26,450,276	T	A	5' UTR variant	MF	-	-	Yes
KN361_3g006497	Chr 3	66,809,018	A	G	intron variant	MF	-	-	Yes
KN361_3g007820	Chr 3	102,243,459	C	G	downstream gene variant	MF	-	Yes	-
KN361_4g000635	Chr 4	5,145,890	G	T	intron variant	MF	Yes	-	-
KN361_4g000770	Chr 4	6,503,532	T	C	upstream gene variant	MF	-	Yes	-
KN361_4g005313	Chr 4	92,916,015	T	C	upstream gene variant	MF	-	Yes	-
KN361_4g005480	Chr 4	93,884,398	GGTGA	G	splice donor, region & intron variants	HG	Yes	Yes	-
KN361_4g007389	Chr 4	105,116,364	A	G	5' UTR variant	MF	-	-	-
KN361_4g007594	Chr 4	106,282,251	A	G	upstream gene variant	MF	Yes	-	-
KN361_0g000020	Un	10,857	TGG	AAA	3' UTR variant	MF	Yes	-	Yes
KN361_0g000025	Un	419	A	C	3' UTR variant	MF	-	-	-
KN361_0g000026	Un	5,058	G	A	5' UTR variant	MF	-	-	-
KN361_0g000028	Un	29,258	GG	AA	5' UTR variant	MF	Yes	-	-
KN361_0g000564	Un	7,679	G	A	3' UTR variant	MF	Yes	-	Yes

OXPHOS= Oxidative phosphorylation; Ref= Reference; Alt=Alternative; CW=Cluster wild type; CM=Cluster mutant; MR = Moderate; MF = Modifier; HG = High; LW = Low

Discussion

This study presents a comprehensive profiling of morphological and transcriptomic changes during *P. ovata* capsule development that are relevant to seed shattering in this species.

Three abscission sites are likely to contribute to *P. ovata* seed shattering

Even though many plant species have different fruit types with variable shattering modes, they have one thing in common: developing abscission layers in specific sites that facilitate seed dispersal. Lamba and Gupta (1981) briefly documented cell structures in the region of dehiscence that start to develop in the undifferentiated zone in the middle of the *P. ovata* ovary (capsule) wall, providing line drawings. With the application of modern microscopy techniques much more detail has been provided here and it is clear that the capsule structure is quite complex.

The DZ divides the capsule into two valves: upper and lower (red arrows in Figures 1d and Figure 2) that have quite different compositions and structures. The layers in the lower valve were composed of elongated parenchyma cells, potentially stretched by the downward pressure of the developing seeds as the capsule expands, whose walls contain pectins and cellulose (Figure 6) but only xylan on the very outer epidermal surface (Figure 5). The base becomes highly wrinkled as the capsule ages (Figure 4a) and remains attached to the inflorescence after dehiscence. The upper valve was the operculum or lid and this was the more complicated half of the capsule, containing structures which potentially orchestrate dehiscence. The lid has more cell layers than the base (Figure 3) and these have defined architecture and composition. As well as containing pectins and cellulose there was strong labelling for xylan in the lid (Figures 5 and 6). This was seen on the outer surface, as for the base, but also along the thickened edge of the blocky cells that comprise a sclerenchyma layer (Figure 3). This was also the only cell layer in the capsule that was lignified (Figure 7) and this secondary thickening is likely to play a key role in capsule strength and dehiscence. The sclerenchyma layer stretches from the apex

of the operculum and joins onto the capsule base in the DZ in a structure that we have called the operculum hook (Figure 3) since it seems to curve around a circumscissile groove that runs around the capsule equator. For the first time we have some information about how dehiscence might occur for the *P. ovata* capsule. As the capsule ages the two layers of very small cells that sit at the join of the outer parenchyma cell layers, a bit like cartilage in a knee joint, and about the circumscissile groove, appear to separate where they join at the DZ (Figure 3 and Figure 4b). It is not clear how this abscission works, whether the cell walls undergo changes that facilitate clean detachment or the walls were ripped apart by physical pressure - this is an area that should be explored in the future using transmission electron microscopy (TEM) to examine the cell surfaces more closely and with immunolabelling to detect changes in polysaccharide composition, particularly of pectins. This event leaves the capsule partially open (Figure 4a, b and d) and full dehiscence does not occur until the end of the operculum hook detaches from the base and the lid comes off, leaving a muricate opercular surface on the lid (Figure 4c, e and f). The mechanism and forces that drive this final step in capsule dehiscence were also unclear at this stage and will be investigated using TEM, but it is clear that *P. ovata* dehiscence is a two-stage process rather than a simple separation event.

Lamba and Gupta (1981) suggested that pressure exerted by the fully formed seeds facilitates fruit dehiscence. However, images in Figure 1e and Supplementary Figure S2 show that the seed reduces in size during the ripening or drying process after 23 DPA so it is likely that pressure on the capsule wall by the seeds is not the only factor triggering capsule opening and seed dispersal, at least in the dry state. This may not be true when the seeds absorb moisture from the environment and the mucilage polysaccharides begin to rehydrate and swell (Cowley et al, 2022). Other factors may therefore contribute to shattering of dry capsules, such as unequal drying. By observing capsule development (Figure 1d and Figure 2), expansion of the two valves appears to occur at different speeds. The upper valve dries more slowly and was more pigmented than the lower valve. These differences were likely to create tension between

the two valves with separation triggered in response to external physical pressures where the capsule fractures along the weakest region, namely the layers of small cushioning cells at the DZ in the first stage. Then additional factors contribute to detachment of the operculum hook and when this happens the upper valve, the seeds, and the maternal disc all detach, leaving the lower valve still attached to the mother plant. Two other abscission sites were also involved in shattering (Figure 1g and h). One sits at the join between the maternal disk and the lower valve, which remained attached to the plant, and the other was the connection between the maternal disk and the seed surface where the developing seed was fed by the mother plant, forming a characteristic oval depression in the seed surface (hence names for this species include Isabgol or "horse's ear") visible after separation (Figure 1h). Although detachment of seeds from the maternal disc may lead them to push against the capsule from the inside it is more likely that the internal abscission zones were poised on a hair trigger, waiting for physical contact or drying to release, but at this stage we have no idea about timing or sequence of events. Deciphering the nature and sequence of cell separations at these points will shed even more light on the mechanics of capsule shattering and potentially can provide targets for reducing shattering severity. Currently, the final detachment of the opercular hook from the base would seem to be an obvious target for modification and it will be fascinating to further explore the key mechanisms governing *P. ovata* shattering.

The *accelerato* mutant

Ace was selected from the mutant collection because its capsules showed more rapid development than those of the WT. This included size and pigmentation and also earlier shattering relating to maturity stage. Morphological examination and immunolabelling studies (data not shown) indicated only capsule size differences but no structural differences to WT so it is likely that the shattering mechanism is also conserved, although this might require scrutiny at the TEM level to confirm. However, *ace* has served as a useful resource for comparative

genetic studies providing valuable information to narrow down a set of key genes involved in capsule development because of the difference in timing of expression.

Many Arabidopsis genes related to pod dehiscence zone (DZ) formation were not detected in the *P. ovata* capsule DEG list

The description of capsule shattering in *P. ovata* above makes it clear that this process is quite different, and involves novel cellular structures, from both silique dehiscence in Arabidopsis and pod shatter in legumes. Therefore, it was not surprising to find that many genes that control the development of DZ in Arabidopsis, namely *SHP1/2*, *RPL*, *FUL*, *IND*, and *ALC* were not present in the list of *P. ovata* DEGs between young and old capsules (1,722 genes, Comparison I, 'youngvsold') nor DEGs between WT and *ace* (1,045, Comparison IV, 'wtvsmt') (Figure 9c). They were also not on the lists from presence and absence expression analysis between WT and *ace* (Figure 9a). Dong et al. (2017) reported that they did not detect these genes in pod shattering-related gene sets in *Vicia sativa* L. either. There are a few possible reasons why we would not expect to see expression changes for these genes. The first is primarily that genes related to DZ cell fate were probably expressed at a very early stage in capsule development, even before fertilisation, so they were not transcribed from 8 to 14 DPA. Then of course, *P. ovata* may not have these genes or they were not expressed in the capsule, since different genes control shattering because there was no replum or other structures in this species similar to Arabidopsis (Chapter 2). Lastly, the *ace* mutation does not directly affect such genes, so there were no differences between WT and mutant.

Only *AP2*, *NST1*, and *ADPG1* were expressed in *P. ovata* capsules. Two copies of *AP2* are present in the *P. ovata* genome, and both are on chromosome 3 (KN361_3g000640 and KN361_3g005489) (Chapter 3). One copy (KN361_3g000640) was upregulated in both WT and *ace* (cluster 1 WT and mutant), while the other copy (KN361_3g005489) was present in *ace* cluster 2, which showed a downregulated trend (Figure 9d). *NST1* or *NAC domain*

containing 43 is also on chromosome 3 (KN361_3g005346). This gene showed upregulated expression in cluster 1 of WT and mutant. However, both *AP2* and *NST1* were not on the DEG list meaning their changes were insignificant. Only expression of *ADPG1* (KN361_4g004369) is upregulated in WT and mutant (cluster 1) and listed in the DEG lists. In WT, *PoADPG1* has a logFC of 2.7, while in the mutant it has 3.2 logFC. *ADPG1* is notable because it encodes an enzyme that hydrolyses cell wall pectin and is reported to be involved in cell separation in the final stages of pod and anther dehiscence (Ogawa et al., 2009). *PoADPG1* may contribute to cell separation in all or some of the abscission zones and in the final dehiscence of the mature *P. ovata* capsule when the operculum is released. *In situ* hybridisation to map transcript location and temporal expression patterns in WT and *ace* capsule sections are obvious future experiments to define the role of this gene product.

Secondary cell wall genes were enriched in older capsules

Secondary cell wall (SCW) genes were predominantly upregulated and enriched in older capsule tissue (Figure 11 and Table 2) and many of these relate to xylan which was shown to be a key polysaccharide component of the capsule cell walls in a number of locations (Figures 5 and 6). There were 7 SCW-specific hemicellulose biosynthetic genes corresponding to xylan biosynthesis: *PoIRX10*, *PoIRX7*, *PoGAUT12/IRX8*, *PoGUX2*, *PoGUX1*, *PoGUX3*, and *PoTBL34*, plus three SCW-specific cellulose synthase genes that classically work together in many plant species (Phan et al., 2016): *PoCEAS8/IRX1*, *PoCESA7/IRX3*, and *PoCESA4/IRX5*. Only one lignin biosynthetic gene is present in the enriched set, *laccase-22 PoLAC22/IRX12,* but this is likely to be significant since the sclerenchyma layer which ends in the operculum hook is notably lignified. Four master regulators were identified: *PoMYB26*, *PoNAC007/VND4* (*VASCULAR RELATED NAC-DOMAIN 4*), *PoNAC043/NST1*, and *PoMYB83* (Figure 11d).

In Arabidopsis, *MYB26* expression was mainly related to anther development and lignin biosynthesis (Yang et al., 2007). *AtMYB26* was strongly expressed in anthers but not in the

silique (pod) (Yang et al., 2007; Takahashi et al., 2020). Takahashi et al. (2020) found that *MYB26* was strongly expressed in seed pods in many legumes. They found that azuki bean (*Vigna angularis*) and yard-long bean (*Vigna unguiculata* cv-gr. Sesquipedalis) have become shattering resistant due to a truncation of the MYB26 protein. Di Vittori et al. (2021) reported that fine regulation of *PvMYB26* was responsible for pod dehiscence resistance in common beans.

MYB26 regulates *AtNST1*, *AtNST2*, and *AtNST3* (Zhao and Dixon, 2011; Nakano et al., 2015). In Arabidopsis, *AtNST1* and *AtNST3* redundantly regulate secondary wall synthesis in the stem, whereas *AtNST1* and *AtNST2* function redundantly to promote secondary wall formation in anthers (Zhao and Dixon, 2011; Nakano et al., 2015). However, the model legume *Medicago truncatula* appears to have a single *NST* gene, and *Mtnst1* loss-of-function mutants show secondary wall biosynthesis defects in both stem and anthers (Zhao et al., 2010). *P. ovata* has two *NST* genes, one copy (KN361_3g005346) identified as *NAC043/NST1*, KN361_1g004611 identified as *NAC012/NST3* and *NAC043/NST1*. This KN361_1g004611 gene has two protein products due to splice variants. Only KN361_3g005346 showed upregulated expression (Cluster 1 WT and mutant), making *PoNST1* a strong candidate to direct the cell wall thickening in *P. ovata* capsules and a clear gene target to explore in future experiments.

In Arabidopsis, transcription factor *MYB83* works together with *MYB46* to directly regulate cellulose, xylan and lignin genes (Zhong and Ye, 2012; Ko et al., 2014; Nakano et al., 2015). Their expression has been reported to be under *NST1* and *VND4* control. Double mutation of these genes shows reduced secondary wall thickening. No noticeable phenotype changes were seen in single *myb83* or *myb46* mutations, while the double mutation *myb46myb83* generated severe growth arrest (Ko et al., 2014). These may not be such strong candidates for further characterisation as *PoNST1*, but they should be added to the list.

Upregulated expression of the *SEEDSTICK* gene may accelerate shattering in the mutant

Another regulatory gene that may be important is *AGAMOUS-LIKE 11* (*AGL11*) or *SEEDSTICK* (*STK),* and this is the only agamous-like gene in the current gene list of the mutant (Supplementary Table S3). It was only expressed at 8 DPA in WT samples while high expression levels were seen in all mutant samples (Figure 12b). Huang et al. (2017) observed that *AGL11* in tomatoes increased its expression level in seeds and central tissues of young fruits, and it was mainly expressed during early fruit development. They reported that downregulated *AGL11* in tomatoes shows a limited effect, as a slight decrease in seed size, while overexpression of this gene led to dramatic modifications, including extreme softening before the onset of ripening. Pinyopich et al. (2003) found that *STK* works redundantly with *SHP1/2* and *AP2* in carpel (fruit) and ovule (seed) development. In Arabidopsis, expression of the *STK* gene was required for normal development of the funiculus or placenta that connects developing seeds to the fruit (Mizzotti et al., 2012). The loss-of-function of *stk* resulted in disruption of the abscission zone, so the seed failed to detach from the fruit (Pinyopich et al., 2003; Huang et al., 2017). As *AtSTK* has a role in seed detachment from fruit at the later stage of seed and fruit development in Arabidopsis, the *STK* gene may also be expressed at the different abscission sites in *P. ovata*. The role of the *AGL11/STK* gene product in cell wall modification, specifically involving pectic polysaccharides, will be interesting to define at the abscission and dehiscence zones, including at the operculum hook junction.

Mutation in oxidative phosphorylation genes may shed light on seed shattering genes

Respiration (oxidative phosphorylation) and photosynthesis were two pathways enriched in mutant capsules compared to WT capsules (Figure 10b), potentially linked to the more rapid development and maturation of the *ace* capsules. As the *P. ovata* capsules get old, they become pigmented and then brownish and it is likely that senescence pathways initiate, as indicated by photosynthesis capacity declining as the genes related to photosynthesis were downregulated,

especially in WT. However, in *ace*, not all expression of photosynthesis-related genes was downregulated in older capsules. Zhu et al. (2018) reported that pod photosynthetic capacity contributes to Arabidopsis seed yield. Higher and continued expression of *P. ovata* photosynthesis genes could also affect seed yield so that the mutant seed was bigger than the WT, but in this case the advantage seems to have flowed to the capsules which were significantly larger than the WT capsules throughout development (Figure 8). This trait may thus not be helpful for increasing seed yield directly, especially of the husk component.

There were no mutations detected in *PoADPG1*, *PoSTK* and or any of the photosynthetic genes (Figure 12). OXPHOS was the only pathway enriched in mutants regardless of capsule age (Figure 4b & d). Wang et al. (2017) found that downregulation of OXPHOS-related genes accelerated fruit ripening in strawberries by promoting the accumulation of sugar and plant hormones such as ABA, ethylene, and polyamine. Using this as a clue, mutations in the *ace* background were sought in genes included in the OXPHOS pathway. The *ace* line has consistently been selfed through to the M6 generation but has not yet been backcrossed to a WT parent since the crossing process was technically challenging for *P. ovata*. There is therefore a mutation background present in this line, potentially in quite a few genes, that needs to be taken into account. However, there were three genes with mutations, two with frameshift variations and one with a premature stop, that have a high chance of being disruptive (Table 3). There were two genes on Chromosome 1, KN361_1g001751 which is a cytochrome b-c1 complex subunit 6-1 and KN361_1g005227 which is an external alternative NADPH-ubiquinone oxidoreductase B2 and one on Chromosome 2 KN361_2g002464 that has homology to a V-type proton ATPase subunit E, with all three of the genes present in the downregulated cluster 2 for the mutant. Although these genes were not obviously connected to cell wall polysaccharide remodelling, they may play a role in an important pathway that has downstream influence on such processes. Such network interaction/s will only be revealed through future experimental work.

Conclusions

For the first time, detailed capsule development and the events that lead to seed shattering are described for *P. ovata*. The approaches consisted of morphological observation using different microscopy methods and transcriptomic analysis from two genotypes at two different stages. Within four weeks after anthesis, capsules and seeds undergo many changes divided into four stages: fruit set, fruit growth, maturation, and ripening. Cell wall polysaccharides were tracked through capsule development, revealing that xylan is a key component of some cell layers, particularly in the thickened walls of a sclerenchyma layer that joins the lid and base of the capsule. A set of cell wall genes associated with polysaccharide presence and enriched in the old capsule stage were identified. Events leading to shattering were described, seemingly a two-phase process with abscission occurring between two of the outer cell layers before the operculum hook detaches to facilitate full dehiscence. Mechanisms underlying these abscissions will be the target of future research.

The more rapid development of the capsule in the *ace* mutant is likely to be linked to alterations in the oxidative phosphorylation pathway, where potentially causal mutations were detected in a number of candidate genes. These include *AGL11/SEEDSTICK*, already known in Arabidopsis where it controls oil concentration in the seeds, which is a negative correlator of mucilage biosynthesis. Future work will validate the expression of the candidate genes identified from the RNAseq data using qPCR and *in situ* localisation of these transcripts in developing capsules. Using CRISPR/CAS9 and a newly developed transformation system for *P. ovata*, we can target these genes to downregulate their expression and study the effect on shattering. Investigation of such candidate genes associated with seed shattering in this study would benefit the molecular breeding program of not only *P. ovata* but also other plants, especially crops with dry dehiscent fruits, to increase yield.

References

Abu-Jamous B, Kelly SA-O (2018) Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data.

Andrews S (2017) FastQC: a quality control tool for high throughput sequence data. 2010.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30:** 2114-2120

Burton RA, Jobling S, Harvey A, Shirley N, Mather D, Bacic A, Fincher GB (2008) The genetics and transcriptional profiles of the cellulose synthase-like HvCslF gene family in barley. Plant physiology **146:** 1821-1833

Bushnell B (2014) BBMap: a fast, accurate, splice-aware aligner. *In*. Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States)

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J (2021) eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Molecular Biology and Evolution **38:** 5825-5829

Chen H, Boutros PC (2011) VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics **12:** 35

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly **6:** 80-92

Coudray A, Battenhouse AM, Bucher P, Iyer VR (2018) Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data. PeerJ **6:** e5362-e5362

Cowley JM, Burton RA (2021) The goo-d stuff: *Plantago* as a myxospermous model with modern utility. New Phytol **229:** 1917-1923

Cowley JM, Herliana L, Neumann KA, Ciani S, Cerne V, Burton RA (2020) A small-scale fractionation pipeline for rapid analysis of seed mucilage characteristics. Plant Methods **16:** 1-12

Di Vittori V, Bitocchi E, Rodriguez M, Alseekh S, Bellucci E, Nanni L, Gioia T, Marzario S, Logozzo G, Rossato M, De Quattro C, Murgia ML, Ferreira JJ, Campa A, Xu C, Fiorani F, Sampathkumar A, Fröhlich A, Attene G, Delledonne M, Usadel B, Fernie AR, Rau D, Papa R (2021) Pod indehiscence in common bean is associated with the fine regulation of PvMYB26. Journal of Experimental Botany **72:** 1617-1633

Di Vittori V, Gioia T, Rodriguez M, Bellucci E, Bitocchi E, Nanni L, Attene G, Rau D, Papa R (2019) Convergent Evolution of the Seed Shattering Trait. Genes **10**

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29:** 15-21

Dong R, Dong D, Luo D, Zhou Q, Chai X, Zhang J, Xie W, Liu W, Dong Y, Wang Y, Liu Z (2017) Transcriptome Analyses Reveal Candidate Pod Shattering-Associated Genes Involved in the Pod Ventral Sutures of Common Vetch (*Vicia sativa* L.). Frontiers in Plant Science **8:** 649

Dong Y, Wang Y-Z (2015) Seed shattering: from models to crops. Frontiers in Plant Science **6:** 476

Dong Y, Yang X, Liu J, Wang B-H, Liu B-L, Wang Y-Z (2014) Pod shattering resistance associated with domestication is mediated by a NAC gene in soybean. Nature Communications **5:** 3352

Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics **32:** 3047-3048

Ferrándiz C (2002) Regulation of fruit dehiscence in Arabidopsis. Journal of Experimental Botany **53:** 2031-2038

Gu Q, Ferrandiz C, Yanofsky MF, Martienssen R (1998) The FRUITFULL MADS-box gene mediates cell differentiation during Arabidopsis fruit development. Development **125:** 1509-1517

Govt India Dept of Commerce (2021) 'Psyllium seed (isobgul) 12119013 Export:Commodity-wise.

Govt India Dept of Commerce (2022) 'Psyllium seed (isobgul) 12119013 Export:Commodity-wise.

Huang B, Routaboul J-M, Liu M, Deng W, Maza E, Mila I, Hu G, Zouine M, Frasse P, Vrebalov JT, Giovannoni JJ, Li Z, van der Rest B, Bouzayen M (2017) Overexpression of the class D MADS-box gene Sl-AGL11 impacts fleshy tissue differentiation and structure in tomato fruits. Journal of Experimental Botany **68:** 4869-4884

Kassambara A, Mundt F (2017) Factoextra R Package: Easy Multivariate Data Analyses and Elegant Visualization.

Ko JH, Jeon HW, Kim WC, Kim JY, Han KH (2014) The MYB46/MYB83-mediated transcriptional regulatory programme is a gatekeeper of secondary wall biosynthesis. Annals of Botany **114:** 1099-1107

Konishi S, Izawa T, Lin Shao Y, Ebana K, Fukuta Y, Sasaki T, Yano M (2006) An SNP Caused Loss of Seed Shattering During Rice Domestication. Science **312:** 1392-1396

Kuai J, Sun Y, Liu T, Zhang P, Zhou M, Wu J, Zhou G (2016) Physiological Mechanisms behind Differences in Pod Shattering Resistance in Rapeseed (Brassica napus L.) Varieties. PLOS ONE **11:** e0157341

Lamba LC and Gupta V (1981) Anatomy of circumscissile dehiscence in *Plantago ovata* Forsk. Current science (Bangalore) **50:** 541-543

Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H (2014) UpSet: Visualization of Intersecting Sets. IEEE Transactions on Visualization and Computer Graphics **20:** 1983-1992

Li C, Zhou A, Sang T (2006) Rice domestication by reducing shattering. Science **311:** 1936-1939

Li F, Komatsu A, Ohtake M, Eun H, Shimizu A, Kato H (2020) Direct identification of a mutation in OsSh1 causing non-shattering in a rice (*Oryza sativa* L.) mutant cultivar using whole-genome resequencing. Scientific Reports **10:** 14936

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:** 2078-2079

Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30:** 923-930

Liljegren SJ, Roeder AHK, Kempin SA, Gremski K, Østergaard L, Guimil S, Reyes DK, Yanofsky MF (2004) Control of fruit patterning in Arabidopsis by INDEHISCENT. Cell **116:** 843-853

Lin Z, Griffith Me Fau - Li X, Li X Fau - Zhu Z, Zhu Z Fau - Tan L, Tan L Fau - Fu Y, Fu Y Fau - Zhang W, Zhang W Fau - Wang X, Wang X Fau - Xie D, Xie D Fau - Sun C, Sun C (2007) Origin of seed shattering in rice (*Oryza sativa* L.).

Lv S, Wu W, Wang MA-O, Meyer RS, Ndjiondjop MN, Tan LA-O, Zhou H, Zhang JA-O, Fu Y, Cai HA-O, Sun C, Wing RA, Zhu ZA-O (2018) Genetic control of seed shattering during African rice domestication.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research **20:** 1297-1303

Mitsuda N, Ohme-Takagi M (2008) NAC transcription factors NST1 and NST3 regulate pod shattering in a partially redundant manner by promoting secondary wall formation after the establishment of tissue identity.

Mizzotti C, Mendes MA, Caporali E, Schnittger A, Kater MM, Battaglia R, Colombo L (2012) The MADS box genes SEEDSTICK and ARABIDOPSIS Bsister play a maternal role in fertilization and seed development. The Plant Journal **70:** 409-420

Nakano Y, Yamaguchi M, Endo H, Rejab NA, Ohtani M (2015) NAC-MYB-based transcriptional regulation of secondary cell wall biosynthesis in land plants. Frontiers in Plant Science **6:** 288

Ogawa M, Kay P, Wilson S, Swain SM (2009) ARABIDOPSIS DEHISCENCE ZONE POLYGALACTURONASE1 (ADPG1), ADPG2, and QUARTET2 are Polygalacturonases required for cell separation during reproductive development in Arabidopsis. The Plant cell **21:** 216-233

Olsen KM (2012) One gene's shattering effects. Nature Genetics **44:** 616-617

Pabón-Mora N, Wong GK-S, Ambrose BA (2014) Evolution of fruit development genes in flowering plants. Frontiers in plant science **5:** 300-300

Patel D, Patel H, Patel P, Patel H, Amin A (2018) Evaluation of stable and non shattering isabgol cultivar-Gujarat isabgol 4. JOSAC**:** 88-90

Phan JL, Cowley JM, Neumann KA, Herliana L, O'Donovan LA, Burton RA (2020) The novel features of *Plantago ovata* seed mucilage accumulation, storage and release. Sci Rep **10:** 1-14

Pinyopich A, Ditta GS, Savidge B, Liljegren SJ, Baumann E, Wisman E, Yanofsky MF (2003) Assessing the redundancy of MADS-box genes during carpel and ovule development. Nature **424:** 85-88

RACP (2016) Value Chain Analysis: Isabgol.

Rajani S, Sundaresan V (2001) The Arabidopsis myc/bHLH gene ALCATRAZ enables cell separation in fruit dehiscence. Current Biology **11:** 1914-1922

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research **43:** e47-e47

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England) **26:** 139-140

Roeder AHK, Ferrándiz C, Yanofsky MF (2003) The role of the REPLUMLESS homeodomain protein in patterning the Arabidopsis fruit. Current Biology **13:** 1630-1635

Ruel K, Nishiyama Y, Joseleau JP (2012) Crystalline and amorphous cellulose in the secondary walls of Arabidopsis. Plant Science **193–194:** 48-61

Ruprecht C, Bartetzko MP, Senf D, Dallabernadina P, Boos I, Andersen MC, Kotake T, Knox JP, Hahn MG, Clausen MH, Pfrengle F (2017) A synthetic glycan microarray enables epitope mapping of plant cell wall glycan-directed antibodies. Plant physiology. **175**(3): 1094-104

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research **13:** 2498-2504

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences **102:** 15545

Takahashi Y, Kongjaimun A, Muto C, Kobayashi Y, Kumagai M, Sakai H, Satou K, Teruya K, Shiroma A, Shimoji M, Hirano T, Isemura T, Saito H, Baba-Kasai A, Kaga A, Somta P, Tomooka N, Naito K (2020) Same locus for non-shattering seed pod in two independently domesticated legumes, *Vigna angularis* and *Vigna unguiculata*. Frontiers in Genetics **11:** 748

Taylor-Teeples M, Lin L, de Lucas M, Turco G, Toal TW, Gaudinier A, Young NF, Trabucco GM, Veling MT, Lamothe R, Handakumbura PP, Xiong G, Wang C, Corwin J, Tsoukalas A, Zhang L, Ware D, Pauly M, Kliebenstein DJ, Dehesh K,

Tagkopoulos I, Breton G, Pruneda-Paz JL, Ahnert SE, Kay SA, Hazen SP, Brady SM (2015) An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature **517:** 571-575

Tucker M, Ma C, Phan J, Neumann K, Shirley N, Hahn M, Cozzolino D, Burton RA (2017) Dissecting the genetic basis for seed coat mucilage heteroxylan biosynthesis in *Plantago ovata* using gamma irradiation and infrared spectroscopy. Frontiers in Plant Science **8**

Verhertbruggen Y, Marcus SE, Haeger A, Ordaz-Ortiz JJ, Knox JP (2009) An extended set of monoclonal antibodies to pectic homogalacturonan. Carbohydrate Research **344**: 1858-1862

Wang Q-H, Zhao C, Zhang M, Li Y-Z, Shen Y-Y, Guo J-X (2017) Transcriptome analysis around the onset of strawberry fruit ripening uncovers an important role of oxidative phosphorylation in ripening. Scientific reports **7:** 41477-41477

Yang C, Xu Z, Song J, Conner K, Vizcay Barrena G, Wilson ZA (2007) Arabidopsis MYB26/MALE STERILE35 Regulates secondary thickening in the endothecium and is essential for anther dehiscence. The Plant Cell **19:** 534-548

Yoon J, Cho L-H, Antt HW, Koh H-J, An G (2017) KNOX Protein OSH15 induces grain shattering by repressing lignin biosynthesis genes. Plant physiology **174:** 312-325

Yoon J, Cho Lh Fau - Kim SL, Kim Sl Fau - Choi H, Choi H Fau - Koh H-J, Koh Hj Fau - An G, An G (2014) The BEL1-type homeobox gene SH5 induces seed shattering by enhancing abscission-zone development and inhibiting lignin biosynthesis.

Young MD, Wakefield MJ, Smyth GK, Oshlack A (2012) goseq: Gene Ontology testing for RNA-seq datasets. R Bioconductor **8:** 1-25

Zhao Q, Dixon RA (2011) Transcriptional networks for lignin biosynthesis: more complex than we thought? Trends in Plant Science **16:** 227-233

Zhao Q, Gallego-Giraldo L, Wang H, Zeng Y, Ding S-Y, Chen F, Dixon RA (2010) An NAC transcription factor orchestrates multiple features of cell wall development in Medicago truncatula. The Plant Journal **63:** 100-114

Zhong R, Ye Z-H (2012) MYB46 and MYB83 Bind to the SMRE Sites and Directly Activate a Suite of Transcription Factors and Secondary Wall Biosynthetic Genes. Plant and Cell Physiology **53:** 368-380

Zhou Y, Lu D, Li C, Luo J, Zhu B-F, Zhu J, Shangguan Y, Wang Z, Sang T, Zhou B, Han B (2012) Genetic control of seed shattering in rice by the APETALA2 transcription factor shattering abortion1. The Plant cell **24:** 1034-1048

Zhu X, Zhang L, Kuang C, Guo Y, Huang C, Deng L, Sun X, Zhan G, Hu Z, Wang H, Hua W (2018) Important photosynthetic contribution of silique wall to seed yield-related traits in Arabidopsis thaliana.

Acknowledgements

The authors thank Dr Fabien Voisin for his technical support in using the Phoenix-HPC. We acknowledge Dr Gwen Mayo for her assistance with microscopy, Melissa Pickering for plant care, and Amanda Philpot for her attempt to collect data. This study was supported by the Australian Research Council (ARC) Centres of Excellence in Plant Cell Walls (CE110001007) and Plant Energy Biology (CE140100008) and an ARC Linkage Grant (LP180100971). This work was also supported with supercomputing resources provided by the Phoenix HPC service, Undercroft Glasshouse University of Adelaide, and Adelaide Microscopy at the University of Adelaide. The RNA sequencing was performed at the Flinders Genomics Facility. LH is supported by the University of Adelaide's Adelaide Graduate Research Scholarship (AGRS) and The National Research and Innovation Agency (BRIN-Indonesia).

Authors' information

Affiliations

School of Agriculture, Food and Wine, Waite Research Institute, University of Adelaide, Waite Campus, Urrbrae, SA, Australia

Lina Herliana, James M. Cowley, Lisa A. O'Donovan, Tycho R. Neumann, Shi Fang Khor, Tina Bianco-Miotto & Rachel A. Burton

South Australian Genomics Centre (SAGC), SA, Australia

Nathan S. Watson-Haigh

Australian Genome Research Facility, Victorian Comprehensive Cancer Centre, Melbourne, VIC 3000, Australia

Nathan S. Watson-Haigh

IP Australia, PO Box 200, Woden, ACT, 2606, Australia

Tycho R. Neumann

Contributions

R.A.B. conceived the project. R.A.B., N.S.W., and T.B. supervised the study and revised the manuscript. L.H. and N.S.W. developed workflows. L.H. staged and harvested fresh WT seeds and capsules, processed and analysed the data and wrote the draft manuscript. T.R.N., and S.F.K collected materials and performed RNA extraction. J.M.C performed observation on fresh WT and mutant capsules. L.A.O. performed microscopy work on fixed material. All authors read, edited, and approved the manuscript.

Corresponding authors

Correspondence to Rachel A. Burton

Additional information

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare no competing interests.

Supplementary material

Supplementary Table S1. Non-coding RNA genes transcribed in WT capsules but not in *ace* tissues.

Gene id	Chr	Start	End	Protein product
KN361_nc011	Chr 1	54490123	54493149	long non-coding RNA
KN361_nc017	Chr 1	112074982	112083609	long non-coding RNA
KN361_nc018	Chr 1	112894943	112903975	long non-coding RNA
KN361_nc019	Chr 1	118472553	118481440	long non-coding RNA
KN361_nc071	Chr 1	78010936	78013200	long non-coding RNA
KN361_nc102	Chr 1	83963081	83972147	long non-coding RNA
KN361_nc110	Chr 1	95451714	95461378	long non-coding RNA
KN361_nc118	Chr 1	102540835	102542520	long non-coding RNA
KN361_nc184	Chr 1	7988229	7988979	long non-coding RNA
KN361_nc014	Chr 1	62699088	62699213	small nuclear RNA U1
KN361_nc329	Chr 2	21825936	21827671	long non-coding RNA
KN361_nc370	Chr 2	89137909	89139351	long non-coding RNA
KN361_nc457	Chr 3	45590631	45596756	long non-coding RNA
KN361_nc458	Chr 3	13117498	13118709	long non-coding RNA
KN361_nc465	Chr 3	95345899	95351221	long non-coding RNA
KN361_nc474	Chr 3	98590868	98593141	long non-coding RNA
KN361_nc478	Chr 3	57384738	57386640	long non-coding RNA
KN361_nc508	Chr 3	64710965	64711827	long non-coding RNA
KN361_nc595	Chr 4	61258105	61259724	long non-coding RNA
KN361_nc613	Chr 4	42660701	42669239	long non-coding RNA
KN361_nc670	Chr 4	54232058	54240414	long non-coding RNA
KN361_nc660	Chr 4	101755663	101755776	small nucleolar RNA SNORD14
KN361_nc608	Chr 4	104708221	104708339	small nucleolar RNA U3
KN361_nc749	Un	11774	16791	long non-coding RNA
KN361_nc753	Un	2332	4797	long non-coding RNA

Supplementary Table S2. Non-coding RNA genes transcribed in *ace* capsules but not in wild-type tissues.

Gene id	Chr	Start	End	Protein product
KN361_nc022	Chr 1	5755021	5755237	small nucleolar RNA U3
KN361_nc029	Chr 1	17762285	17762419	small nuclear RNA U1
KN361_nc115	Chr 1	104432399	104434651	long non-coding RNA
KN361_nc156	Chr 1	90298326	90308172	long non-coding RNA
KN361_nc163	Chr 1	57439487	57441037	long non-coding RNA
KN361_nc364	Chr 2	126503976	126505389	long non-coding RNA
KN361_nc582	Chr 4	2195476	2195606	small nuclear RNA U1
KN361_nc583	Chr 4	2202780	2202961	small nuclear RNA U2
KN361_nc638	Chr 4	64872509	64874659	long non-coding RNA
KN361_nc655	Chr 4	50499632	50501724	long non-coding RNA
KN361_nc730	Un	21	999	long non-coding RNA

Supplementary Table S3. A list of expressed genes specifically in *ace* capsules at 8-14 DPA.

Gene id	Chr	Start	End	Protein product
KN361_1g000043	Chr 1	321909	325137	putative auxin efflux carrier component 1c;putative auxin efflux carrier component 1b
KN361_1g000248	Chr 1	1750535	1753729	protein QUIRKY
KN361_1g000266	Chr 1	1919503	1920866	beta-glucosidase 47
KN361_1g000456	Chr 1	3108327	3115770	protein IWS1 1
KN361_1g000487	Chr 1	3321861	3322984	F-box/kelch-repeat-containing protein
KN361_1g000572	Chr 1	3878157	3878733	F-box protein
KN361_1g000630	Chr 1	4199500	4208740	hypothetical protein
KN361_1g000658	Chr 1	4360251	4361175	hypothetical protein
KN361_1g000667	Chr 1	4419161	4420321	early nodulin-like protein 1

KN361_1g000729	Chr 1	4860159	4863933	serine carboxypeptidase-like 34
KN361_1g000750	Chr 1	4975840	4976647	pEARLI1-like lipid transfer protein 2
KN361_1g000807	Chr 1	5359751	5360507	hypothetical protein
KN361_1g000839	Chr 1	5511139	5511853	hypothetical protein
KN361_1g000855	Chr 1	5614968	5616758	hypothetical protein
KN361_1g000939	Chr 1	5985935	5988523	pentatricopeptide repeat-containing protein
KN361_1g001017	Chr 1	6495218	6496734	E3 ubiquitin-protein ligase SINA-like 7
KN361_1g001018	Chr 1	6495983	6499358	E3 ubiquitin-protein ligase SINA-like 7
KN361_1g001020	Chr 1	6507787	6509813	transcription factor MYB101
KN361_1g001046	Chr 1	6651118	6651771	protein SPH29
KN361_1g001101	Chr 1	7053881	7055868	helicase
KN361_1g001187	Chr 1	7462496	7463040	hypothetical protein
KN361_1g001308	Chr 1	8122267	8123006	auxin-responsive protein SAUR24
KN361_1g001337	Chr 1	8233829	8235167	hypothetical protein
KN361_1g001577	Chr 1	9910526	9913065	NOTCH family protein
KN361_1g001585	Chr 1	9968301	9970360	protein STICHEL
KN361_1g001671	Chr 1	10503597	10504753	WEB family protein
KN361_1g001682	Chr 1	10599138	10600553	serine/threonine-protein kinase-like protein
KN361_1g001697	Chr 1	10673963	10676022	MLO-like protein 4
KN361_1g001720	Chr 1	10813046	10814515	3-ketoacyl-CoA synthase 7
KN361_1g001782	Chr 1	11185837	11186019	hypothetical protein
KN361_1g001818	Chr 1	11392862	11393434	S locus-related glyco P
KN361_1g001864	Chr 1	11656027	11658899	putative inorganic phosphate transporter 1-7
KN361_1g001949	Chr 1	12196512	12198003	putative F-box protein
KN361_1g002033	Chr 1	12716679	12719942	putative LRR receptor-like serine/threonine-protein kinase;hypothetical protein
KN361_1g002071	Chr 1	12931897	12937240	putative late blight resistance protein R1B-16
KN361_1g002079	Chr 1	13000343	13001532	zinc-finger homeodomain-containing protein 5

KN361_1g002102	Chr 1	13160408	13161395	bZIP transcription factor 44
KN361_1g002193	Chr 1	13735315	13736903	hypothetical protein
KN361_1g002304	Chr 1	14358576	14360198	F-box/WD-40 repeat-containing protein 1
KN361_1g002318	Chr 1	14448085	14448876	hypothetical protein
KN361_1g002484	Chr 1	15521881	15524745	fringe-like;hypothetical protein
KN361_1g002494	Chr 1	15567729	15569814	L-type lectin-domain-containing protein containing receptor kinase S.4
KN361_1g002503	Chr 1	15612854	15613431	EB module domain-containing protein
KN361_1g002599	Chr 1	16316929	16319836	heat shock cognate protein 70
KN361_1g002604	Chr 1	16358732	16360018	RING-H2 finger protein ATL29
KN361_1g003009	Chr 1	19123217	19134606	hypothetical protein
KN361_1g003022	Chr 1	19199189	19201204	protein IQ-domain-containing protein IQD20;hypothetical protein
KN361_1g003042	Chr 1	19351811	19355946	2-oxoglutarate-dependent dioxygenase 33
KN361_1g003202	Chr 1	20525892	20527520	calcium-dependent protein kinase 20
KN361_1g003212	Chr 1	20582959	20587420	cytokinin riboside 5'-monophosphate phosphoribohydrolase LOG7
KN361_1g003277	Chr 1	21075547	21078293	hypothetical protein
KN361_1g003293	Chr 1	21160757	21163726	glycine-rich protein A3
KN361_1g003332	Chr 1	21403269	21404655	TATA-box-binding protein
KN361_1g003399	Chr 1	21872608	21873746	protein kirola
KN361_1g003471	Chr 1	22367225	22367960	Cotton fiber expressed protein
KN361_1g003472	Chr 1	22368380	22370588	putative pectinesterase inhibitor 12
KN361_1g003566	Chr 1	23086110	23086880	hypothetical protein
KN361_1g003698	Chr 1	23980501	23981005	EB module domain-containing protein
KN361_1g004009	Chr 1	25908842	25911409	hypothetical protein
KN361_1g004030	Chr 1	25996675	25998540	hypothetical protein
KN361_1g004059	Chr 1	26186005	26187212	hypothetical protein
KN361_1g004065	Chr 1	26236278	26237721	PDDEXK-like protein;oxidoreductase molybdopterin binding domain-containing protein
KN361_1g004081	Chr 1	26373785	26375511	zinc finger protein CCCH domain-containing protein 2

KN361_1g004086	Chr 1	26455222	26457534	reticulon-like protein B18;reticulon-like protein B17
KN361_1g004114	Chr 1	26669642	26672508	beta-D-xylosidase 1
KN361_1g004151	Chr 1	26922687	26923676	hypothetical protein
KN361_1g004259	Chr 1	27735894	27737288	ethylene-responsive transcription factor ERF003
KN361_1g004266	Chr 1	27784005	27785180	zinc finger protein 3
KN361_1g004285	Chr 1	27937088	27939517	flotillin-like protein 3;flotillin-like protein 1
KN361_1g004287	Chr 1	27943679	27946091	flotillin-like protein 1;flotillin-like protein 3
KN361_1g004407	Chr 1	28791779	28796459	primary amine oxidase
KN361_1g004455	Chr 1	29120500	29122323	UDP-glycosyltransferase 73C3
KN361_1g004553	Chr 1	29710407	29714114	hypothetical protein;ZIP zinc transporter
KN361_1g004666	Chr 1	30453450	30457322	transcription factor LHW
KN361_1g004772	Chr 1	31270274	31272423	xylan arabinosyltransferase 3 XAT2
KN361_1g004780	Chr 1	31300682	31302483	FBD domain-containing protein;F-box protein
KN361_1g004945	Chr 1	32477987	32479456	transcription termination factor MTERF8
KN361_1g005184	Chr 1	34068811	34070739	hypothetical protein
KN361_1g005270	Chr 1	34714675	34715749	hypothetical protein
KN361_1g005296	Chr 1	34922296	34935602	gibberellin 2-beta-dioxygenase 8
KN361_1g005366	Chr 1	35432979	35434328	putative F-box protein
KN361_1g005398	Chr 1	35724267	35725204	F-box/kelch-repeat-containing protein
KN361_1g005407	Chr 1	35788602	35789435	non-specific lipid-transfer protein
KN361_1g005447	Chr 1	36022142	36022689	cystein proteinase inhibitor 5
KN361_1g005463	Chr 1	36135689	36137797	putative pumilio 8
KN361_1g005497	Chr 1	36340509	36344530	nuclear transcription factor Y subunit B-6
KN361_1g005537	Chr 1	36579561	36581885	cytochrome P450
KN361_1g005555	Chr 1	36667986	36670377	serine/threonine-protein kinase AGC1-5
KN361_1g005614	Chr 1	37222169	37222663	hypothetical protein
KN361_1g005799	Chr 1	38401447	38403378	hypothetical protein

KN361_1g005812	Chr 1	38481944	38483647	Mif2/CENP-C like;hypothetical protein
KN361_1g005846	Chr 1	38722598	38724851	hypothetical protein;laminin domain-containing protein II
KN361_1g005924	Chr 1	39273793	39284444	Delta(12)-fatty-acid desaturase FAD2
KN361_1g006062	Chr 1	40490584	40492044	3-beta-glucosidase
KN361_1g006163	Chr 1	41216492	41218321	hypothetical protein
KN361_1g006214	Chr 1	41537759	41543688	bZIP transcription factor TGA10
KN361_1g006219	Chr 1	41565647	41567872	Ras-related protein RABA4d
KN361_1g006234	Chr 1	41734881	41736864	hypothetical protein
KN361_1g006276	Chr 1	42008898	42010541	non-specific lipid-transfer protein 3
KN361_1g006307	Chr 1	42205162	42206344	hypothetical protein
KN361_1g006346	Chr 1	42493251	42494340	hypothetical protein
KN361_1g006378	Chr 1	42742109	42743322	transposase
KN361_1g006386	Chr 1	42780083	42782402	beta-amyrin 28-monooxygenase
KN361_1g006388	Chr 1	42786862	42789325	protein kinesin light chain-related KLCR2
KN361_1g006545	Chr 1	44022873	44024050	Snakin-2
KN361_1g006662	Chr 1	44762711	44765913	putative late blight resistance protein R1B-11
KN361_1g006703	Chr 1	44956228	44956904	auxin-responsive protein SAUR23
KN361_1g006704	Chr 1	44959560	44962458	putative late blight resistance protein R1A-6
KN361_1g006756	Chr 1	45322016	45323164	26S proteasome regulatory subunit S10B-B
KN361_1g006809	Chr 1	45697181	45699692	retrotransposon gag protein
KN361_1g006964	Chr 1	46635209	46635613	ribosome-inactivating protein PMRIPm
KN361_1g006967	Chr 1	46685769	46690615	inositol-tetrakisphosphate 1-kinase 2;inositol-tetrakisphosphate 1-kinase 3
KN361_1g006973	Chr 1	46799559	46805629	zinc-binding in reverse transcriptase
KN361_1g006975	Chr 1	46808870	46813456	jasmonate-induced protein 60
KN361_1g006978	Chr 1	46855367	46856549	receptor-like protein kinase 2
KN361_1g006979	Chr 1	46857175	46861275	jasmonate-induced protein 60
KN361_1g006985	Chr 1	46910544	46918575	putative glucomannan 4-beta-mannosyltransferase 14;hypothetical protein

KN361_1g006991	Chr 1	46982514	46985256	protein phosphatase 2A regulatory B subunit B56 family protein
KN361_1g006997	Chr 1	47020979	47028272	hypothetical protein;GRF zinc finger protein
KN361_1g006999	Chr 1	47029511	47029976	hypothetical protein
KN361_1g007093	Chr 1	47788076	47790123	membrane steroid-binding protein 1
KN361_1g007101	Chr 1	47863369	47865349	transcription factor bHLH25;transcription factor bHLH18
KN361_1g007131	Chr 1	48180264	48181759	hypothetical protein
KN361_1g007281	Chr 1	49535076	49536670	TSC-22/dip/bun family protein
KN361_1g007293	Chr 1	49653709	49657406	retrotransposon gag protein;hypothetical protein
KN361_1g007362	Chr 1	50243042	50244928	classical arabinogalactan protein 26
KN361_1g007363	Chr 1	50244248	50246049	protein NETWORKED 3A
KN361_1g007384	Chr 1	50455732	50456598	monothiol glutaredoxin-S2
KN361_1g007424	Chr 1	50816222	50817805	hypothetical protein
KN361_1g007621	Chr 1	52702615	52704087	putative bifunctional methylthioribulose-1-phosphate dehydratase/enolase-phosphatase E1 2;putative bifunctional methylthioribulose-1-phosphate dehydratase/enolase-phosphatase E1
KN361_1g007625	Chr 1	52713578	52715754	polyphenol oxidase 2;(+)-larreatricin hydroxylase
KN361_1g007638	Chr 1	52784313	52786608	ATP carrier protein 3
KN361_1g007670	Chr 1	53079638	53081264	trans-resveratrol di-O-methyltransferase
KN361_1g007687	Chr 1	53222528	53226852	putative LRR receptor-like serine/threonine-protein kinase
KN361_1g007764	Chr 1	54209909	54212501	11-oxo-beta-amyrin 30-oxidase
KN361_1g007904	Chr 1	55676060	55677623	protein translation factor SUI1
KN361_1g007905	Chr 1	55677138	55682127	hypothetical protein;glycine-rich RNA-binding protein RZ1C;serine/arginine-rich splicing factor SR45a
KN361_1g007952	Chr 1	56042212	56046776	hypothetical protein
KN361_1g007964	Chr 1	56192656	56193763	vascular-related protein 1
KN361_1g007991	Chr 1	56437667	56439199	hypothetical protein
KN361_1g008004	Chr 1	56593037	56595519	hypothetical protein

KN361_1g008010	Chr 1	56667030	56671142	retrotransposon gag protein
KN361_1g008016	Chr 1	56741224	56748310	kinesin-like protein KIN-14N
KN361_1g008074	Chr 1	57551455	57553421	putative late blight resistance protein R1A-6
KN361_1g008158	Chr 1	59797801	59802314	SWIM zinc finger protein
KN361_1g008159	Chr 1	59805355	59808440	chromo chromatin organization modifier domain-containing protein
KN361_1g008163	Chr 1	59824009	59826008	hypothetical protein
KN361_1g008218	Chr 1	61413898	61424937	glutathione S-transferase T3
KN361_1g008238	Chr 1	61762688	61768475	protein FAR1-related sequence FRS5
KN361_1g008239	Chr 1	61768654	61770850	succinate--CoA ligase ADP-forming subunit beta
KN361_1g008240	Chr 1	61769332	61774317	succinate--CoA ligase ADP-forming subunit beta
KN361_1g008366	Chr 1	64144827	64146328	putative ribonuclease H protein
KN361_1g008372	Chr 1	64324367	64325034	hypothetical protein
KN361_1g008373	Chr 1	64327831	64329615	hypothetical protein
KN361_1g008434	Chr 1	67235603	67242276	serine hydroxymethyltransferase 4
KN361_1g008459	Chr 1	68062502	68064288	histone H3-lysine
KN361_1g008479	Chr 1	68458663	68463708	retrovirus-related Pol polyprotein from transposon TNT 1-94
KN361_1g008484	Chr 1	68569350	68571916	hypothetical protein
KN361_1g008572	Chr 1	71590009	71592763	hypothetical protein
KN361_1g008584	Chr 1	72134769	72137088	retrovirus-related Pol polyprotein from transposon TNT 1-94
KN361_1g008602	Chr 1	72680545	72688070	retrovirus-related Pol polyprotein from transposon TNT 1-94
KN361_1g008609	Chr 1	72893462	72896495	hypothetical protein
KN361_1g008611	Chr 1	72936030	72940381	1-acylglycerol-3-phosphate O-acyltransferase
KN361_1g008614	Chr 1	73139784	73144678	RNA-directed DNA polymerase
KN361_1g008618	Chr 1	73162469	73169625	retrotransposon gag protein
KN361_1g008640	Chr 1	73951668	73952876	hypothetical protein
KN361_1g008686	Chr 1	75494033	75498179	hypothetical protein

KN361_1g008708	Chr 1	76746437	76749466	hypothetical protein
KN361_1g008710	Chr 1	76840407	76841115	retrotransposon gag protein
KN361_1g008772	Chr 1	79022767	79026740	putative ribonuclease H protein
KN361_1g008800	Chr 1	79335731	79337098	hypothetical protein
KN361_1g008840	Chr 1	80801206	80806213	hypothetical protein
KN361_1g008941	Chr 1	83988300	83988681	hypothetical protein
KN361_1g008949	Chr 1	84233427	84236859	exocyst complex component EXO70A1
KN361_1g009045	Chr 1	88168782	88171008	bZIP transcription factor
KN361_1g009102	Chr 1	90805939	90808114	B3 domain-containing protein
KN361_1g009294	Chr 1	97421613	97424654	retrovirus-related Pol polyprotein from transposon TNT 1-94
KN361_1g009400	Chr 1	101407139	101416629	hypothetical protein
KN361_1g009402	Chr 1	101422915	101424372	hypothetical protein
KN361_1g009403	Chr 1	101442325	101445663	serine--tRNA ligase
KN361_1g009494	Chr 1	104232655	104238226	hypothetical protein
KN361_1g009499	Chr 1	104243237	104249198	protein FAR1-related sequence FRS5
KN361_1g009510	Chr 1	104447623	104452140	ferric reduction oxidase 5;ferric reduction oxidase 4
KN361_1g009525	Chr 1	104617081	104628635	hypothetical protein;plant transposase Ptta or En or Spm family protein;preprotein translocase subunit SCY2
KN361_1g009541	Chr 1	105064532	105067919	plant transposase Ptta or En or Spm family protein;aminotransferase class I and II
KN361_1g009542	Chr 1	105097403	105098885	SopA-like central domain-containing protein
KN361_1g009556	Chr 1	106446216	106447743	integrase core domain-containing protein;hypothetical protein
KN361_1g009557	Chr 1	106451933	106454126	hypothetical protein
KN361_1g009615	Chr 1	108861481	108864308	hypothetical protein
KN361_1g009742	Chr 1	114609865	114617232	methyl-CpG-binding domain-containing protein 4-like protein
KN361_1g009869	Chr 1	117886298	117887303	hypothetical protein
KN361_1g009873	Chr 1	117993196	117997096	tropomyosin;hypothetical protein
KN361_1g009909	Chr 1	119461249	119465177	geranylgeranyl pyrophosphate synthase 7

KN361_1g009932	Chr 1	120156817	120158494	DNA-directed RNA polymerase V subunit 7
KN361_1g009935	Chr 1	120179156	120180117	hypothetical protein
KN361_1g009941	Chr 1	120240756	120243859	putative ribonuclease H protein
KN361_1g009947	Chr 1	120549738	120554793	hypothetical protein
KN361_1g009964	Chr 1	120750149	120753000	hypothetical protein
KN361_1g009968	Chr 1	120843339	120850777	late blight resistance protein R1-A
KN361_1g009970	Chr 1	120959507	120962678	putative late blight resistance protein R1B-16
KN361_1g009971	Chr 1	121028505	121031726	hypothetical protein
KN361_1g010010	Chr 1	121933076	121935188	hypothetical protein
KN361_1g010016	Chr 1	122263360	122264355	hypothetical protein
KN361_1g010082	Chr 1	123885878	123887452	hypothetical protein
KN361_1g010124	Chr 1	124852764	124854062	hypothetical protein
KN361_1g010137	Chr 1	125013576	125016804	hypothetical protein
KN361_1g010144	Chr 1	125192332	125198344	macoilin family protein
KN361_1g010152	Chr 1	125378676	125391285	putative late blight resistance protein R1B-17;late blight resistance protein R1-A;disease resistance protein PIK6-NP
KN361_1g010155	Chr 1	125391860	125395476	putative L-type lectin-domain-containing protein containing receptor kinase S.7;hypothetical protein
KN361_1g010191	Chr 1	126249281	126252973	hypothetical protein
KN361_1g010221	Chr 1	126518259	126519175	hypothetical protein
KN361_1g010261	Chr 1	127020814	127025536	hypothetical protein
KN361_1g010390	Chr 1	128647328	128649292	protein AGENET domain-containing protein P1
KN361_1g010408	Chr 1	128791594	128793432	hypothetical protein
KN361_1g010434	Chr 1	128988705	128992442	DNA-directed RNA polymerase V subunit 7
KN361_1g010442	Chr 1	129043361	129047637	DNA-directed RNA polymerase V subunit 7
KN361_1g010469	Chr 1	129249220	129252031	wall-associated receptor kinase 5
KN361_1g010538	Chr 1	129774476	129778254	mini-chromosome maintenance complex-binding protein

KN361_1g010591	Chr 1	130077331	130080410	CBL-interacting protein kinase 2
KN361_1g010825	Chr 1	132162584	132163306	hypothetical protein
KN361_1g010884	Chr 1	132607958	132611270	hypothetical protein
KN361_1g010941	Chr 1	132864498	132868528	meiosis-specific protein ASY1
KN361_1g010971	Chr 1	133178911	133182463	hypothetical protein
KN361_1g011044	Chr 1	133643975	133645223	putative protein phosphatase 2C 63
KN361_1g011047	Chr 1	133655478	133655920	nitrogen permease regulator of amino acid transport activity 3
KN361_1g011057	Chr 1	133743345	133746666	putative cyclin-A3-1;hypothetical protein;cyclin-A3-2
KN361_1g011243	Chr 1	135106298	135109293	hypothetical protein
KN361_1g011295	Chr 1	135504935	135505244	hypothetical protein
KN361_1g011311	Chr 1	135617570	135619670	hypothetical protein
KN361_1g011313	Chr 1	135640937	135642711	protein EXORDIUM-like 3
KN361_1g011316	Chr 1	135645564	135646067	GCM motif protein
KN361_1g011396	Chr 1	136292167	136292884	zinc-binding in reverse transcriptase
KN361_1g011434	Chr 1	136535179	136537803	TFIIFbeta subunit HTH domain-containing protein
KN361_nc022	Chr 1	5755021	5755237	small nucleolar RNA U3
KN361_nc029	Chr 1	17762285	17762419	small nuclear RNA U1
KN361_nc115	Chr 1	104432399	104434651	long non-coding RNA
KN361_nc156	Chr 1	90298326	90308172	long non-coding RNA
KN361_nc163	Chr 1	57439487	57441037	long non-coding RNA
KN361_2g000053	Chr 2	322988	324899	O-fucosyltransferase 23
KN361_2g000119	Chr 2	675469	678333	patellin-4
KN361_2g000121	Chr 2	684116	685916	berberine bridge enzyme-like 21
KN361_2g000204	Chr 2	1205378	1207114	3-galactosyltransferase 20
KN361_2g000339	Chr 2	2009107	2009489	hypothetical protein
KN361_2g000503	Chr 2	3100970	3102017	putative ribonuclease H protein
KN361_2g000507	Chr 2	3109871	3110727	hypothetical protein
KN361_2g000668	Chr 2	4020475	4021086	pentatricopeptide repeat-containing protein

KN361_2g000743	Chr 2	4560363	4560717	defensin-like protein 1
KN361_2g000771	Chr 2	4748388	4753733	DNA cross-link repair protein SNM1
KN361_2g000826	Chr 2	5049239	5050763	peroxidase 9
KN361_2g000843	Chr 2	5119658	5120486	protein kirola
KN361_2g000879	Chr 2	5257384	5259805	E3 ubiquitin-protein ligase SINA-like 2;E3 ubiquitin-protein ligase SINA-like 1
KN361_2g000881	Chr 2	5264900	5268290	putative E3 ubiquitin-protein ligase SINA-like 6
KN361_2g000969	Chr 2	5702059	5703168	hypothetical protein
KN361_2g001107	Chr 2	6510023	6514036	lysine-specific demethylase JMJ25
KN361_2g001110	Chr 2	6524253	6528487	hypothetical protein
KN361_2g001198	Chr 2	7098361	7101341	gibberellin 20 oxidase 5;gibberellin 20 oxidase 3;gibberellin 20 oxidase 1-B;gibberellin 20 oxidase 2
KN361_2g001255	Chr 2	7420345	7420965	hypothetical protein
KN361_2g001293	Chr 2	7655029	7657120	hypothetical protein
KN361_2g001349	Chr 2	7973340	7981966	retrovirus-related Pol polyprotein from transposon RE1
KN361_2g001377	Chr 2	8148000	8148841	hypothetical protein
KN361_2g001408	Chr 2	8302947	8305063	retrovirus-related Pol polyprotein from transposon RE1
KN361_2g001513	Chr 2	8956024	8957162	MaoC like domain-containing protein
KN361_2g001547	Chr 2	9208114	9208716	septum formation topological specificity factor MinE
KN361_2g001584	Chr 2	9511546	9513297	hypothetical protein
KN361_2g002176	Chr 2	13487919	13497799	acetylajmalan esterase;GDSL esterase/lipase
KN361_2g002250	Chr 2	13966838	13969630	carbohydrate binding domain-containing protein
KN361_2g002461	Chr 2	15423051	15425936	G2/mitotic-specific cyclin-1;G2/mitotic-specific cyclin-2
KN361_2g002584	Chr 2	16231229	16234095	NAC domain-containing protein 89;NAC domain-containing protein 40
KN361_2g002763	Chr 2	17572473	17574432	transcription factor MYB36
KN361_2g002857	Chr 2	18242488	18246241	mitogen-activated protein kinase NTF6
KN361_2g002899	Chr 2	18619028	18620557	hypothetical protein

KN361_2g003183	Chr 2	20661046	20662505	octanoyltransferase LIB
KN361_2g003365	Chr 2	21842850	21844481	F-box/kelch-repeat-containing protein
KN361_2g003375	Chr 2	21896184	21898283	hypothetical protein
KN361_2g003404	Chr 2	22076217	22078200	retrovirus-related Pol polyprotein from transposon RE1
KN361_2g003429	Chr 2	22271051	22276433	thioredoxin domain-containing protein PLP3B
KN361_2g003432	Chr 2	22282246	22292071	LIM domain-containing protein PLIM2a;LIM domain-containing protein PLIM2c
KN361_2g003744	Chr 2	24785291	24790472	cystathionine beta-lyase
KN361_2g004055	Chr 2	29610280	29611092	hypothetical protein
KN361_2g004067	Chr 2	29800320	29800787	defensin-like protein 306
KN361_2g004122	Chr 2	30825277	30830251	mediator of RNA polymerase II transcription subunit 15a
KN361_2g004126	Chr 2	30996825	30997334	hypothetical protein
KN361_2g004281	Chr 2	37957279	37957876	hypothetical protein
KN361_2g004405	Chr 2	43116870	43117830	hypothetical protein
KN361_2g004543	Chr 2	50864440	50866206	hypothetical protein
KN361_2g004864	Chr 2	65911470	65913724	NB glycoprotein;chromatin assembly factor 1 subunit A;hypothetical protein
KN361_2g004866	Chr 2	65918314	65920179	retrovirus-related Pol polyprotein from transposon TNT 1-94
KN361_2g005361	Chr 2	87225758	87226070	hypothetical protein
KN361_2g005420	Chr 2	89190896	89197095	hypothetical protein
KN361_2g005424	Chr 2	89240317	89245763	hypothetical protein
KN361_2g005440	Chr 2	89510850	89511486	gamma-thionin family protein
KN361_2g005460	Chr 2	90162893	90164978	hypothetical protein
KN361_2g005503	Chr 2	91186086	91190100	polyadenylate-binding protein-interacting protein 9
KN361_2g005520	Chr 2	91419749	91420845	hypothetical protein
KN361_2g005546	Chr 2	92397071	92397811	S locus-related glyco P
KN361_2g005807	Chr 2	101764249	101768610	NRT1/PTR family protein 2.11

KN361_2g005967	Chr 2	105129547	105130912	heavy metal-associated isoprenylated plant protein 35;heavy metal-associated isoprenylated plant protein 36
KN361_2g006104	Chr 2	106824887	106826265	hypothetical protein
KN361_2g006114	Chr 2	106907342	106915299	hypothetical protein
KN361_2g006174	Chr 2	107420421	107427741	3-beta-glucosidase 2
KN361_2g006241	Chr 2	107823524	107825424	trihelix transcription factor ASR3
KN361_2g006253	Chr 2	107901589	107902913	F-box protein CPR1
KN361_2g006278	Chr 2	108170066	108171254	BAG family protein molecular chaperone regulator 5
KN361_2g006284	Chr 2	108214275	108216724	retrovirus-related Pol polyprotein from transposon RE1
KN361_2g006409	Chr 2	109403187	109404201	hypothetical protein
KN361_2g006436	Chr 2	109585136	109598348	receptor-like protein 35;LRR receptor-like serine/threonine-protein kinase EFR
KN361_2g006467	Chr 2	109817858	109818431	hypothetical protein
KN361_2g006520	Chr 2	110154337	110156415	hypothetical protein
KN361_2g006529	Chr 2	110216588	110218353	CASP-like protein 5C3;CASP-like protein 1F2
KN361_2g006581	Chr 2	110669610	110670097	S locus-related glyco P
KN361_2g006615	Chr 2	110897495	110901567	scopoletin glucosyltransferase
KN361_2g006671	Chr 2	111271948	111274676	tubby-like protein 8
KN361_2g006693	Chr 2	111430803	111431358	hypothetical protein
KN361_2g006734	Chr 2	111734992	111740276	plasma membrane ATPase 4
KN361_2g006835	Chr 2	112289044	112293069	protein SDA1;hypothetical protein
KN361_2g006910	Chr 2	112768365	112769431	putative serine/threonine-protein kinase CST;putative serine/threonine-protein kinase PIX13
KN361_2g006929	Chr 2	112847283	112854906	exosome complex exonuclease RRP44 A
KN361_2g006949	Chr 2	112965497	112969039	cell division cycle-associated protein 8
KN361_2g007009	Chr 2	113343101	113344406	hypothetical protein;protein PPP4R2
KN361_2g007060	Chr 2	113681396	113682695	hypothetical protein
KN361_2g007137	Chr 2	114101641	114110679	ubiquitin-like domain-containing protein CIP73
KN361_2g007145	Chr 2	114183455	114186720	neutral ceramidase;neutral ceramidase 2

KN361_2g007150	Chr 2	114203399	114205948	lysine histidine transporter-like 8
KN361_2g007169	Chr 2	114333161	114338543	high mobility group B protein 9
KN361_2g007235	Chr 2	114767182	114769316	receptor protein kinase CLAVATA1
KN361_2g007465	Chr 2	116273626	116275180	zeatin O-glucosyltransferase
KN361_2g007476	Chr 2	116348099	116349297	SPX domain-containing protein membrane protein
KN361_2g007534	Chr 2	116738622	116739673	protein NDR1
KN361_2g007585	Chr 2	117220812	117222624	zinc-finger homeodomain-containing protein 1
KN361_2g007627	Chr 2	117451747	117455723	hypothetical protein
KN361_2g007644	Chr 2	117569771	117571385	dof zinc finger protein DOF5.7
KN361_2g007833	Chr 2	118945791	118946459	hypothetical protein
KN361_2g007905	Chr 2	119327660	119329056	iridoid oxidase
KN361_2g007939	Chr 2	119511290	119512664	putative pentatricopeptide repeat-containing protein
KN361_2g007944	Chr 2	119543648	119547071	rop guanine nucleotide exchange factor 7
KN361_2g007998	Chr 2	119913373	119916717	cytochrome P450
KN361_2g008082	Chr 2	120450661	120451944	D-amino acid oxidase activator
KN361_2g008136	Chr 2	120755455	120760093	histidine-containing phosphotransfer protein 1
KN361_2g008140	Chr 2	120788377	120790148	hypothetical protein
KN361_2g008189	Chr 2	121078753	121080128	pathogenesis-related thaumatin-like protein 3.5
KN361_2g008338	Chr 2	122114156	122115216	hypothetical protein
KN361_2g008345	Chr 2	122172779	122175977	wall-associated receptor kinase 5;wall-associated receptor kinase 2
KN361_2g008364	Chr 2	122303221	122305924	NRT1/PTR family protein 5.10
KN361_2g008374	Chr 2	122367354	122368734	cytochrome P450
KN361_2g008563	Chr 2	123605277	123606012	hypothetical protein
KN361_2g008581	Chr 2	123727498	123729014	dof zinc finger protein DOF5.3;dof zinc finger protein DOF5.6
KN361_2g008882	Chr 2	125417047	125418527	putative xyloglucan endotransglucosylase/hydrolase protein 33
KN361_2g008936	Chr 2	125725437	125729490	sugar transporter ERD6-like 16

KN361_2g008949	Chr 2	125791380	125792569	BON1-associated protein 2
KN361_2g008979	Chr 2	125903447	125904081	monothiol glutaredoxin-S10
KN361_2g008988	Chr 2	125963160	125964140	2S albumin
KN361_2g009006	Chr 2	126040803	126042334	transcription factor JAMYB;transcription factor MYB78
KN361_2g009060	Chr 2	126379407	126381079	protein PHYLLO
KN361_2g009067	Chr 2	126430050	126431492	UDP-glycosyltransferase 83A1
KN361_2g009085	Chr 2	126543321	126545459	glyoxysomal fatty acid beta-oxidation complex protein MFP-a
KN361_2g009114	Chr 2	126685792	126687680	pirin-like protein
KN361_2g009128	Chr 2	126737054	126739000	putative aldo-keto reductase 2;putative aldo-keto reductase 5
KN361_2g009161	Chr 2	126954476	126955961	hypothetical protein
KN361_2g009235	Chr 2	127451734	127453855	putative (S)-N-methylcoclaurine 3'-hydroxylase isozyme 2;(S)-N-methylcoclaurine 3'-hydroxylase isozyme 1
KN361_2g009243	Chr 2	127486451	127486917	hypothetical protein
KN361_2g009302	Chr 2	127773559	127777856	putative pectinesterase 68
KN361_2g009321	Chr 2	127875963	127879465	putative aspartyl protease
KN361_2g009327	Chr 2	127905120	127906719	root meristem growth factor 6
KN361_2g009335	Chr 2	127936933	127937847	pEARLI1-like lipid transfer protein 2
KN361_2g009441	Chr 2	128616303	128619884	hypothetical protein;LXG domain-containing protein of WXG superfamily protein
KN361_nc364	Chr 2	126503976	126505389	long non-coding RNA
KN361_3g000006	Chr 3	27855	28470	hypothetical protein
KN361_3g000281	Chr 3	1718684	1719695	CLAVATA3/ESR-related protein 25
KN361_3g000290	Chr 3	1768019	1769123	pathogenesis-related protein 5
KN361_3g000332	Chr 3	2091656	2092609	plant invertase/pectin methylesterase inhibitor
KN361_3g000352	Chr 3	2243649	2249494	amino acid transporter AVT1J
KN361_3g000383	Chr 3	2467291	2474951	phosphate transporter PHO1 4;phosphate transporter PHO1 9
KN361_3g000457	Chr 3	2893206	2895670	hypothetical protein
KN361_3g000478	Chr 3	2986385	2992136	putative receptor-like protein kinase

KN361_3g000516	Chr 3	3280011	3280904	basic leucine zipper 43
KN361_3g000568	Chr 3	3540763	3542875	cyclin-D5-3;cyclin-D5-1
KN361_3g000775	Chr 3	4992395	4993075	Syd protein SUKH-2
KN361_3g000922	Chr 3	5986207	5987218	catalase isozyme 3
KN361_3g000950	Chr 3	6117619	6120385	hypothetical protein
KN361_3g000974	Chr 3	6268492	6270966	lysine histidine transporter 2;lysine histidine transporter 1
KN361_3g001024	Chr 3	6536825	6538071	hypothetical protein
KN361_3g001052	Chr 3	6664843	6667694	scarecrow-like protein 28
KN361_3g001225	Chr 3	7697903	7701125	G-type lectin S-receptor-like serine/threonine-protein kinase LECRK3
KN361_3g001261	Chr 3	7925310	7926171	hypothetical protein
KN361_3g001325	Chr 3	8282163	8282648	hypothetical protein
KN361_3g001407	Chr 3	8773589	8777367	triacylglycerol lipase 2
KN361_3g001471	Chr 3	9095606	9101085	nuclear/nucleolar GTPase 2
KN361_3g001714	Chr 3	10498418	10499205	putative ribonuclease H protein
KN361_3g001758	Chr 3	10763490	10765128	polygalacturonase
KN361_3g001791	Chr 3	10938476	10939032	F-box protein
KN361_3g001792	Chr 3	10945192	10946688	hypothetical protein
KN361_3g001793	Chr 3	10946772	10948925	hypothetical protein
KN361_3g001794	Chr 3	10951372	10954851	60S acidic ribosomal protein P0-1;60S acidic ribosomal protein P0-2
KN361_3g001795	Chr 3	10959249	10961335	9-cis-epoxycarotenoid dioxygenase NCED6
KN361_3g001842	Chr 3	11191629	11195870	protein NLP8
KN361_3g001942	Chr 3	11820572	11824225	hypothetical protein
KN361_3g002042	Chr 3	12400568	12402724	hypothetical protein
KN361_3g002101	Chr 3	12700256	12701411	werner syndrome-like exonuclease
KN361_3g002218	Chr 3	13445044	13446435	hypothetical protein
KN361_3g002398	Chr 3	14556994	14561990	protein IQ-domain-containing protein IQD20;hypothetical protein
KN361_3g002402	Chr 3	14567578	14570396	hypothetical protein

KN361_3g002619	Chr 3	15837071	15838915	mitogen-activated protein kinase kinase NPK1
KN361_3g002678	Chr 3	16115997	16117707	F-box protein CPR1
KN361_3g002770	Chr 3	16586185	16589240	B3 domain-containing protein
KN361_3g002861	Chr 3	17101694	17103058	pollen allergen Che a 1
KN361_3g002867	Chr 3	17136671	17139207	hypothetical protein
KN361_3g002999	Chr 3	18109720	18110322	protein ymf17
KN361_3g003003	Chr 3	18127514	18128837	1-cys peroxiredoxin
KN361_3g003055	Chr 3	18433738	18441031	protein detoxification DTX28
KN361_3g003124	Chr 3	18904121	18905088	hypothetical protein
KN361_3g003136	Chr 3	18959933	18961997	hypothetical protein
KN361_3g003271	Chr 3	19751910	19754364	germin-like protein 9-3
KN361_3g003276	Chr 3	19762071	19763147	protein FAF-like
KN361_3g003578	Chr 3	21796051	21797987	glutathione S-transferase U17
KN361_3g003750	Chr 3	22972938	22974207	dormancy-associated protein 1
KN361_3g003802	Chr 3	23342929	23345343	RimP N-terminal domain-containing protein
KN361_3g003810	Chr 3	23408518	23410247	anthocyanin 5-O-glucoside-6"-O-malonyltransferase
KN361_3g004204	Chr 3	26073443	26074216	hypothetical protein
KN361_3g004215	Chr 3	26124749	26129053	cyclin-D1-1
KN361_3g004283	Chr 3	26559405	26562047	zinc finger protein C3HC4 type RING finger;E3 ubiquitin-protein ligase ORTHRUS 1;hypothetical protein
KN361_3g004381	Chr 3	27164784	27166564	protein altered xyloglucan 4-like
KN361_3g004423	Chr 3	27446300	27448175	beta-amyrin 28-monooxygenase
KN361_3g004519	Chr 3	28129091	28130432	transcription factor UPBEAT1
KN361_3g004520	Chr 3	28153769	28156170	UDP-glycosyltransferase 83A1
KN361_3g004574	Chr 3	28539575	28541792	proline dehydrogenase 2
KN361_3g004703	Chr 3	29356810	29363284	zinc knuckle domain-containing protein;hypothetical protein
KN361_3g004729	Chr 3	29489393	29490430	ribulose bisphosphate carboxylase small subunit 3
KN361_3g004949	Chr 3	31133539	31137373	hypothetical protein

KN361_3g004996	Chr 3	31416092	31419423	protein gamete expressed GEX1
KN361_3g005053	Chr 3	31801143	31804095	hypothetical protein
KN361_3g005298	Chr 3	33779304	33784488	LRR receptor-like serine/threonine-protein kinase FEI 2;LRR receptor-like serine/threonine-protein kinase FEI 1
KN361_3g005368	Chr 3	34682765	34685928	protein argonaute 1A;protein argonaute 1
KN361_3g005532	Chr 3	37494861	37505467	putative serine/threonine-protein kinase CST;putative serine/threonine-protein kinase PIX7;putative serine/threonine-protein kinase PBL2
KN361_3g005534	Chr 3	37509188	37520361	putative leucine-rich repeat-containing protein receptor-like protein kinase;putative serine/threonine-protein kinase PBL2;putative serine/threonine-protein kinase CST
KN361_3g005659	Chr 3	39757331	39759090	hypothetical protein
KN361_3g005819	Chr 3	42219390	42220847	retrotransposon gag protein
KN361_3g005837	Chr 3	42576715	42579101	hypothetical protein
KN361_3g005840	Chr 3	42700739	42702994	retroviral aspartyl protease
KN361_3g005952	Chr 3	46269803	46271105	heavy metal-associated isoprenylated plant protein 20
KN361_3g006052	Chr 3	48623311	48629452	protease do-like 7
KN361_3g006055	Chr 3	48764106	48769508	gag-polypeptide of LTR copia-type
KN361_3g006103	Chr 3	51174138	51177547	protein ABIL2;ferredoxin-dependent glutamate synthase
KN361_3g006157	Chr 3	53384561	53387362	hypothetical protein
KN361_3g006260	Chr 3	57129191	57130567	retrovirus-related Pol polyprotein from transposon RE1
KN361_3g006381	Chr 3	61469989	61472472	scarecrow-like protein 6
KN361_3g006423	Chr 3	63029209	63030641	hypothetical protein
KN361_3g006610	Chr 3	69919051	69920551	hypothetical protein
KN361_3g006670	Chr 3	72289050	72291359	G2/mitotic-specific cyclin-1
KN361_3g006673	Chr 3	72592506	72594564	retrotransposon gag protein
KN361_3g006765	Chr 3	75885870	75887087	putative ribonuclease H protein
KN361_3g006906	Chr 3	82183696	82184773	hypothetical protein

KN361_3g007263	Chr 3	92884687	92888098	nuclear transcription factor Y subunit A-1;nuclear transcription factor Y subunit A-7
KN361_3g007270	Chr 3	93123457	93125245	hypothetical protein
KN361_3g007671	Chr 3	100544994	100546096	putative ribonuclease H protein
KN361_3g007689	Chr 3	100763106	100763744	hypothetical protein
KN361_3g007777	Chr 3	101646320	101648037	protein rice salt sensitive RSS3
KN361_3g007828	Chr 3	102352304	102357349	hypothetical protein
KN361_3g007875	Chr 3	102760321	102761572	hypothetical protein
KN361_3g007879	Chr 3	102769763	102769995	hypothetical protein
KN361_3g008066	Chr 3	104743714	104744767	early nodulin-like protein 1
KN361_3g008309	Chr 3	106900401	106903414	gag protein p24 N-terminal domain-containing protein
KN361_3g008311	Chr 3	106909558	106912723	hypothetical protein
KN361_3g008374	Chr 3	107542355	107544053	polygalacturonase
KN361_3g008403	Chr 3	107851082	107853446	FT-interacting protein 3
KN361_3g008406	Chr 3	107871874	107874376	hypothetical protein
KN361_3g008451	Chr 3	108145322	108145923	putative disease resistance protein RDL5
KN361_3g008517	Chr 3	108582318	108583654	elicitor-responsive protein 3
KN361_3g008547	Chr 3	108814548	108818032	transposase mutator family protein;MuDR family protein transposase
KN361_3g008557	Chr 3	108884779	108886733	PDDEXK-like protein
KN361_3g008645	Chr 3	109697561	109700599	pectin acetylesterase 7
KN361_3g008689	Chr 3	110094463	110096573	retrovirus-related Pol polyprotein from transposon TNT 1-94
KN361_3g008705	Chr 3	110356047	110361487	small RNA 2'-O-methyltransferase
KN361_3g008709	Chr 3	110385129	110387108	GATA transcription factor 2
KN361_3g008718	Chr 3	110506534	110508250	squamosa promoter-binding-like protein 8
KN361_3g008753	Chr 3	110816335	110820072	hypothetical protein
KN361_3g008980	Chr 3	113023963	113026463	outer capsid protein VP7;40S ribosome biogenesis protein Tsr1 and BMS1 C-terminal
KN361_3g009005	Chr 3	113274435	113276384	aldehyde oxidase GLOX1
KN361_3g009032	Chr 3	113497427	113501936	C1 domain-containing protein

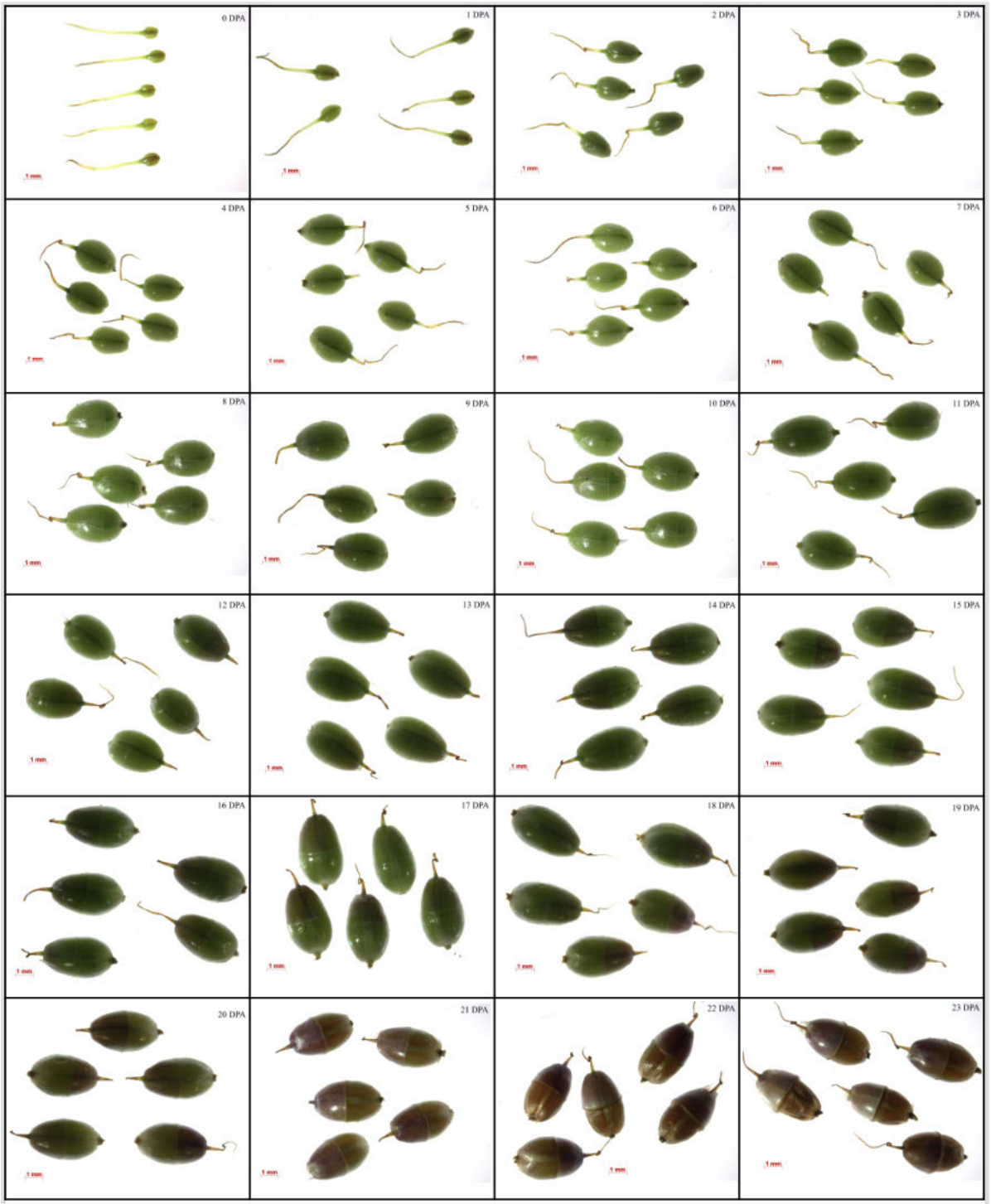
KN361_3g009054	Chr 3	113666440	113669935	protein NSP-interacting kinase NIK2
KN361_3g009110	Chr 3	114438708	114439590	hypothetical protein
gene-1177	Chr 4	49537601	49537670	tRNA-Phe
KN361_4g000022	Chr 4	109284	110422	hypothetical protein
KN361_4g000174	Chr 4	1363127	1364289	RING-variant domain-containing protein
KN361_4g000175	Chr 4	1381329	1382426	hypothetical protein
KN361_4g000257	Chr 4	2065701	2066880	hypothetical protein
KN361_4g000269	Chr 4	2162150	2163374	hypothetical protein
KN361_4g000347	Chr 4	2844310	2847330	MOTHER of FT and TFL1
KN361_4g000482	Chr 4	3965779	3966910	hypothetical protein
KN361_4g000510	Chr 4	4137327	4139033	non-functional NADPH-dependent codeinone reductase 2;methylecgonone reductase
KN361_4g000542	Chr 4	4420027	4421924	callose synthase 3
KN361_4g000570	Chr 4	4582968	4595630	SNF2 domain-containing protein CLASSY 3;SNF2 domain-containing protein CLASSY 4
KN361_4g000637	Chr 4	5158048	5161801	agamous-like MADS-box protein AGL11
KN361_4g000646	Chr 4	5229695	5230872	hypothetical protein
KN361_4g000654	Chr 4	5344867	5346784	palmitoyl-acyl carrier protein thioesterase
KN361_4g000664	Chr 4	5397484	5399888	11-beta-hydroxysteroid dehydrogenase-like 6;11-beta-hydroxysteroid dehydrogenase A
KN361_4g000773	Chr 4	6554480	6556194	fructose-bisphosphate aldolase class-I
KN361_4g000870	Chr 4	7371128	7373895	cytochrome P450
KN361_4g000911	Chr 4	7669439	7671248	putative xyloglucan galactosyltransferase GT12
KN361_4g000997	Chr 4	8364828	8365539	protein SPH27
KN361_4g001090	Chr 4	9259396	9261352	putative disease resistance protein RXW24L
KN361_4g001107	Chr 4	9345864	9350830	hypothetical protein;PHD finger protein ALFIN-like 9
KN361_4g001217	Chr 4	10620808	10625392	FCS-like zinc finger protein 10
KN361_4g001267	Chr 4	11226237	11229647	axial regulator YABBY 4
KN361_4g001330	Chr 4	11937634	11938898	PHD finger protein

KN361_4g001342	Chr 4	12146211	12147110	hypothetical protein
KN361_4g001356	Chr 4	12376900	12377642	hypothetical protein
KN361_4g001379	Chr 4	12624919	12627928	hypothetical protein
KN361_4g001443	Chr 4	13903415	13904650	microtubule-associated protein 6
KN361_4g001495	Chr 4	15173422	15174365	protein FAR1-related sequence FRS5
KN361_4g001627	Chr 4	20214969	20215457	hypothetical protein
KN361_4g001671	Chr 4	21716885	21719187	plant transposase Ptta or En or Spm family protein
KN361_4g001684	Chr 4	22134443	22142647	hypothetical protein
KN361_4g001773	Chr 4	24530442	24533520	retrovirus-related Pol polyprotein from transposon TNT 1-94
KN361_4g001997	Chr 4	32889268	32892067	RNA-directed DNA polymerase
KN361_4g002080	Chr 4	35755758	35757839	retrovirus-related Pol polyprotein from transposon RE1
KN361_4g002081	Chr 4	35760948	35763872	signal recognition particlealpha subunit N-terminal
KN361_4g002170	Chr 4	38807658	38809457	hypothetical protein
KN361_4g002220	Chr 4	41203438	41205519	putative transcription factor
KN361_4g002295	Chr 4	43116712	43119946	PIF1-like helicase;helicase
KN361_4g002512	Chr 4	50585776	50587216	hypothetical protein
KN361_4g002881	Chr 4	63672890	63676670	hypothetical protein
KN361_4g002905	Chr 4	64873182	64877470	hypothetical protein
KN361_4g002988	Chr 4	67027321	67031722	pleckstrin y domain-containing protein 1;serine--tRNA ligase
KN361_4g003076	Chr 4	69406917	69407881	hypothetical protein
KN361_4g003271	Chr 4	76602246	76604499	hypothetical protein
KN361_4g003295	Chr 4	76984023	76985552	transcription factor MYB1
KN361_4g003345	Chr 4	77573150	77575425	UDP-glycosyltransferase 87A2
KN361_4g003385	Chr 4	78640371	78642523	receptor-like protein EIX2
KN361_4g003397	Chr 4	78857234	78858477	hypothetical protein
KN361_4g003435	Chr 4	79492822	79494834	quinone-oxidoreductase QR1;protein QORH
KN361_4g003823	Chr 4	83061159	83063237	putative microtubule-binding protein TANGLED

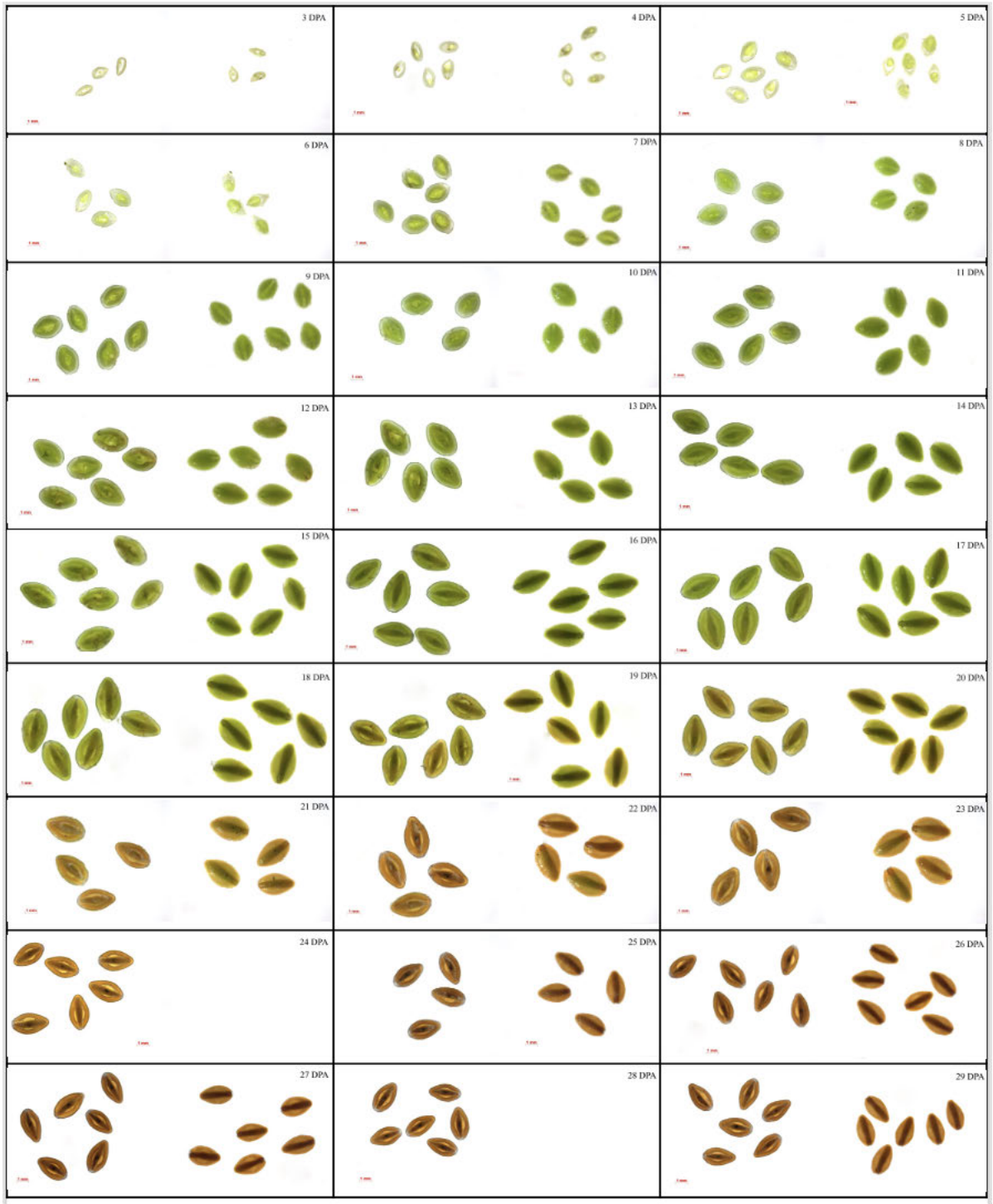
KN361_4g003892	Chr 4	83519715	83520720	xyloglucan endotransglucosylase/hydrolase protein 3
KN361_4g004271	Chr 4	86350092	86352958	cell division control protein 2
KN361_4g004459	Chr 4	87469422	87470151	hypothetical protein
KN361_4g004468	Chr 4	87489116	87490626	F-box protein
KN361_4g004477	Chr 4	87504702	87506032	F-box protein PR1
KN361_4g004478	Chr 4	87508002	87509319	F-box protein PR1
KN361_4g004537	Chr 4	87928241	87929392	ethylene-responsive transcription factor ERF012
KN361_4g004657	Chr 4	88685925	88687681	F-box/kelch-repeat-containing protein
KN361_4g004696	Chr 4	88947818	88949098	hypothetical protein
KN361_4g004740	Chr 4	89219790	89220172	hypothetical protein
KN361_4g004757	Chr 4	89326212	89326809	hypothetical protein
KN361_4g004836	Chr 4	89913263	89915318	galactinol synthase 2
KN361_4g004901	Chr 4	90293974	90295294	transcription factor MYB102
KN361_4g005054	Chr 4	91301553	91306000	putative LRR receptor-like serine/threonine-protein kinase
KN361_4g005077	Chr 4	91431187	91432395	mitochondrial import inner membrane translocase subunit TIM50
KN361_4g005088	Chr 4	91523603	91528034	pentatricopeptide repeat-containing protein
KN361_4g005304	Chr 4	92874205	92875169	transcription factor MYB3
KN361_4g005364	Chr 4	93207099	93209790	putative protein XLY2;putative beta-D-xylosidase 2;beta-D-xylosidase 1
KN361_4g005437	Chr 4	93609731	93611177	hypothetical protein
KN361_4g005448	Chr 4	93676035	93676917	PerC transcriptional activator
KN361_4g005506	Chr 4	94037794	94044616	N-acetylglucosaminyltransferase-IV GnT-IV conserved region
KN361_4g005525	Chr 4	94137911	94138498	hypothetical protein
KN361_4g005536	Chr 4	94182318	94183233	glycerol-3-phosphate dehydrogenase (NAD(+)) GPDHC1
KN361_4g005731	Chr 4	95583725	95585770	aspartic proteinase NANA
KN361_4g005745	Chr 4	95656441	95658024	periplasmic sensor domain-containing protein extracellular;major pollen allergen Ole e 10
KN361_4g005754	Chr 4	95704891	95707553	hypothetical protein

KN361_4g005881	Chr 4	96377332	96378052	protease inhibitor/seed storage/LTP family protein
KN361_4g005940	Chr 4	96706707	96707172	hypothetical protein
KN361_4g006129	Chr 4	97850865	97852967	cytochrome P450
KN361_4g006132	Chr 4	97875619	97880850	putative transcription factor bHLH041
KN361_4g006261	Chr 4	98671626	98675810	7-deoxyloganetin glucosyltransferase;UDP-glycosyltransferase 85A8
KN361_4g006262	Chr 4	98676827	98685313	UDP-glycosyltransferase 85A8
KN361_4g006336	Chr 4	99135496	99136581	vacuolar iron transporter 4
KN361_4g006431	Chr 4	99585658	99586681	pentatricopeptide repeat-containing protein
KN361_4g006491	Chr 4	99920286	99921378	pentatricopeptide repeat-containing protein
KN361_4g006543	Chr 4	100187799	100189637	hypothetical protein
KN361_4g006580	Chr 4	100418088	100419726	expansin-A9
KN361_4g006599	Chr 4	100592079	100595459	GRF1-interacting factor 1
KN361_4g006843	Chr 4	102017982	102019865	LOB domain-containing protein 1
KN361_4g006848	Chr 4	102035991	102037770	hypothetical protein
KN361_4g006926	Chr 4	102474751	102475999	putative ribonuclease H protein
KN361_4g007178	Chr 4	103796300	103797311	ethylene-responsive transcription factor ERF106
KN361_4g007196	Chr 4	103867226	103869239	hypothetical protein
KN361_4g007201	Chr 4	103919350	103921307	F-box protein
KN361_4g007206	Chr 4	103939663	103941011	putative aquaporin TIP5-1
KN361_4g007215	Chr 4	103991695	103995252	NAC domain-containing protein 7
KN361_4g007314	Chr 4	104656892	104659805	ADP-ribosylation factor-like protein 13B
KN361_4g007330	Chr 4	104782357	104783839	heavy metal-associated isoprenylated plant protein 31
KN361_4g007365	Chr 4	104977074	104979616	protein PATRONUS 2
KN361_4g007371	Chr 4	105031537	105033060	putative septum site-determining protein minD
KN361_4g007478	Chr 4	105625965	105626990	pentatricopeptide repeat-containing protein
KN361_4g007508	Chr 4	105831583	105837903	SPFH domain-containing protein/Band 7 family protein;hypersensitive-induced response protein 3

KN361_4g007603	Chr 4	106345453	106345917	hypothetical protein
KN361_nc582	Chr 4	2195476	2195606	small nuclear RNA U1
KN361_nc583	Chr 4	2202780	2202961	small nuclear RNA U2
KN361_nc638	Chr 4	64872509	64874659	long non-coding RNA
KN361_nc655	Chr 4	50499632	50501724	long non-coding RNA
KN361_0g000140	Un	2273	4776	hypothetical protein
KN361_0g000156	Un	5	769	hypothetical protein
KN361_0g000210	Un	7205	9040	mitotic spindle checkpoint protein MAD2
KN361_0g000287	Un	4	2241	retrovirus-related Pol polyprotein from transposon RE2
KN361_0g000289	Un	9497	12877	putative late blight resistance protein R1B-16
KN361_0g000292	Un	46975	56462	hypothetical protein
KN361_0g000301	Un	109697	112499	hypothetical protein
KN361_0g000302	Un	112594	114238	gag-polypeptide of LTR copia-type
KN361_0g000304	Un	120921	124003	retrovirus-related Pol polyprotein from transposon RE1
KN361_0g000307	Un	23612	32687	RNA-directed DNA polymerase;hypothetical protein
KN361_0g000313	Un	8644	10663	hypothetical protein
KN361_0g000478	Un	1405	2747	hypothetical protein
KN361_0g000486	Un	18285	21791	ABC transporter G family protein member 9
KN361_0g000495	Un	3105	7590	pyridine nucleotide-disulfide oxidoreductase;hypothetical protein
KN361_0g000517	Un	5511	6099	hypothetical protein
KN361_0g000518	Un	9303	10943	hypothetical protein
KN361_nc730	Un	21	999	long non-coding RNA



Supplementary Figure S1. *P. ovata* WT developing fruits from anthesis to 23 DPA.



Supplementary Figure S2. *P. ovata* WT developing seeds (two sets per development stage) from 3 to 29 DPA.

Chapter 5

Screening gamma-irradiated putative mutants for higher mucilage yields

ꕥ•ꕥ

Introduction

Seeds from the myxospermous species *Plantago ovata* release a heteroxylan-enriched mucilage when exposed to moisture. The mucilage is the hydrated husk, commonly called psyllium, that has been used for many health, food, and industrial applications (Phan et al., 2020; Cowley et al., 2021). This plant has also been promoted as a model species for studying heteroxylan biosynthesis (Jensen et al., 2013; Jensen et al., 2014; Phan et al., 2016; Tucker et al., 2017). Despite the importance of *P. ovata* in the global market and its potential as a model system for plant cell wall research, seed, and therefore mucilage/husk production is limited or negatively affected by adverse environmental conditions (Karimzadeh and Omidbaigi, 2004; Dhar et al., 2005; Cowley et al., 2022). Progress in *P. ovata* breeding programs in generating well-adapted cultivars with improved yield and resistance to biotic and abiotic stress is hindered by low genetic diversity or variation (Dhar et al., 2005; Singh et al., 2009; Fougat et al., 2014). Variation can be achieved by exploring wild relatives (Phan et al., 2016; Cowley and Burton, 2021) or by induced mutation (Lal et al., 2020).

Induced mutation has been reported to improve seed propagated crops, including rice, wheat, barley and beans and vegetatively propagated ornamental plants such as chrysanthemum and rose (Ahloowalia and Maluszynski, 2001). DNA mutation can be achieved by chemical treatments or gamma radiation (Ahloowalia and Maluszynski, 2001). Gamma irradiation successfully induced mutation of *P. ovata* while EMS was not optimal (Dhar et al., 2005; Tucker et al., 2017; Lal et al., 2020). Many studies have used Arabidopsis mutants with a reduced amount and/or altered mucilage composition to identify the genes involved (Western et al., 2001; Voiniciuc et al., 2018). This study focuses on selecting a mutant line with a higher mucilage yield than wild type (WT) but with no observed pleiotropic phenotype, for analysis to help identify genes related to mucilage production, as described in Chapter 6. The candidate mutant was selected from the gamma-irradiated mutant population described by Tucker et al. (2017).

Material and Methods

Plant materials and sample collections

The 201 gamma-irradiated *P. ovata* putative mutant lines used in this screening were generated as described by Tucker et al. (2017) and had been propagated by selfing to the M5 generation. Eight selected mutant lines and one wild type with ten replicates were grown in the field at the Frank Wise Institute, Kununurra, Western Australia, in 2019. One candidate mutant 970-1 (*raya*) and the wild type with three replicates were grown in a growth chamber at the University of Adelaide, with a controlled temperature of 23°C and photoperiod 16h light and 8h dark. Harvested seeds were stored in the oven at 37°C for 72h before future experiments. An overview of the screening process can be seen in Figure 1.

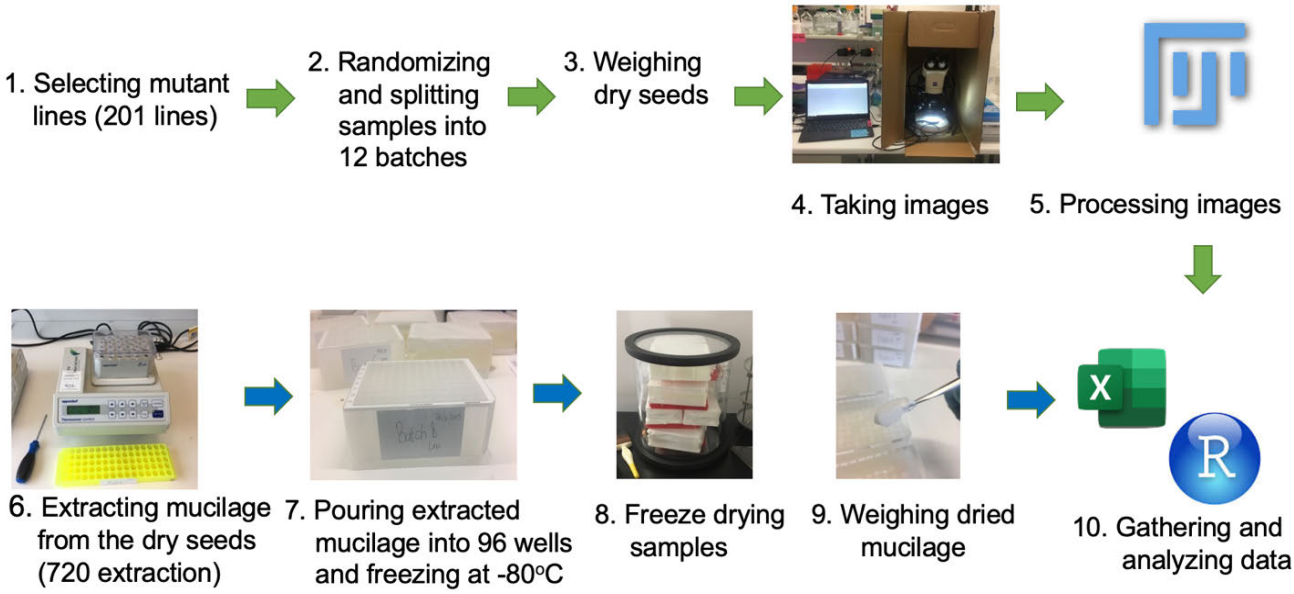


Figure 1. Steps for screening candidate gamma-irradiated mutant lines for a higher mucilage yield genotype.

Seed and plant measurements

Before mucilage extraction, twenty seeds for each mutant line were weighed using an analytical balance and photographed under a dissecting microscope (Zeiss Stemi 2000-C, Germany) equipped with a colour digital camera (Zeiss AxioCam ERc 5s, Germany). The seed

measurements were processed using Fiji ImageJ v1.51. The measured parameters included weight (mg), area (mm^2), diameter (mm), perimeter (mm), circularity, and roundness. The dried weights of field trial plants and seed yield were also taken to obtain their biomass (mg).

Seed staining

Seeds were stained with ruthenium red (0.01% w/v) without prior soaking or directly added onto a microscope slide (Rowe GM2715, Australia) with a cover slip before being photographed (Cowley et al., 2020).

Mucilage extraction and monosaccharide profile

Mucilage seed extraction and monosaccharide analysis have been described previously in Cowley et al. (2020). There are three main products of the extraction pipeline, namely cold water-extractable (CWE), hot water extractable (HWE) and intense extraction resistant (IAE) fractions. All three separate mucilage extractions were analysed for monosaccharide profiles using reverse-phase high performance liquid chromatography (RP-HPLC) of 1-phenyl-3-methyl-5-pyrazoline (PMP) derivates. Major monosaccharides were measured, including rhamnose, xylose, and arabinose.

Data analysis

Data collected during this study included mucilage yield (mg/mg dry weight), mucilage yield per individual seed (mg), mucilage yield per individual plant (mg), average seed weight (mg), seed size parameters, seed weight per plant, plant weight and monosaccharide composition. Correlation analysis was used to find the relationship between mucilage yield and seed size parameters. Principal Component Analysis was used to reduce the number of variables and cluster mutant lines according to their similarities. T-test and Anova one-way analysis were used to compare wild-type and mutant line traits.

Results

Seeds of mutant lines and wild type were analysed for a range of seed traits by measuring their weight, analysing seed images, and using biochemical techniques to identify lines with increased mucilage yield, not simply proportional changes due to seed size modifications.

Plantago ovata seed measurements

Figure 2A shows that a large proportion (78%) of gamma-irradiated mutant lines (201) had a reduction in seed weight at or below 1.97 mg per seed, which is the WT seed weight. This trend was also observed as a reduction in mucilage yield per seed (Figure 2B). About 68% of the mutants had reduced mucilage amounts whilst 31% had an increased mucilage amount, with the WT mucilage yield at about 0.44 mg per seed (Figure 2B). The proportion of mucilage yield per dry seed weight shows the opposite trend. More mutant lines have a higher proportion of mucilage, about 73% of the population (Figure 2C). The WT mucilage yield is about 23% of seed weight (Figure 2C).

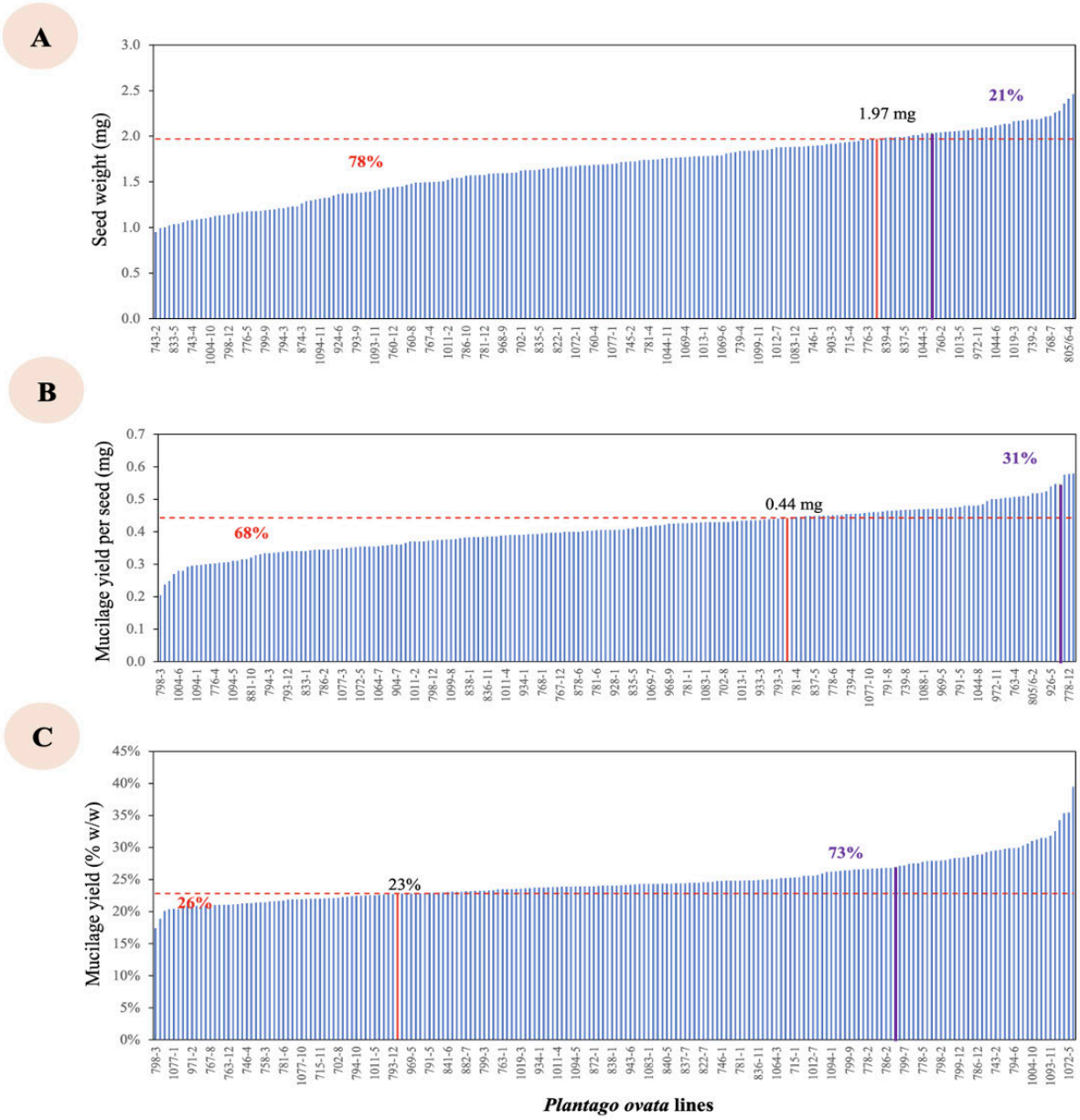
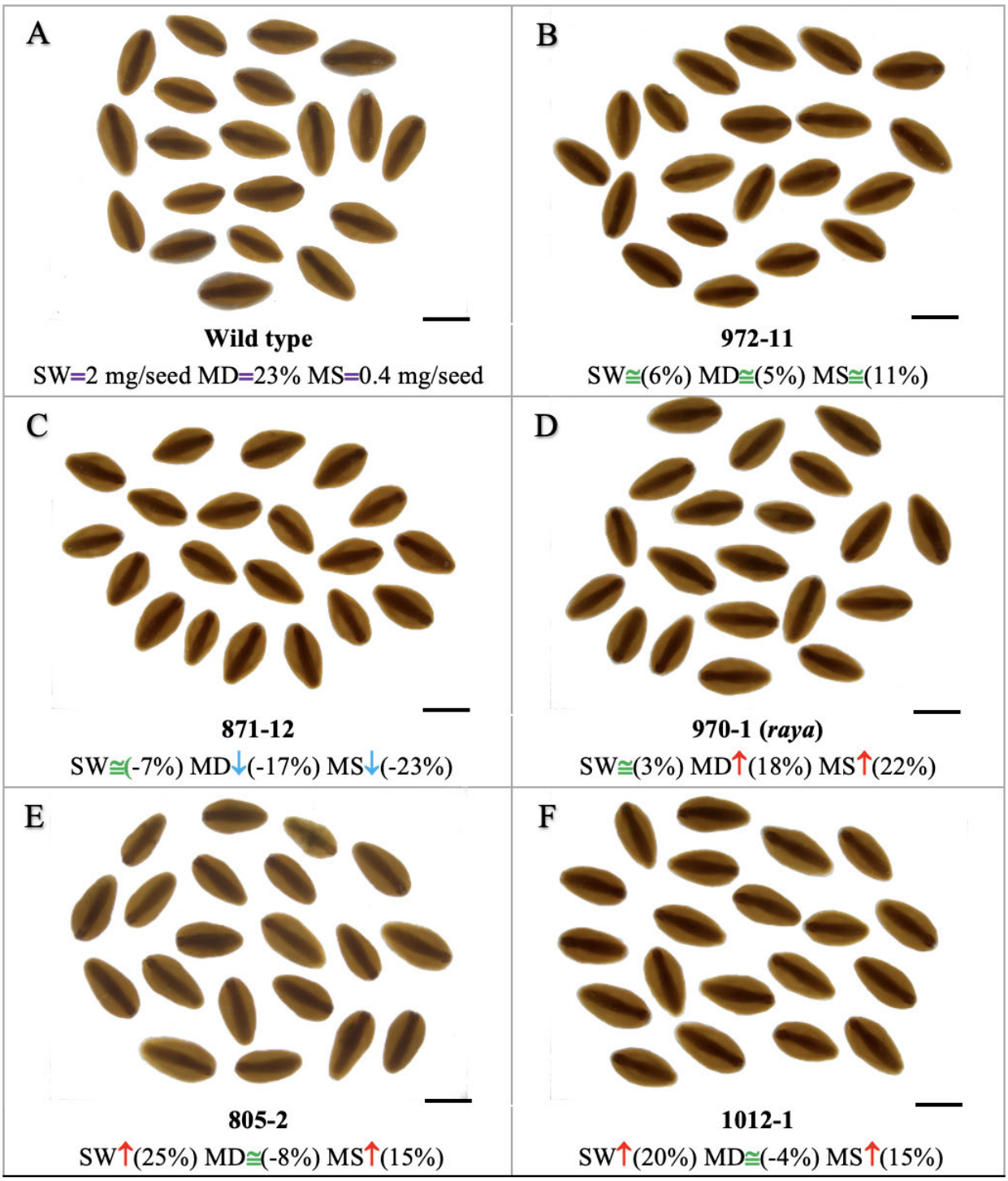
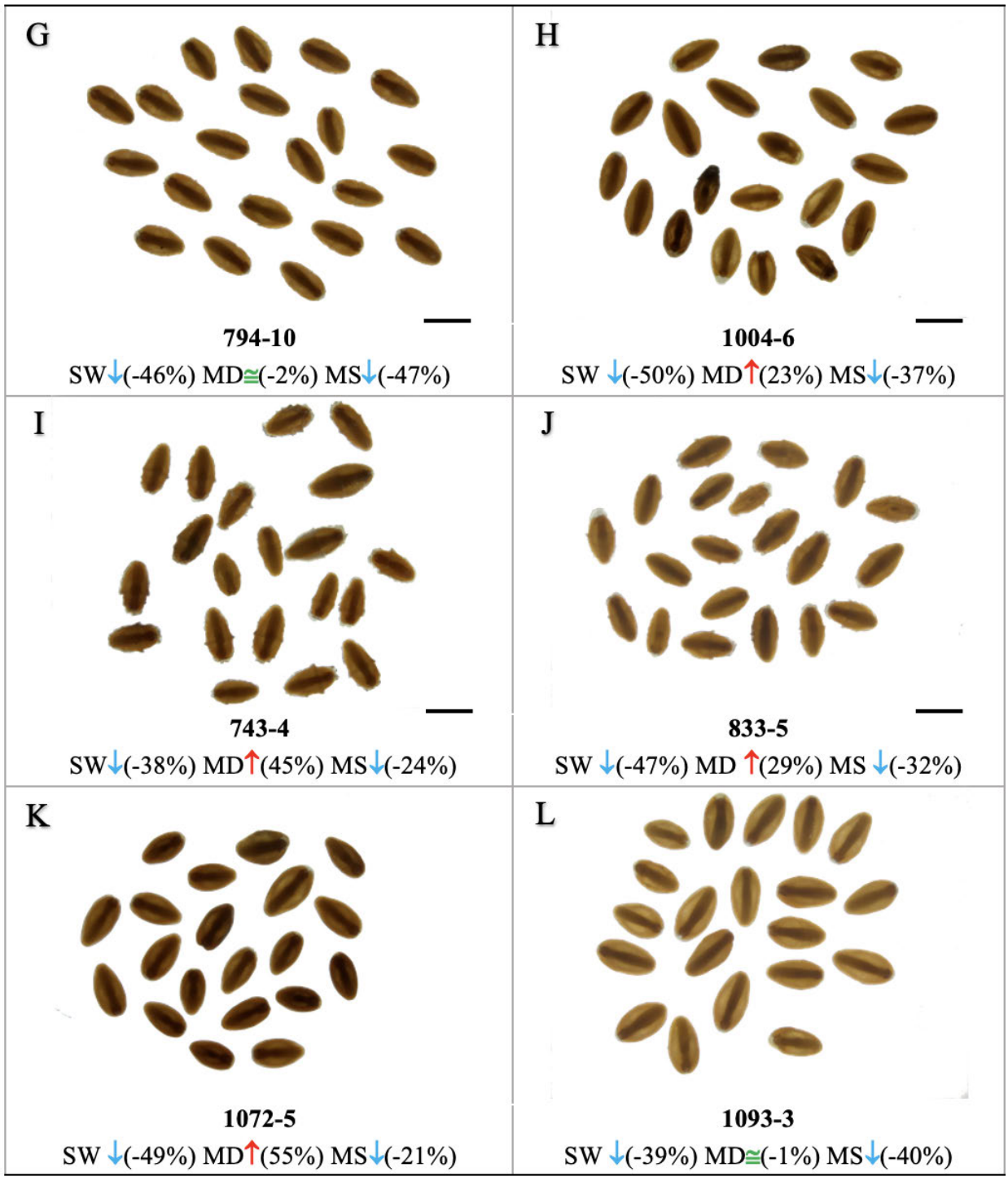


Figure 2. Comparisons between seed weight (A), mucilage yield per seed (B) and mucilage yield per dry seed weight (C). The red bar with value above on each graph indicates WT and the purple bar refers to candidate putative mutant *ray*. Percentages of putative mutant lines above baseline were coloured purple while below the line were coloured red

Before performing mucilage extraction, all seeds from WT and putative mutant lines were photographed to capture phenotypic variation in seed dimensions. A representative sample from 18 individual mutants is presented in Figure 3. Lines 972-11 (Figure 3B), 871-12 (Figure 3C),

and 970-1 (Figure 3D) are difficult to distinguish from the WT (Figure 3A) visually. Line 972-11 has similar seed weight and mucilage yield to the WT. Line 871-12 has a lower mucilage yield, while line 970-1 produces a higher amount of mucilage than WT. No defect was observed in 805-2 and 1012-1 (Figures E and F), which have heavier seeds. It was noted that seed phenotypes like colour and shape are more likely to change when the seed weight decreases (Figures 3G – 3R). The abnormal shape with an apparent horizontal line given as an example in Figure 3I is likely to have arisen as the seed was crushed in the dehiscence zone on the capsule (Chapter 3). This abnormality was observed in line 794-10 (Figure 3G), some seeds of 1004-6 (Figure 3H), and many seeds of 743-4 and 833-5 (Figures 3I and 3J, respectively). Lines 1004-6 and 743-4 have darker seed colours, while the seeds from line 833-5 are yellowish. Darker seed with a standard shape but lighter in weight was found for lines 1072-5 (Figure 3K) and 803-7 (Figure 3M), while yellowish seed with a standard shape but lighter in weight was detected in 1093-3 (Figure 3L). One mutant line, 1088-1 (Figure 3N), has four different seed phenotypes (Figures 3O-3R) where 1088-1a has a wild-type-like appearance and seed weight but has more mucilage than WT; 1088-1b is lighter in weight, darker in colour and produces more mucilage than WT; 1088-1c seeds were not fully developed with a more prominent transparent layer and smaller embryo and endosperm parts. The seeds were extraordinarily light in weight and the proportion of mucilage was exceptionally high (Figure 3Q). Meanwhile, mucilage per seed was still relatively similar to the WT (Figure 3A). The last type, 1088-1d (Figure 3R), has a reduced seed weight and greenish colour, but the seeds produce more mucilage.


A
Wild type
SW=2 mg/seed MD=23% MS=0.4 mg/seed
B
972-11
SW≅(6%) MD≅(5%) MS≅(11%)
C
871-12
SW≅(-7%) MD↓(-17%) MS↓(-23%)
D
970-1 (raya)
SW≅(3%) MD↑(18%) MS↑(22%)
E
805-2
SW↑(25%) MD≅(-8%) MS↑(15%)
F
1012-1
SW↑(20%) MD≅(-4%) MS↑(15%)


G
794-10
SW↓(-46%) MD≅(-2%) MS↓(-47%)
H
1004-6
SW ↓(-50%) MD↑(23%) MS↓(-37%)
I
743-4
SW↓(-38%) MD↑(45%) MS↓(-24%)
J
833-5
SW ↓(-47%) MD ↑(29%) MS ↓(-32%)
K
1072-5
SW ↓(-49%) MD↑(55%) MS↓(-21%)
L
1093-3
SW ↓(-39%) MD≅(-1%) MS↓(-40%)

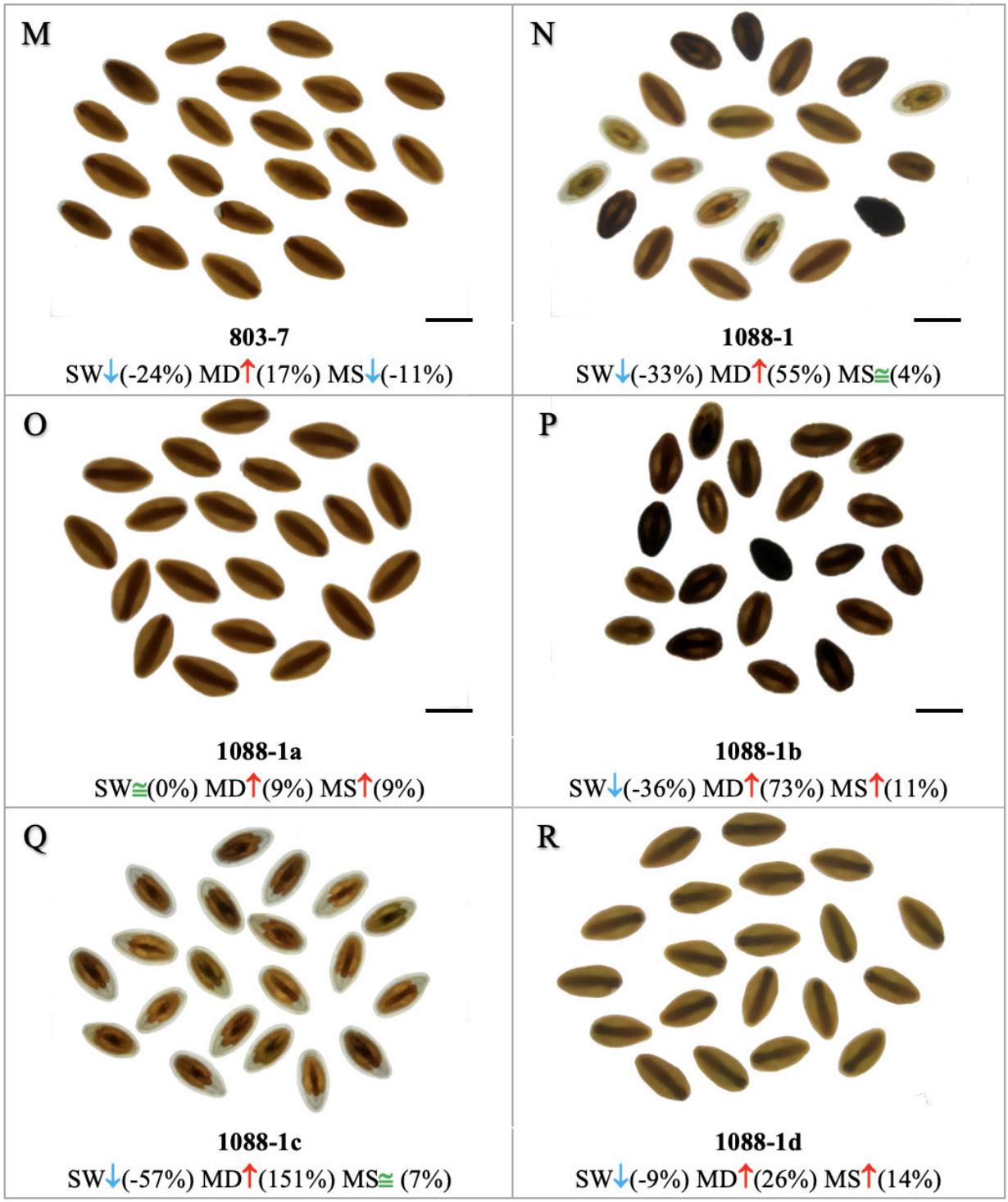
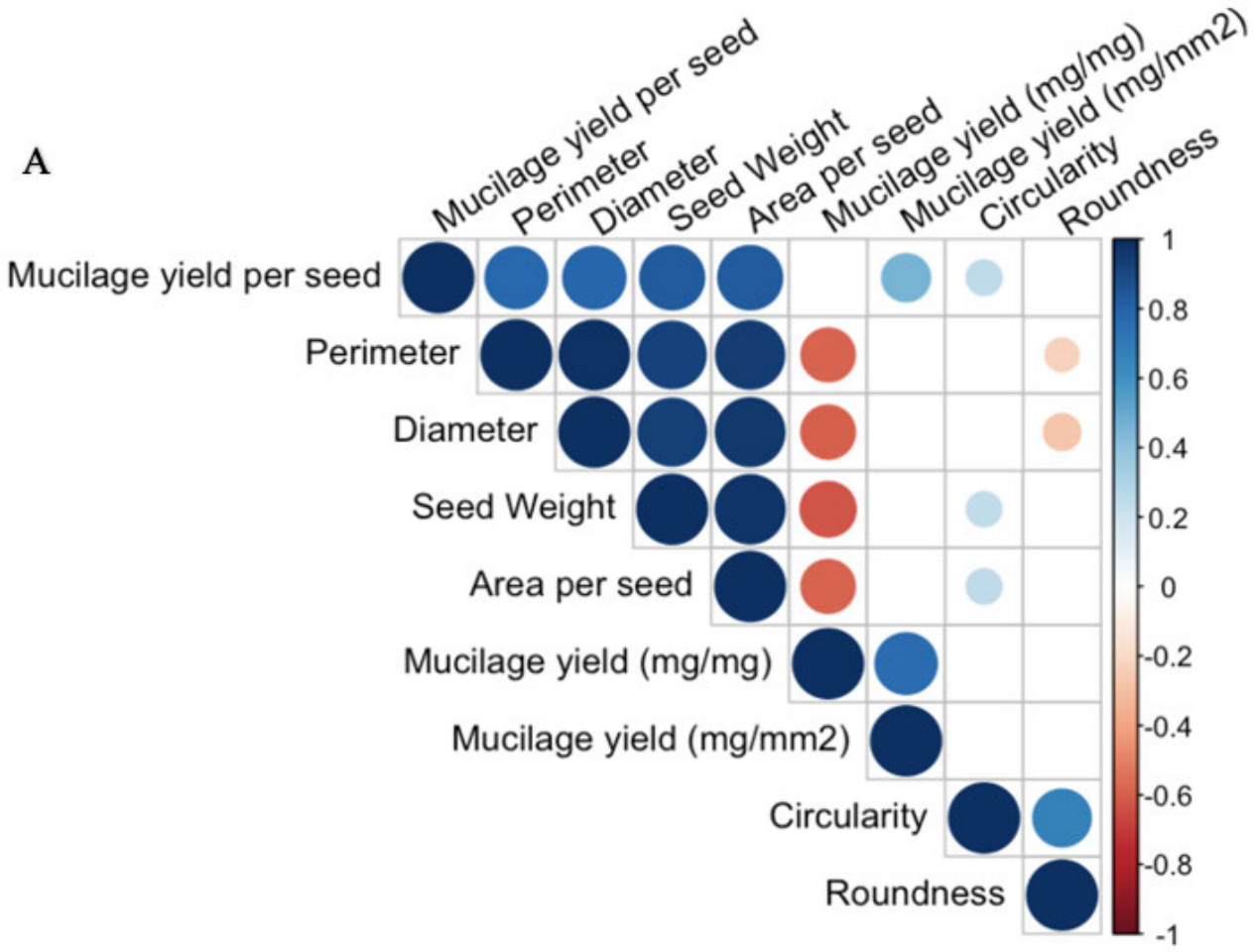

M
803-7
SW↓(-24%) MD↑(17%) MS↓(-11%)
N
1088-1
SW↓(-33%) MD↑(55%) MS≅(4%)
O
1088-1a
SW≅(0%) MD↑(9%) MS↑(9%)
P
1088-1b
SW↓(-36%) MD↑(73%) MS↑(11%)
Q
1088-1c
SW↓(-57%) MD↑(151%) MS≅ (7%)
R
1088-1d
SW↓(-9%) MD↑(26%) MS↑(14%)

Figure 3. Representative seed images that were processed using ImageJ. WT and line names are under each image with SW = Seed weight, MD = Mucilage yield per dry seed weight, MS = Mucilage yield per seed. A blue downward arrow represents measurements of mutant lines lower than the wild-type value while a red upwards arrow means the mutant measurement is higher than the wild type. A green symbol shows that the mutant and WT values are equal. Scale bar = 2 mm.

Univariate and multivariate analysis on seed parameters

Correlation analysis was performed on ten seed measurements (Figure 4). Of these ten, three measurements are shown in Figure 2 and seven variables were obtained from image analysis are perimeter, diameter, area per seed, circularity, and roundness. One variable obtained indirectly from the image is mucilage yield per area. Mucilage yield per seed has a positive and robust correlation with seed weight (R=0.84), area (R=0.84), perimeter (R=0.78), and diameter (R=0.79) (Figure 4A and 4B). It also has a positive but weak correlation with mucilage yield per area (R=0.47) and circularity (R=0.25). No significant correlation was seen between mucilage yield per seed and roundness (P=0.73) and mucilage yield (mg/mg) (P=0.12) (Figures 4A and B). In contrast, mucilage yield per seed weight was mostly negatively correlated with seed dimension variables (Figures 4A and C). The coefficient correlation (R) with a $P<0.01$ ranges from -0.59 (diameter) to -0.62 (seed weight). Mucilage yield per area positively correlates with mucilage yield per seed (R=0.47) and dry seed weight (R=0.77) (Figure 4A). Besides mucilage yield per seed, mucilage yield per dry seed weight does not significantly correlate with roundness (P=0.41) (Figure 4A, B, and C).

A



B

Variable	R	*P*<0.01
Seed weight (mg)	0.84	0.00
Area (mm^2)	0.84	0.00
Perimeter (mm)	0.78	0.00
Circularity	0.25	0.00
Diameter (mm)	0.79	0.00
Roundness	-0.02	0.73
Mucilage yield (mg/mg)	-0.11	0.12
Mucilage yield per area	0.47	0.00

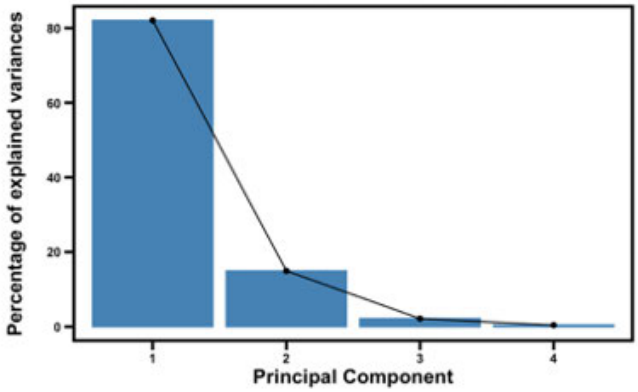
C

Variable	R	*P*<0.01
Seed weight (mg)	-0.62	0.00
Area (mm^2)	-0.59	0.00
Perimeter (mm)	-0.58	0.00
Circularity	-0.05	0.49
Diameter (mm)	-0.59	0.00
Roundness	0.06	0.41
Mucilage yield per seed	-0.11	0.12
Mucilage yield per area	0.77	0.00

Figure 4. Correlations between seed measurements visualised using Corrplot in R (A) and tables shows correlation coefficient and p-value between mucilage yield per seed and other variables (B) and between mucilage yield per dry weight and other variables (C).

While a correlation plot helps investigate pairwise correlations, PCA multivariate analyses help reduce data dimensions with more than two variables (Figure 5). Ten variables were reduced into two principal components (PC1 and PC2) (Figures 5A and B). These two PCs explain 97% of the variation in the data, with PC1 at 82.1% and PC2 at 15% (Figure 5). Principal Component 1 has an inverse correlation with seed weight, area, perimeter, diameter and mucilage yield per seed (Figure 5B). A positive correlation was seen only with mucilage yield per dry seed weight (Figure 5B). Only mucilage yield per seed and mucilage yield per dry seed weight correlate with PC2, and they have a positive correlation (Figure 5B). Variables of seed weight, perimeter, diameter, and area per seed overlap (Figure 5C). They are both aligned with PC1, so seed weight can be used as a representation of seed dimension (Figure 5C). Thus, there are three different directions of variables to the principal components (Figure 5 C).

A



B

Variables	PC1 (82.1%)	PC2 (15%)
Seed weight (mg)	-0.98	0.00
Area (mm^2)	-0.99	0.00
Perimeter (mm)	-0.97	0.00
Diameter (mm)	-0.98	0.00
Mucilage yield per seed	-0.83	0.54
Mucilage yield (mg/mg)	0.62	0.78

C

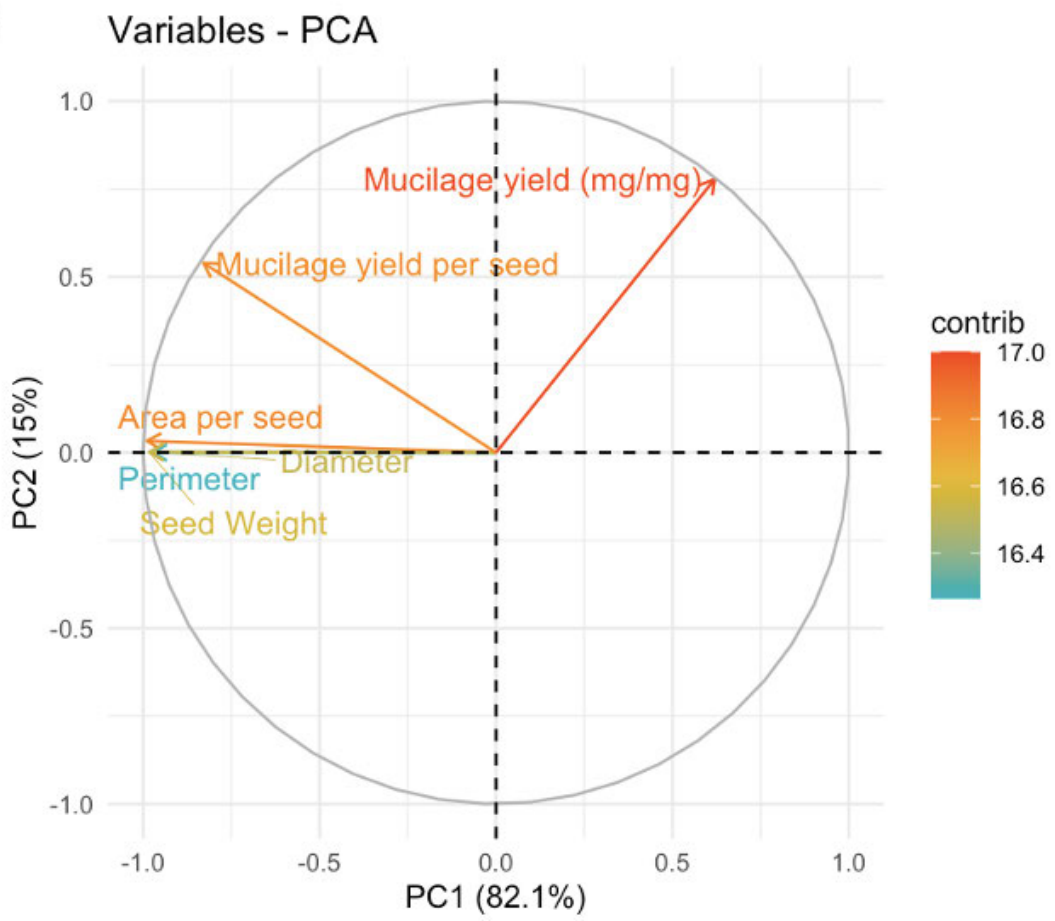


Figure 5. Two Principal Components 1 and 2 (PC1 and PC2) explain more than 97% of the variation in *P. ovata* seed measurements using the factoextra library package in R.

Grouping samples separately according to their weight and mucilage yield

Individual samples (WT and mutant lines) were visualised in PCA plots with three different coloured groups (Figure 6). Figure 6A is coloured according to seed weight, Figure 6B by mucilage yield per seed and Figure 6C by mucilage yield per dry seed weight. Each group's numbers and cut-offs were decided by comparing raw data in Microsoft Excel and PCA plots. Samples with similar profiles were grouped. Each variable (seed weight, mucilage yield per seed and dry weight) has four groups, from the lowest in Group 1 to the highest in Group 4, by comparing the mutant lines to the WT values for each measurement (Table 1 and Figure 6). The group containing mutant lines with similar profiles to the wild type were called the standard group. Group 3 (standard) was assigned for seed weight and mucilage yield per seed, while Group 2 (standard) is in mucilage yield per dry weight. Group 1 in the seed weight group has the lightest seeds (less than -27% of wild-type weight) and does not produce mucilage yield per seed more than the standard group (Table 1). Nevertheless, they produce similar or more mucilage per dry weight than the standard (Table 1).

Table 1. Twenty-two combinations of three different groups observed in the mutant population.

Seed weight	Group	Mucilage yield/seed	Group	Mucilage/dry weight	Group
1. less than -27%	Light	1. less than -21%	Lower	2. -10% to 12%	Standard
1. less than -27%	Light	2. -21% to -8%	Low	2. -10% to 12%	Standard
1. less than -27%	Light	1. less than -21%	Lower	3. 12% to 32%	High
1. less than -27%	Light	2. -21% to -8%	Low	3. 12% to 32%	High
1. less than -27%	Light	1. less than -21%	Lower	4. more than 32%	Higher
1. less than -27%	Light	2. -21% to -8%	Low	4. more than 32%	Higher
1. less than -27%	Light	3. -8% to 8%	Standard	4. more than 32%	Higher
2. -27% to -8%	Light	1. less than -21%	Lower	1. less than -10%	Low
2. -27% to -8%	Light	1. less than -21%	Lower	2. -10% to 12%	Standard
2. -27% to -8%	Light	2. -21% to -8%	Low	2. -10% to 12%	Standard
2. -27% to -8%	Light	3. -8% to 8%	Standard	2. -10% to 12%	Standard
2. -27% to -8%	Light	2. -21% to -8%	Low	3. 12% to 32%	High
2. -27% to -8%	Light	3. -8% to 8%	Standard	3. 12% to 32%	High
2. -27% to -8%	Light	3. -8% to 8%	Standard	4. more than 32%	Higher
3. -8% to 15%	Standard	1. less than -21%	Lower	1. less than -10%	Low
3. -8% to 15%	Standard	3. -8% to 8%	Standard	1. less than -10%	Low
3. -8% to 15%	Standard	2. -21% to -8%	Low	2. -10% to 12%	Standard
3. -8% to 15%	Standard	3. -8% to 8%	Standard	2. -10% to 12%	Standard
3. -8% to 15%	Standard	4. more than 8%	High	2. -10% to 12%	Standard
3. -8% to 15%	Standard	3. -8% to 8%	Standard	3. 12% to 32%	High
3. -8% to 15%	Standard	4. more than 8%	High	3. 12% to 32%	High
4. more than 15%	Heavy	4. more than 8%	High	2. -10% to 12%	Standard

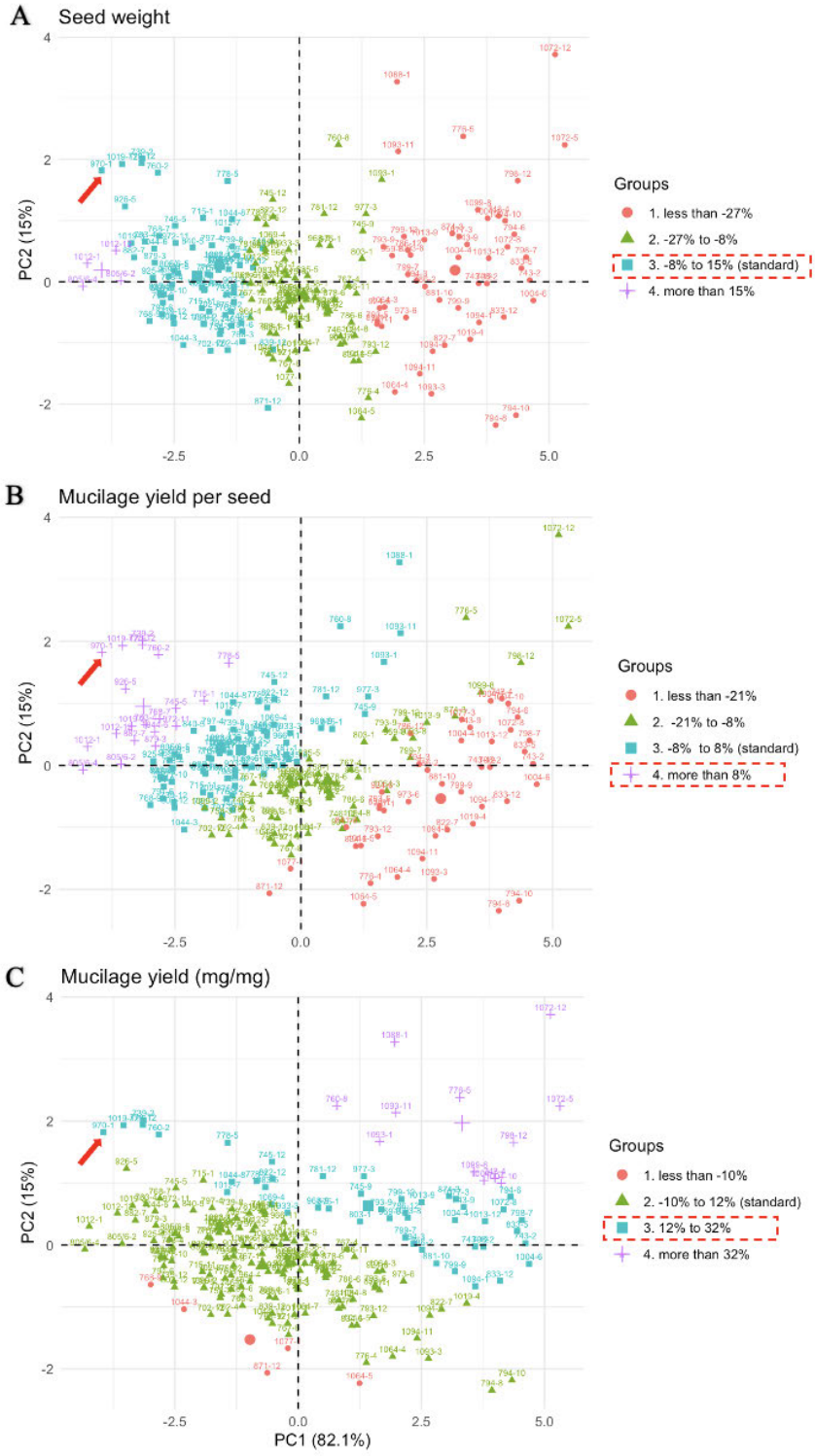


Figure 6. Three PCA plots represent three different groupings based on (A) seed weight, (B) mucilage yield per seed and (C) mucilage yield per dry seed weight. Red arrows indicate *ray* samples.

Similarly, Group 2 in the seed weight group shows a seed reduction between -27% to -8%. They do not have a mucilage yield per seed of more than 8% (Group 4) (Table 1). However, members of Group 2 seed weight were recorded to include all groups of mucilage yield per dry weight from the lowest (Group 1, less than -10%) to the highest (Group 4, more than 32%) (Table 1). Even though mutant lines have a similar seed weight to the wild type (Group 3, -8% to 15%), they have varying seed mucilage yields (Table 1), varying from less than -21% (Group 1) to more than 8% (Group 4) (Table 1). However, their mucilage yield per dry weight is only a maximum of 32% of the WT value. Some giant mutant seeds (Group 4, more than 15%) have a mucilage yield per seed of more than 8% (Group 4), but the mucilage yield per dry weight is similar to the WT (Group 2, standard) (Table 1).

Mucilage extrusion pattern stained with ruthenium red does not reflect mucilage amount

Variations of mucilage extrusion patterns can be found in the gamma-irradiated mutant lines (Figure 7). However, mucilage patterns are not accurate for predicting mucilage amount. Lines 743 and 1072-12 are in a similar seed weight group (Table 2), at less than -27% of seed weight. Both mutant lines produce more mucilage than the WT, at 38% extra for line 743 and 73% more for 1072-12 but they both release less mucilage per seed than the WT (less than -21%, Table 2). However, their mucilage extrusion patterns are distinct (Figures 7B and G). The pattern for 743-4 (Figure 7B) is more similar to the patterns of lines 834-6 (Figure 7D) and 1064-5 (Figure 7F), which have a spiky outer layer. However, lines 834-6 and 1064-5 have less mucilage than the WT (Table 2). The staining pattern of 768-9 (Figure 7C) was similar to the wild type (Figure 7A). A wild-type pattern was also observed for lines 871-12, 760-2, 778-12, 970-1, 1012-7, 1069-4, and 1083-5 (Figure 7E, J, K, L, N, and P, respectively), but not all lines shared the same seed weight and mucilage yield groups (Table 2). Only 760-2, 778-12 and 970-1 have wild-type seed size and high mucilage yield per seed and dry weight (Table 2). The

mucilage extrusion pattern from 1088-1c seeds (Figure 3Q and 7H) was unique. The mucilage envelope is square with thicker layers at the top and bottom "edges" (Figure 7H)

Table 2. Seed measurement and groups for selected mutant lines presented in Figure 7.

Lines	Seed weigh	Seed weight group	mg/seed	mg/seed group	mg/mg	mg/mg group
743-4	-45.24%	1. less than -27%	-24.40%	1. less than -21%	38%	4. more than 32%
1072-12	-47.96%	1. less than -27%	-10.00%	2. -21% to -8%	73%	4. more than 32%
1064-5	-25.50%	2. -27% to -8%	-34.40%	1. less than -21%	-12%	1. less than -10%
834-6	-23.62%	2. -27% to -8%	-25.60%	1. less than -21%	-2%	2. -10% to 12%
1069-4	-10.10%	2. -27% to -8%	1.20%	3. -8% to 8%	13%	3. 12% to 32%
1083-5	-9.54%	2. -27% to -8%	4.10%	3. -8% to 8%	15%	3. 12% to 32%
871-12	-7.28%	3. -8% to 15%	-23.30%	1. less than -21%	-17%	1. less than -10%
768-9	14.57%	3. -8% to 15%	2.20%	3. -8% to 8%	-10.70%	1. less than -10%
1012-7	-4.77%	3. -8% to 15%	6.70%	3. -8% to 8%	12.10%	3. 12% to 32%
1044-8	-6.60%	3. -8% to 15%	7.00%	3. -8% to 8%	15%	3. 12% to 32%
778-12	10.91%	3. -8% to 15%	28.30%	4. more than 8%	16%	3. 12% to 32%
760-2	3.55%	3. -8% to 15%	21.70%	4. more than 8%	18%	3. 12% to 32%
970-1	3.38%	3. -8% to 15%	21.90%	4. more than 8%	18%	3. 12% to 32%

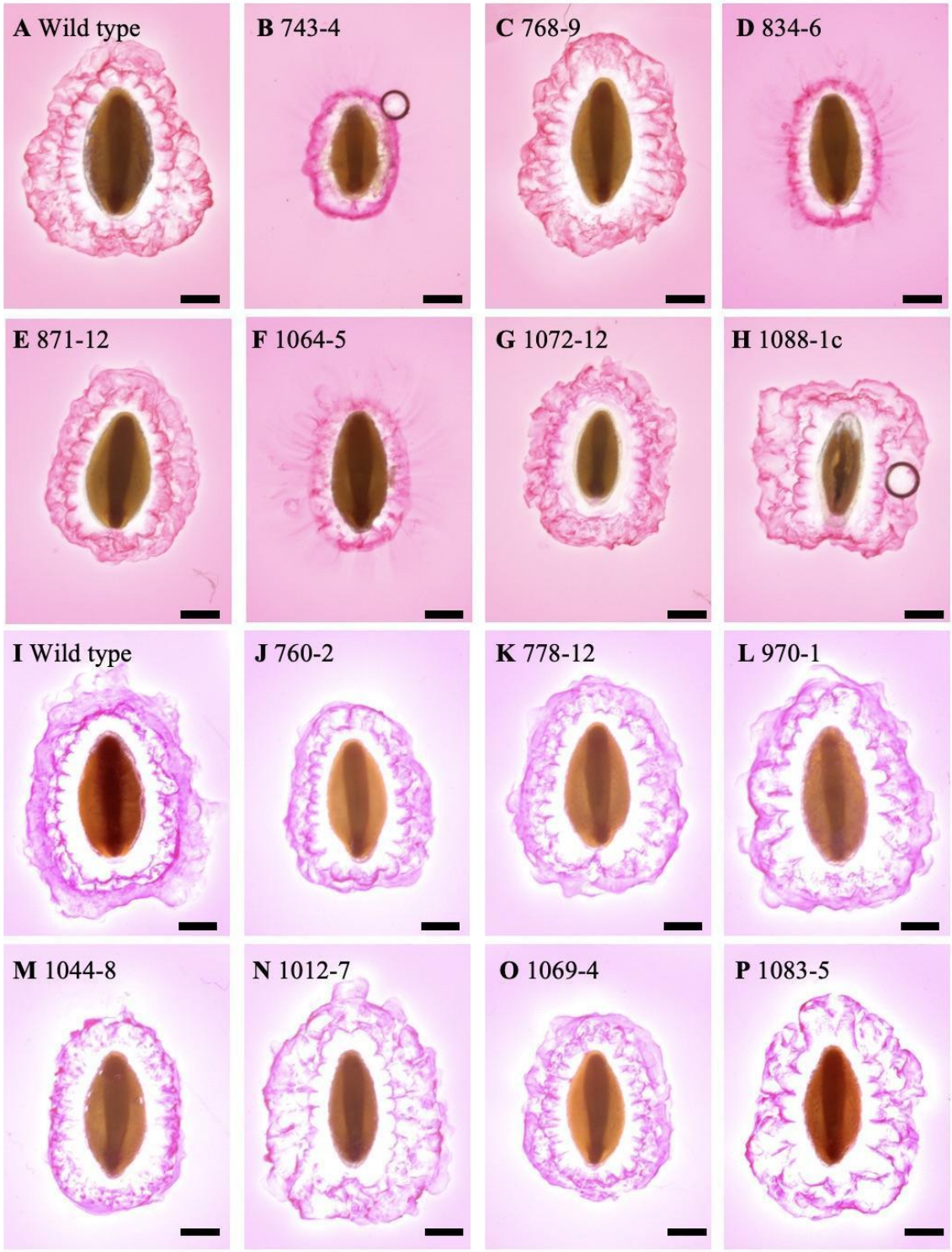


Figure 7. Variation in mucilage extrusion patterns from the mature seeds of *P. ovata* WT and gamma irradiated mutant lines. Seeds were stained using ruthenium red (0.01% w/v). Images A – H and I – P were stained using different stock solutions. Scale bars = 1 mm.

Observation of wild-type and mutant lines grown in the field at Kununurra

Figure 8 shows variation in total seed weight per plant (seed yield) between WT and eight selected putative mutants. Lines in Figure 8A were coloured by seed weight group while in Figure 8B by two pieces of mucilage yield group information from the previous grouping system with a modification for seed weight classification (Table 1). Groups 1 and 2 in the seed weight classification were merged into the group called "seed light" (Table 1). Overall, the selected mutant lines have higher average seed yields compared to the WT (P=0.0099). Plants from six lines produce greater total seed weight than the WT, while two were indistinguishable. The six lines are 768 (****, $P \leq 0.0001$), 871 (**, $P \leq 0.01$), and 794, 1072, 1012 and 778 (*, $P \leq 0.05$). They have different individual seed weight groups consisting of lighter (lines 794 and 1072), standard (768, 778, and 871), and heavier (1012) seeds than the WT. These six mutant lines also have varied mucilage groups. Based on the mucilage per dry weight group, 768 and 871 are in the low group, 794 and 1012 belong in the standard group, 778 is in the high group and 1072 in the higher or highest group. Based on mucilage per seed, lines 871 and 794 belong to the group having lower or lowest mucilage yield, 1072 is in the low mucilage group, 768 is in the standard group and 778 and 1012 show the highest mucilage amount. The line with the highest seed yield, 768, has a standard seed weight and mucilage yield per seed but a low mucilage per dry seed weight (Figure 8). Only lines 970 and 1004 have no significant differences in seed yield per plant compared to the WT (Figure 8). Line 970 showed little variation (0.6 ± 0.11 gram) among the ten plants measured, while 1004 individuals showed huge differences (1.17 ± 1.13 grams) in seed yield. In addition, both lines had different seed weights and mucilage yield classes. Line 1004 had a lighter seed and lower mucilage yield per dry seed but higher mucilage per dry seed weight. Line 970 had a standard seed weight but high mucilage yield per seed and dry seed weight (Figure 8).

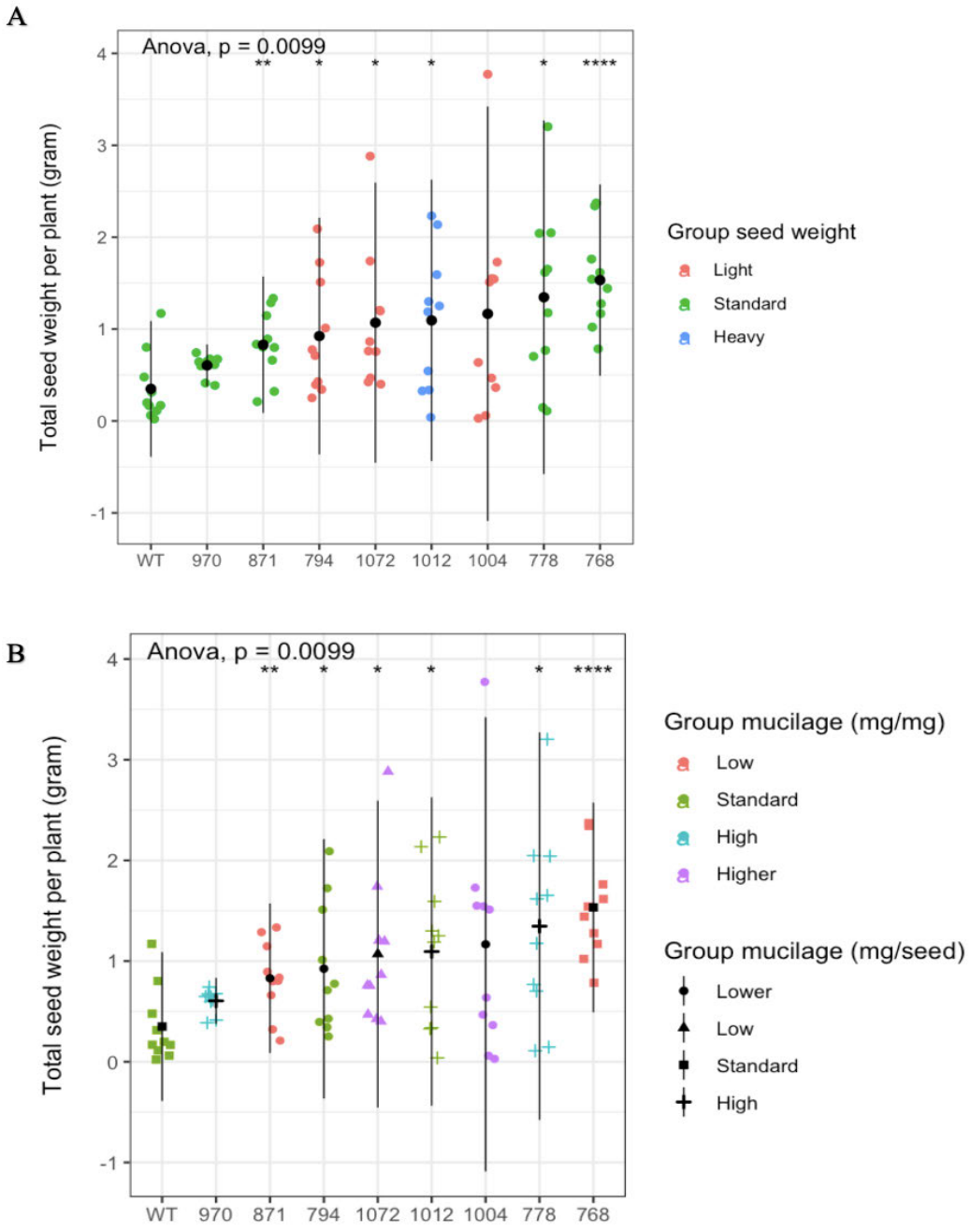


Figure 8. Comparison of total seed weight per plant from nine lines grown in the field. The seed weight for each line is visualised by colour (A), and their groups relating to mucilage per dry weight and per seed are also visualised using different colours and shapes, respectively (B). Genotypes were ordered from the lowest to the highest plant weight. n = 10 biological replicates.

Plant weight (dry biomass) measurements (Figure 9) varied between WT and the eight selected putative mutants ($P = 0.0035$), where selected mutant lines (4.33 – 11.91 grams) had a higher average plant weight than WT (3.21 grams). However, only three out of eight mutant lines differed significantly from wild-type plant weight. They were 768 (**, $P \leq 0.01$), 778 and 794 (*, $P \leq 0.05$). Line 768 had an average or standard weight per seed (2.26 mg) and mucilage yield per seed (0.46 mg) but low mucilage yield per dry weight (20.4%). Line 778 had standard seed weight (1.82 mg) but high mucilage per seed (0.51 mg) and dry seed weight (27.8%). In contrast, line 794 had a light seed (1.06 mg), low mucilage yield per seed (0.24 mg) and standard mucilage yield per dry weight (22.4%). The remaining five mutant lines had a relatively similar weight to the WT (~2 mg). However, they spanned different groups of mucilage yield. For example, lines 970 and 871 had the same standard seed weight but the opposite mucilage yield. Line 970 produced a high amount of mucilage per dry weight (26.9%) and even higher per seed (0.55 mg) compared to the WT (22.8%, 0.45 mg), while line 871 had a low amount per dry weight (18.9%) and even lower per seed (0.35 mg).

Of all the lines that were screened, 970 was chosen as the line that produced consistently larger amounts of mucilage than the WT. This line was called *raya*, which means greater in Indonesian.

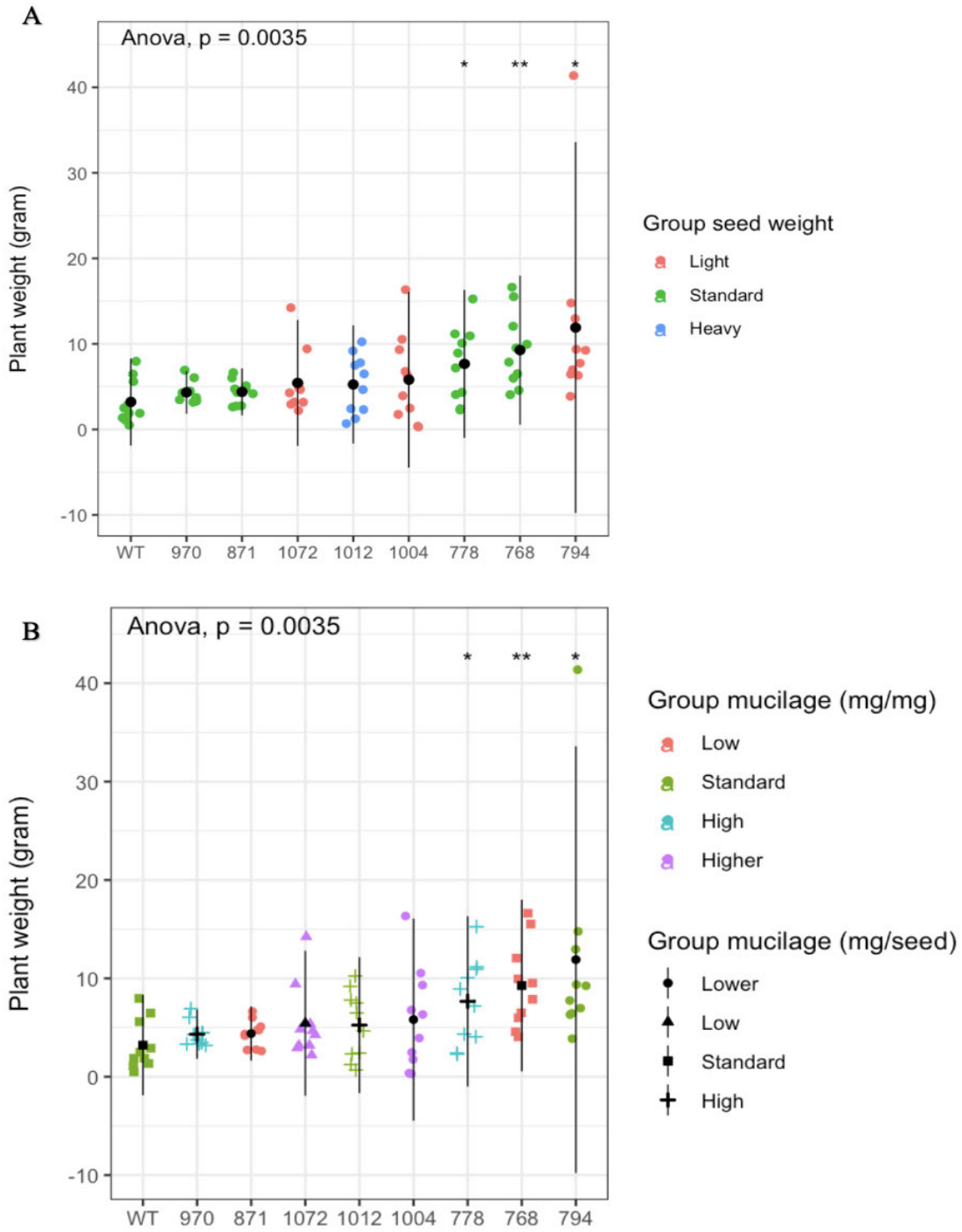


Figure 9. Comparison of dry plant weight from nine lines grown in the field. The seed weight for each line is visualised by colour (A), and the groups reflecting mucilage per dry weight and per seed are visualised using different colours and shapes, respectively (B). Genotypes were ordered from the lowest to the highest plant weight. n = 10 biological replicates for each line.

Phenotypic comparisons between wild-type and *ray*

Wild-type and *ray* plants grew healthily, and no distinct phenotypes were observed (Figure 10). There were no significant differences between plant dry weight (above-ground biomass), seed yield and harvest index (Figure 10). Overall, the mucilage yield of *ray,* grown either in the growth chamber or in the field, was higher than the WT grown at the same time and under the same conditions. However, mucilage yield (mg/mg) was significantly higher from plants grown in the field while mucilage yield per seed was significantly higher from plants grown in the growth chamber (Figure 11).

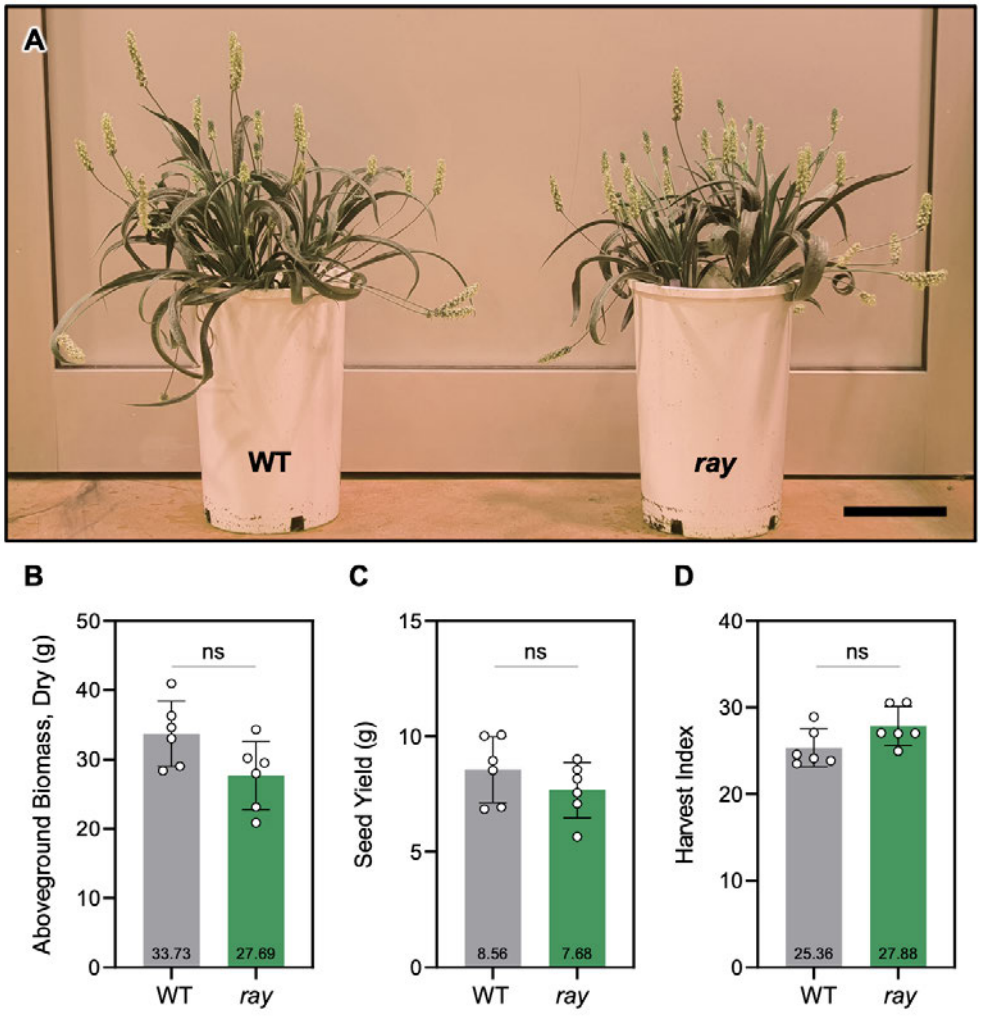


Figure 10. Comparison between WT and *ray* grown in the growth chamber showing visible plant phenotype (A), above-ground dry biomass (B), seed yield (C) and harvest index (D), n = 6 biological replicates.

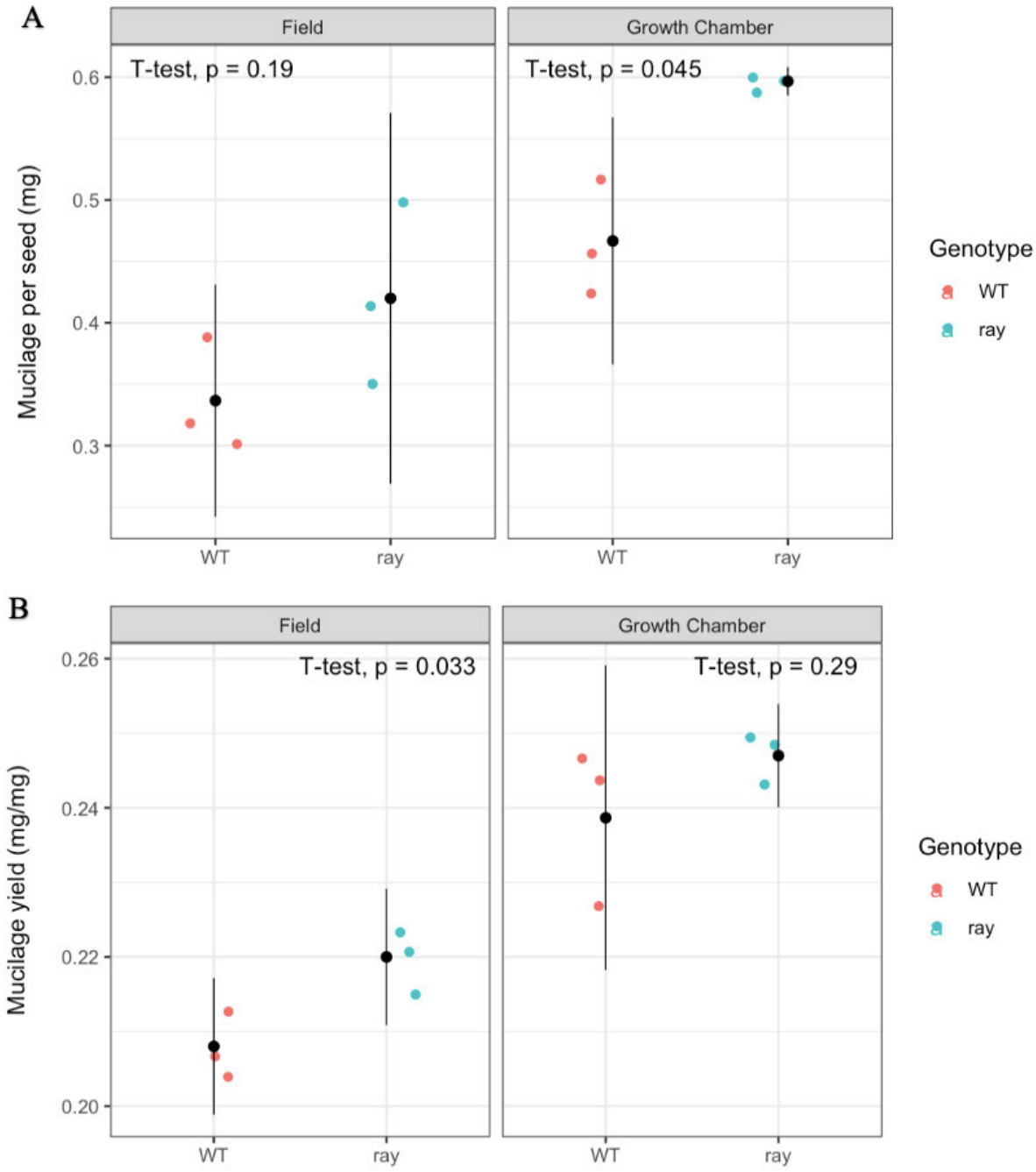


Figure 11. Comparison of mucilage yield between WT and *ray* grown in the growth chamber or under field conditions, showing mucilage per seed (A) and mucilage yield (B).

Mucilage polysaccharide quality

Mucilage polysaccharide composition between the two genotypes were relatively similar (P>0.1) (Figure 12). Of the monosaccharides measured, rhamnose concentration is below 10 M in CWE and barely detected in the HWE and IAE fractions. Xylose was detected in all samples; however, xylose concentration was higher in CWE and IAE (31 – 56.7 M) than in the HWE fraction (19.1 – 27.6 M). In contrast, arabinose was higher in IAE (12.7 – 20.7 M) than in CWE and HWE (4.5 – 11.8 M). Xylose and arabinose combined followed the xylose pattern with CWE and IAE containing more in total than the HWE fraction.

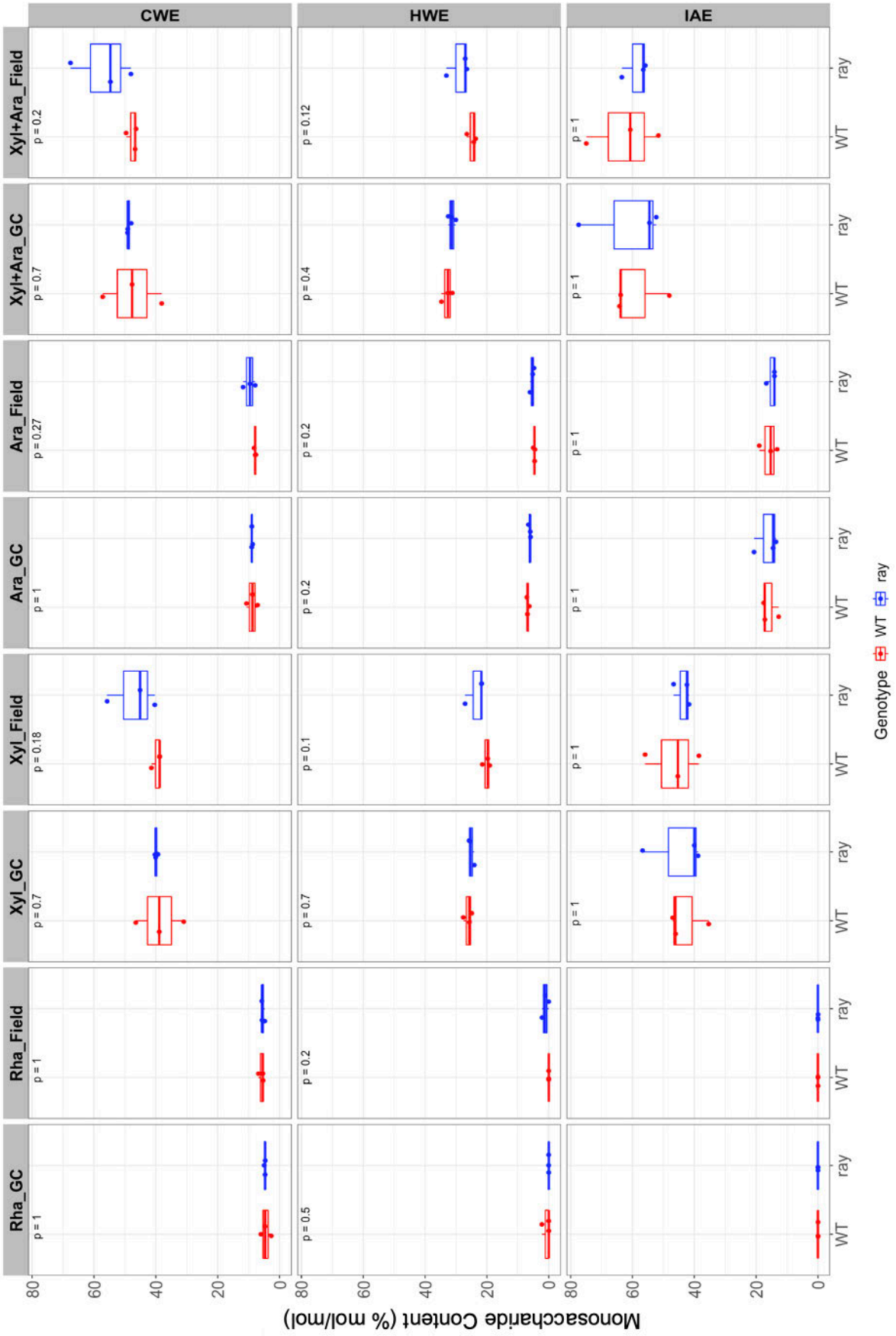

CWE
HWE
IAE
Xyl+Ara_Field
Xyl+Ara_GC
Ara_Field
Ara_GC
Xyl_Field
Xyl_GC
Rha_Field
Rha_GC
p = 0.2
p = 0.12
p = 1
p = 0.7
p = 0.4
p = 1
p = 0.27
p = 0.2
p = 1
p = 1
p = 0.2
p = 1
p = 0.18
p = 0.1
p = 1
p = 0.7
p = 0.7
p = 1
p = 1
p = 0.2
p = 1
p = 0.5
WT
ray
0
20
40
60
80
Monosaccharide Content (% mol/mol)
Genotype
WT
ray

Figure 12. Comparison of major mucilage monosaccharide composition between WT and *ray* grown in the growth chamber or the field. Abbreviations Rha_GC = Rhamnose Growth Chamber, Rha_Field = Rhamnose Field, Xyl_GC = Xylose Growth Chamber, Xyl_Field = Xylose Field, Ara_GC = Arabinose Growth Chamber, Ara_Field = Arabinose Field, Xyl+Ara_GC = Xylose and Arabinose Growth Chamber, Xyl+Ara_Field = Xylose and Arabinose Field, CWE = Cold water-extractable mucilage fraction, HWE = Hot water-extractable fraction, IAE = Intense extraction-resistant fraction.

Discussion

This study aimed to identify seed parameters affecting *P. ovata* mucilage production and select one candidate mutant for studying molecular mechanisms controlling the process. A previous study (Tucker et al., 2017) used the same source of gamma-irradiated mutants. The results show that seed phenotypes that link to a variation in gene expression were involved in heteroxylan biosynthesis (Tucker et al., 2017). Here, the study focused on the amount of mucilage rather than mucilage quality. The candidate mutant needed to have similar seed phenotype characteristics than the WT with no growth-habit defects; the only expected difference being the mucilage amount. When using mutants that have been generated artificially it is likely that there are numerous mutations present in the genome, which persist in the background and can cause pleiotropic effects that are not necessarily related to the trait of interest. The background mutations are often not removed until a comprehensive back crossing and selection process is followed. However, for some crop plants this is technically challenging, both where the physical process of crossing is technically challenging and not efficient and where lack of a genome might make the use of marker-assisted selection impossible. Instead, the identification of a line with a consistent phenotype, still present in the M5 and M6 selfed generations, with minimal background phenotypic defects was sought.

Seed weight and other seed features, namely area per seed, diameter, and perimeter, were found to be strongly correlated (Figures 4 and 5), so seed weight is enough to represent seed dimensions. Mucilage amount can be measured as mucilage yield per seed or as dry seed weight given that seed weights were positively correlated with mucilage yield per seed. In contrast, it negatively correlated with mucilage yield per dry seed weight. This is why the proportions of mutants with smaller seed (78% Figure 2A) and reduced mucilage yield per seed (68% Figure 2B) were similar. Overall, the mutant population has a high mucilage per dry seed weight. However, mucilage yield per dry seed weight does not correlate with mucilage yield per seed (Figure 3A).

Many seed phenotypes with reduced seed weight or lighter seeds have an abnormal appearance, including lines 794-10, 1004-6, 743, 833-5, and an extreme reduction can be observed in 1088-1c (Figure 3). Extreme reductions in seed weight led to a drastic increase in mucilage proportion in 1088-1c (Figure 3Q). Abnormally stained mucilage extrusion was also observed (Figure 7H). These seeds did not germinate as well as other lines suggesting a wide scale effect on fitness (data not shown). Even though reducing seed weight can increase mucilage proportion, it is not a good option, given that deleterious phenotype effects are often linked. In addition, mucilage is produced from the seed coat (husk), the smaller portion of the seed, while the significant portion is the endosperm and embryo (Cowley and Burton, 2021; Cowley et al., 2021). Therefore, increasing seed weight might mean more husk/mucilage but it also increases the waste product, which is considered to be less valuable at the moment.

Mutant lines having wild-type seed weights produce varying amounts of mucilage. The selection then focused on groups that had amounts of mucilage per seed and dry weight that were higher or lower than WT, but still possessed an average seed weight. Two candidates were lines 871-12 (Figure 3C) and 970-1 (Figure 3D), which had normal seed phenotypes. The mucilage yield per seed decreased by 23 % in 871-12 while increasing by 22% in 970-1 (Table 1, Figures 3C and D). The mucilage yield per dry weight decreased by 17% in 871-12 while increasing by 18% for 970-1 (Table 1, Figures 3C and D). Mucilage extrusion patterns are standard for both mutant lines. Seed yield per plant (Figure 8) and plant weight or biomass (Figure 9) of lines 871 and 970 growing in the field were higher than WT but only significantly so for line 871, but then plants of line 871-12 died during flowering.

Line 970-1 was selected as the best line producing more mucilage with minimal effects of background mutations and was then named *raya* (*ray*) which means ‘greater’ in Indonesian. More tests were conducted on *ray*. There were no significant differences between WT and *ray* plant growth (Figure 10 A) nor plant biomass, seed yield and harvest index (Figure 10 B-D). Mucilage per seed was significantly higher from *ray* than WT growing in the growth chamber.

Meanwhile, the proportion of mucilage for a given weight of seed was significantly higher in the field for *ray* than WT (Figure 11). Mucilage composition was also similar for both genotypes growing in the field or growth chamber, indicating no likely effects on downstream quality. Therefore, *ray* was chosen as an excellent candidate to study genes controlling mucilage production, in comparison to WT, as no pleiotropic phenotypes were observed. This work is presented in Chapter 6.

Conclusion

Gamma-irradiated mutants provide valuable resources for genetic improvement. In this study, mutant lines were screened to obtain one candidate mutant suitable for identifying genes related to mucilage production. A combination of 5 variables can explain about 97.1% of the variation among mutant lines. They are seed weight, area perimeter, diameter, mucilage yield per seed, and mucilage yield per dry seed weight. Principal Component Analysis (PCA) helped to visualise and identify groups in the mutant population. Many combinations of seed characteristics can be used to select mutants according to end use. For several reasons, the mutant named *raya* was selected as the candidate mutant for the work presented in Chapter 6. It has a high mucilage yield but similar composition, a relatively similar seed weight but with no obvious defects in plant development.

References

Ahloowalia B, Maluszynski M (2001) Induced mutations–A new paradigm in plant breeding. Euphytica **118:** 167-173

Cowley JM, Burton RA (2021) The goo-d stuff: *Plantago* as a myxospermous model with modern utility. New Phytol **229:** 1917-1923

Cowley JM, Herliana L, Neumann KA, Ciani S, Cerne V, Burton RA (2020) A small-scale fractionation pipeline for rapid analysis of seed mucilage characteristics. Plant Methods **16:** 1-12

Cowley JM, McNeil DL, Lui KY, Barsby JP, Ciani S, Cerne V, Burton RA (2022) Rain events at maturity severely impact the seed quality of psyllium (*Plantago ovata* Forssk.). Journal of Agronomy and Crop Science

Cowley JM, O'Donovan LA, Burton RA (2021) The composition of Australian *Plantago* seeds highlights their potential as nutritionally-rich functional food ingredients. Sci Rep **11:** 12692

Dhar M, Kaul S, Sareen S, Koul A (2005) *Plantago ovata*: Genetic diversity, cultivation, utilization and chemistry. Plant Genetic Resources: characterization and utilization **3:** 252-263

Fougat RS, Joshi C, Kulkarni K, Kumar S, Patel A, Sakure A, Mistry J (2014) Rapid development of microsatellite markers for *Plantago ovata* Forsk.: using next generation sequencing and their cross-species transferability. Agriculture **4:** 199-216

Jensen JK, Johnson N, Wilkerson CG (2013) Discovery of diversity in xylan biosynthetic genes by transcriptional profiling of a heteroxylan containing mucilaginous tissue. Frontiers in plant science **4:** 183-183

Jensen JK, Johnson NR, Wilkerson CG (2014) *Arabidopsis thaliana* IRX10 and two related proteins from psyllium and *Physcomitrella patens* are xylan xylosyltransferases. The Plant Journal **80:** 207-215

Karimzadeh G, Omidbaigi R (2004) Growth and seed characteristics of isabgol (*Plantago ovata* Forsk) as influenced by some environmental factors. Journal of Agricultural Science and Technology **6:** 103-110

Lal RK, Chanotiya CS, Gupta P (2020) Induced mutation breeding for qualitative and quantitative traits and varietal development in medicinal and aromatic crops at CSIR-

CIMAP, Lucknow (India): past and recent accomplishment. International Journal of Radiation Biology **96:** 1513-1527

Phan JL, Cowley JM, Neumann KA, Herliana L, O'Donovan LA, Burton RA (2020) The novel features of *Plantago ovata* seed mucilage accumulation, storage and release. Sci Rep **10:** 1-14

Phan JL, Tucker MR, Khor SF, Shirley N, Lahnstein J, Beahan C, Bacic A, Burton RA (2016) Differences in glycosyltransferase family 61 accompany variation in seed coat mucilage composition in *Plantago* spp. Journal of Experimental Botany **67:** 6481-6495

Singh N, Lal RK, Shasany AK (2009) Phenotypic and RAPD diversity among 80 germplasm accessions of the medicinal plant isabgol (*Plantago ovata*, Plantaginaceae). Genet. Mol. Res **8:** 1273-1284

Tucker M, Ma C, Phan J, Neumann K, Shirley N, Hahn M, Cozzolino D, Burton R (2017) Dissecting the genetic basis for seed coat mucilage heteroxylan biosynthesis in *Plantago ovata* using gamma irradiation and infrared spectroscopy. Frontiers in Plant Science **8**

Voiniciuc C, Engle KA, Günl M, Dieluweit S, Schmidt MH-W, Yang J-Y, Moremen KW, Mohnen D, Usadel B (2018) Identification of key enzymes for pectin synthesis in seed mucilage. Plant Physiology **178:** 1045-1064

Western TL, Burn J, Tan WL, Skinner DJ, Martin-McCaffrey L, Moffatt BA, Haughn GW (2001) Isolation and characterization of mutants defective in seed coat mucilage secretory cell development in *Arabidopsis*. Plant Physiology **127:** 998-101

Chapter 6

Comprehensive transcriptome analysis of *Plantago ovata* seed development related to mucilage production

ꕥ•ꕥ

Abbreviations

DPA	Days post anthesis
GSEA	Gene set enrichment analysis
INT	Integument
KEGG	Kyoto encyclopedia of genes and genomes
KO	KEGG ortholog
LCM	Laser capture microdissection
ML	Mucilage layer
MSC	mucilage secretory cell
MT_INT_2	*ray* integument 2 DPA samples
MT_INT_4	*ray* integument 4 DPA samples
MT_INT_6	*ray* integument 6 DPA samples
MT_MSC_2	*ray* mucilage secretory cell 2 DPA samples
MT_MSC_4	*ray* mucilage secretory cell 4 DPA samples
PCA	Principal component analysis
WT_INT_2	WT integument 2 DPA samples
WT_INT_4	WT integument 4 DPA samples
WT_INT_6	WT integument 6 DPA samples
WT_MSC_2	WT mucilage secretory cell 2 DPA samples
WT_MSC_4	WT mucilage secretory cell 4 DPA samples
RNAseq	RNA sequencing
WT	Wild type

Abstract

Plantago ovata seeds release mucilage upon wetting which is widely used in food and health industries. Unlike the model species *Arabidopsis thaliana*, *P. ovata* seed mucilage is enriched in heteroxylan. Recently, mechanisms guiding mucilage accumulation, storage and release in *P. ovata* have been described but there is still much to discover regarding mucilage production mechanisms. This study used immunolabelling, cryo-sectioning, laser capture microdissection (LCM) and RNA-sequencing to define and compare genes impacting mucilage biosynthesis in two developing seed tissues - the mucilage secretory cells (MSC) and the integument (INT). Total RNA was isolated from these different tissue types from the parental WT and a gamma-irradiated mutant called *raya* that produces a higher yield of mucilage per seed. Presence/absence gene expression analysis, Principal Component Analysis (PCA), differential gene expression analysis using EdgeR and gene set enrichment analysis (GSEA) were performed using GO-terms as biological themes. About 64% of all known *P. ovata* genes were expressed in these developing seed tissues. There were 611 transcripts specific to *raya* samples including seven transcription factors (TF) with only 280 transcripts specific to the WT, including three TFs. Only one out of the seven TFs identified, *GL2,* has been previously shown to be implicated in mucilage production. PCA showed separation between tissue types and time points (2 and 4 DPA), while some samples overlapped between the two genotypes. Twenty-seven pairwise comparisons were made to detect differentially expressed genes (DEGs) and enrichment clusters. DEGs varied from 1 to 8 for genotype comparisons, from 1,046 to 3,443 for tissue comparisons, and from 443 to 3,099 for time-point comparisons. A cell wall cluster was enriched in the INT from *ray* at 2 DPA but was not present in the WT samples at the same age. Like Glycosyltransferase family 47 (*IRX7* and *IRX10*) and GT61 genes, GT43 genes (*IRX9* and *IRX14*) were also expressed in early development but mainly in MSC not in INT tissues and at a higher level in *ray* samples. Enrichment cellular compartment analysis indicates mucilage polysaccharide biosynthesis occurs in the Golgi apparatus and they are then

transported via vesicles and deposited into the vacuole. This study provides valuable information not only for the *P. ovata* breeding program but also for advancing knowledge in polysaccharide biosynthesis relevant to numerous plant species.

Introduction

Plantago ovata has myxospermous seeds that release mucilage polysaccharides upon hydration (Anderson and Fireman, 1935; Hyde, 1970; Yu et al., 2017; Phan et al., 2020; Cowley and Burton, 2021). Dry mucilage or the husk which is milled off the outside of the seed is called psyllium and is used in many applications (Dhar et al., 2005; Rao et al., 2013; Verma and Mogra, 2013; Cowley and Burton, 2021), including in gluten-free products (Franco et al., 2020). Attempts have been made to obtain optimum psyllium yield by cultivating this plant in areas with suitable conditions such as temperature, pH and type of soil, the right season, and by applying nitrogen fertilisers (Karimzadeh and Omidbaigi, 2004; Ghaderi-Far et al., 2012; Kumar, 2015). However, the ratio of husk yield to seed weight is only 25-30% (Kumar, 2015; Cowley et al., 2020) and so 70%, the non-husk components, are discarded or used for animal feed (Cowley and Burton, 2021; Cowley et al., 2021). The challenge is to increase the percentage of seed husk by maximising mucilage polysaccharide production without interfering with seed development. Comprehensive knowledge about seed development and molecular mechanisms controlling mucilage production is required to alter or engineer this characteristic.

The development of *P. ovata* mucilage-producing cells or mucilage secretory cells (MSCs) has been investigated as early as 1970 (Hyde, 1970). However, the detailed characterisation of mucilage accumulation, storage and release has only recently been published (Phan et al., 2020). Following fertilisation, like other plants, ovule integuments differentiate into the mature seed coat or the seed husk in the case of *P. ovata*. The outermost epidermal integument layers become distinct shapes and differentiate into MSCs at 1 DPA (days post-anthesis) (Phan et al., 2020) or upon pollination (Hyde, 1970). The cells contain expanded vacuoles and no evidence

of cell division (Hyde, 1970). From 3 to 5 DPA, MSCs grow and elongate rapidly (Phan et al., 2020). Cell wall remodelling and disintegration start at 6/7 DPA as indicated by changes in carbohydrate-binding module labelling (Phan et al., 2020). Bansal (2018) showed a reduction in soluble sugar content by 82.8%, while starch and cellulose increased significantly in the developing outer tissues seven days after anthesis. By 15 DPA, the outer cell walls of the MSCs have degraded and so the mucilage polysaccharides become exposed as the outer layer on the seed, although at this point the seeds are still inside the capsule (Phan et al., 2020). The thickness of the MSC layer reduces from 80-90 um at 7 DPA to become compressed into a cell-free thin mucilage polysaccharide layer (10-18 um) in mature seeds (Phan et al., 2020). Unlike the model plant Arabidopsis, a secondary cell wall columella does not form in the *P. ovata* seed and the cell walls of the outer seed coat layer of Arabidopsis seeds remains intact, only rupturing when the mature seed is exposed to moisture (Phan et al., 2020; Cowley and Burton, 2021).

The inner integument layers, which sit inside the MSC layer and are likely to supply it with carbon for polysaccharide synthesis, also degrade and become squashed between the expanding embryo and the degraded MSC layer. Later a pigmented layer also develops in this vicinity (Phan et al., 2020). Seed integuments are important tissues and have been studied in a range of plants (Weber et al., 1996; Shea Miller et al., 1999; Haughn and Chaudhury, 2005). Currently, the role of the integuments is not at all understood in *Plantago*, nor are any of the molecular processes likely to be occurring, such as programmed cell death or transport of nutrients to the MSC layer or the growing embryo, defined. We have no indication if the relative size of the integument tissues and efficiency of nutrient transfer might have an impact on the final seed size or the amount of mucilage polysaccharides that are made, which ultimately determines the final husk, or psyllium, yield. Nor do we know if disintegration of the integument tissues as the seed develops is driven by programmed cell death, as with other seeds (Beers et al., 2000; Hierl et al., 2012; Lima et al., 2015; Lopez-Fernandez and Maldonado, 2015), physical pressure

from the growing embryo, or a combination of both. There is much to discover about this tissue in *Plantago* seeds.

P. ovata mucilage is enriched in complex heteroxylan, comprising about 90% of the total with small amounts of pectin and cellulose (Fischer et al., 2004; Guo et al., 2008; Cowley et al., 2020). Heteroxylan is composed of a backbone of xylose residues decorated with a variety of side chains typically comprised of arabinose (Ara), xylose (Xyl), and glucuronic acid (GlcA), and traces of other sugars (Fischer et al., 2004; Yu et al., 2017). In contrast, Arabidopsis seed mucilage is composed primarily of pectin (Naran et al., 2008; Arsovski et al., 2010). Fundamental research into mucilage biosynthesis in Arabidopsis is well established. Many genes have been reported as master regulators of seed coat cell growth and differentiation and genes involved in synthesising mucilage components, including pectin, hemicellulose, cellulose and structural proteins, and mucilage deposition and secretion have been described (Western et al., 2001; Western et al., 2004; Gonzalez et al., 2009; Arsovski et al., 2010; Saez-Aguayo et al., 2013; North et al., 2014; Voiniciuc et al., 2015; Golz et al., 2018). The differences in seed development patterns and mucilage composition indicates that *P. ovata* is likely to have distinctly different molecular mechanisms from Arabidopsis. Therefore, Arabidopsis regulatory networks controlling seed mucilage pathways are not likely to be fully translatable to crop plants like *P. ovata*.

Several studies have been conducted to elucidate *P. ovata* mucilage polysaccharide biosynthesis using RNA sequencing. Jensen et al. (2013) proposed that *P. ovata* only requires the glycosyltransferase (GT) family 47 (GT47) gene *IRX10*, but no GT43s, namely *IRX9* and *IRX14,* for xylan biosynthesis. Although the expression of GT47 genes was very low in their MSC samples at 6, 8, 10 and 12 DPA, four *IRX10* homologs and seven *GT61* genes involved in making the backbone and sidechains of xylan, respectively, were found. Eleven more *GT61* genes were identified by Phan et al. (2016). They also observed peak expression of *IRX10* homologues and six *PoGT61* genes at 13-14 DAP (Phan et al., 2016). Kotwal et al. (2016) have

identified 18 genes from *P. ovata* ovaries at 15 DPA that are also present in Arabidopsis developing seeds, with three of them being transcription factors *APETALA2* (*AP2*), *TRANSPARENT TESTA GLABRA1* (*TTG1*) and *LEUNIG HOMOLOG* (*LUH*). Gupta et al. (2018) analysed the expression pattern of five genes with predicted involvement in arabinoxylan or heteroxylan synthesis from ovules at 3, 7, 11, and 15 DPA. They found that the expression level of three out of five of these genes increased by 5-8-fold at 15 DPA.

While all these results add to our current knowledge about synthesis of *P. ovata* mucilage polysaccharides, understanding of the molecular mechanisms is still far behind the model species Arabidopsis. Therefore, this study aimed to advance our knowledge of *P. ovata* by carefully designing a more targeted RNAseq experiment. Previous *P. ovata* RNAseq data analysis included samples at the later stage of seed development, where most MSCs are already ruptured and contain much mucilage. Mucilage can hinder efficient RNA extraction (Jensen et al., 2013; Kotwal et al., 2016; Phan et al., 2016; Gupta et al., 2018), and many of the transcripts reported in previous studies may not be from the MSCs as whole ovules (seeds) and ovaries (fruit) were used. These organs contain endosperm and embryo tissues which have multiple cell types. MSCs are only one layer (Phan et al., 2020), so using whole seed tissues may also dilute the expression of genes specifically in MSCs, or the integuments. Here, cryo-sectioning and laser capture microdissection techniques on sections from early developmental time points (2, 4, and 6 DPA) were used to tackle this problem, producing much cleaner data sets from specific tissues.

Much progress has been made in understanding mucilage polysaccharide biosynthesis in Arabidopsis using mutants (Western et al. 2001). Therefore, we screened our gamma-irradiated mutant collection (Tucker et al., 2017) to find one with a higher yield of mucilage called *raya* (*ray*) as described in Chapter 5. This mutant was used in comparison with WT as an additional tool to help determine candidate genes that might control the amount of mucilage polysaccharides produced. We hypothesised that the transcriptome profile of integument and

mucilage secretory cells would differ between the WT and *ray* at the same developmental time point. Thus, information from all the RNAseq experiments will aid in identifying gene targets that might be useful in a breeding program to allow manipulation of the mucilage quantity and quality, thus directly impacting downstream applications and economics of psyllium use.

Material and Methods

Plant materials and sample collections

To understand factors contributing to mucilage polysaccharide production during early seed development, samples were collected from two genotypes, two tissue types and three time points. WT and *ray* were obtained from a population previously generated by Tucker et al. (2017) and screened using a method described by Cowley et al. (2020), as outlined in Chapter 5. Seeds were grown from the M5 generation (all selfed) and the *ray* mutant plants appeared phenotypically indistinguishable from the WT, except for mucilage yield (Figure 1), including at the cell wall level inside the developing seeds (Figure 2). Integument and MSC layers were chosen as the target tissues since they are both contributing to mucilage polysaccharide synthesis and might influence final yield. Three developmental time points - 2, 4, and 6 DPA were chosen for the gene expression profiling as they represent key developmental transitions (Phan et al., 2020). In total, there were 40 samples (10 groups with four replicates). Since processing samples took several months, the samples were randomised completely using R.

This study complies with local and national guidelines. Plants were grown in a controlled plant growth room with a 16/8 hrs light cycle and a constant temperature of 22-23^{o}C at the University of Adelaide. Flowers were marked and tagged with the date to harvest at the relevant day post-anthesis (DPA) following Phan et al. (2020). Inflorescences were harvested, and fruits were removed and dissected using a scalpel and fine-tip tweezers under a dissection microscope to remove the floret and bracts. Each fruit was placed into a 2 ml microcentrifuge tube, then immediately frozen in liquid nitrogen and stored at -80^{o}C until required.

RNA-seq experiments

Frozen *P. ovata* fruits were sectioned using a cryomicrotome to a thickness of 7 µm and placed on PEN membrane slides (product code 11600288). Once the slide dried, the region of interest was cut out using a Microdissection Leica LMD Microscope (Figure 3). Total RNA was extracted using the Arcturus Pico RNA extraction kit (Thermo Fisher Scientific, Inc) and then treated with RNAse-free DNAse (1:8 dilution of DNase I in RDD buffer; Qiagen). Total RNA samples were checked for RNA integrity using the TapeSation 2200 and quantity was assessed using a Qubit. All samples were adjusted to a concentration of 0.266 ng/mL with RIN values in the range 6.6–8.3 with an average size of 399-454 bp (Supplementary Table 1). Some samples were concentrated using a Speed Vac. Ribodepletion was performed using a customised ribodepletion kit, TECAN. Total clusters were 1,486 million. Stranded total RNA-seq libraries from forty samples were generated according to the Nugen Universal Plus Total RNA-seq protocol (Part No. 15044223 Rev. B) and included 17 cycles of amplification. Libraries had similar sizes and quantities (Supplementary Table 1). The libraries were sequenced in an MGI FCL flowcell with PE 148 with an additional eight bp UMIs using MGI DNBSEQ-G400 sequencing chemistry, generating up to 1.6 billion reads per flowcell. The raw sequences were deposited at the SRA NCBI database under accession numbers SRR19600243 – SRR19600282. A reviewer link can be found at

https://dataview.ncbi.nlm.nih.gov/object/PRJNA732452?reviewer=pnqfc729ma1fet4tcg5pbia040.

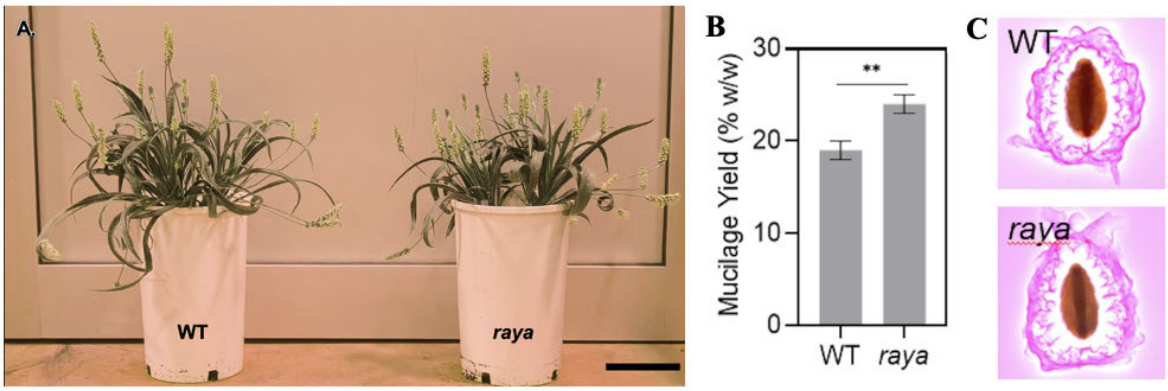


Figure 1. Comparison of two genotypes (WT and *raya*). (a) Both genotypes grew well and plants showed no obvious vegetative defects for *ray*; (b) *ray* seeds produced more mucilage compared to the WT ($P < 0.05$, Student's t-test using Graphpad Prism 8.4); (c) Mucilage released from seeds visualized by staining with ruthenium red which stains acidic polysaccharides like pectin, shows no difference in mucilage morphology between the two genotypes.

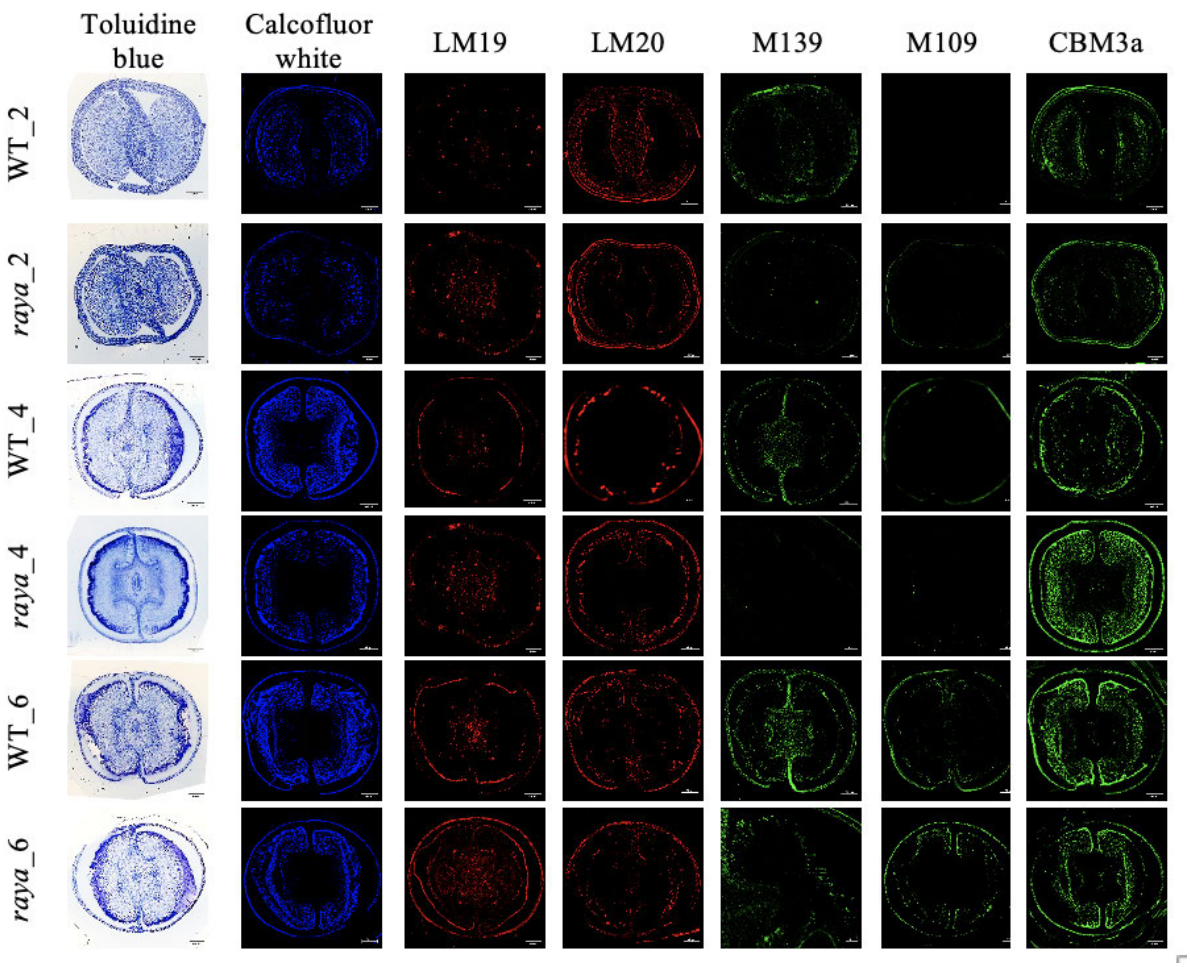


Figure 2. Staining and immunolabelling on *P. ovata* transversal seed and capsule sections show similar cell wall composition between wild type and *raya*. LM19 and LM20 were used to label unesterified and esterified homogalacturonan, respectively. M109 and M139 to detect heteroxylan and branched xylan, respectively. CBM3a to detect crystalline cellulose. Detail method can be found in Chapter 4. WT_2 = wildtype capsule at 2 DPA, *raya*_2 = *raya* capsule at 2 DPA, WT_4 = wildtype capsule at 4 DPA, *raya*_4 = *raya* capsule at 4 DPA, WT_6= wildtype capsule at 6 DPA, *raya*_6 = *raya* capsule at 6 DPA. Scale bar = 200 µm.

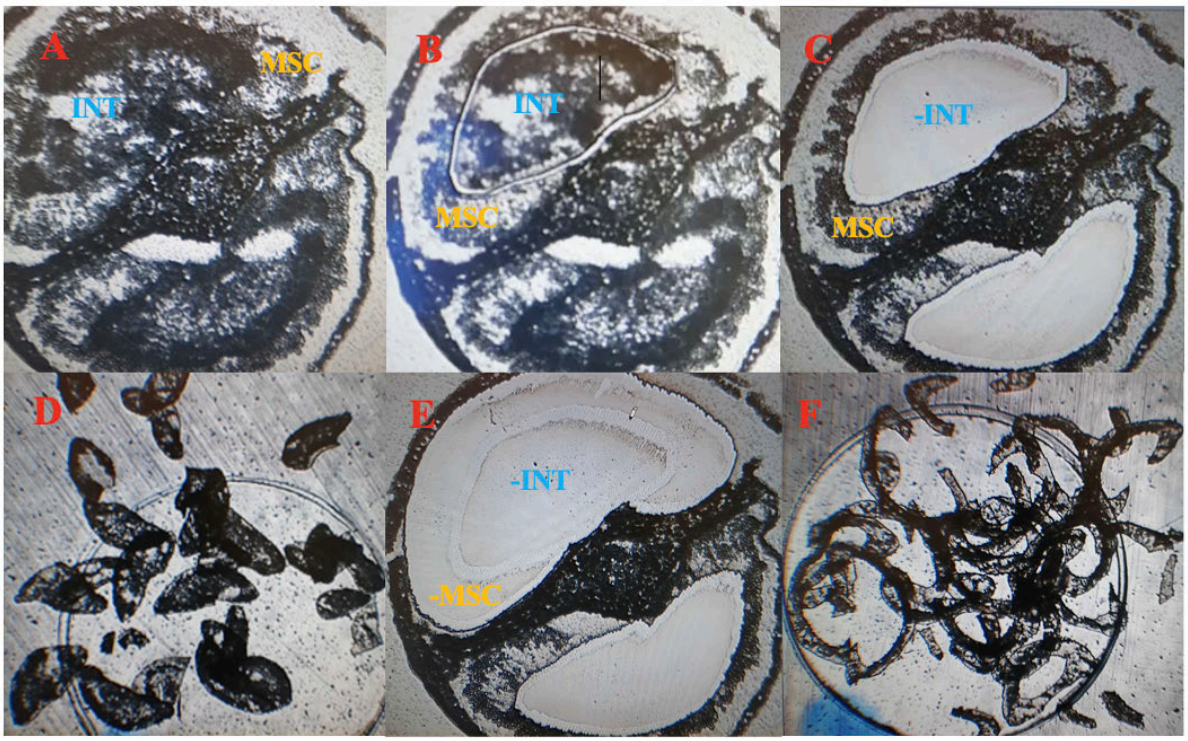


Figure 3. An example of *P. ovata* cryo-sectioned fruit at 4 DPA for isolating INT and MSC layers using the Microdissection Leica LMD Microscope (a) before; (b) during lasering; (c & e) after lasering; (d) integument sections; (f) MSC sections. No scale is provided as there were problems with the software. INT = integument layer, MSC = mucilage secretory cell layer. -INT or -MSC = the tissues were dissected by the laser.

RNAseq analysis

The RNA-seq data were processed by quality checking, trimming, cleaning and aligning reads to the reference genome, followed by deduplication. Quality checking was performed using FastQC v0.11.9 (Andrews, 2017) and MultiQC v1.8 (Ewels et al., 2016) with default parameters. Trimmomatic v0.39 (Bolger et al., 2014) removed adapter and PCR primer fragments. BBDuk, BBmap v38.87 (Bushnell, 2014) removed unwanted reads. Unwanted reads were from the *P. ovata* chloroplast genome, one mitochondrial gene, and ribosomal and transfer RNA (rRNA and tRNA) genes. Clean reads were aligned to the reference genome using STAR v2.7.6a with a 2-pass procedure (Dobin et al., 2013). Before read counting, deduplication using umi_tools v1.1.2 was performed to remove PCR duplicates.

After counting reads with gene-level summarisation using featureCounts (Liao et al., 2014), the first step in differential expression analysis was performed using edgeR (Robinson et al., 2010). Count per million (CPM) and The Trimmed Mean of the M-values (TMM) methods were used to normalise count data. For gene presence and absence analysis, genes from WT and *ray* samples were filtered separately with criteria expression level CPM $\geq$ 0.5 and expressed in more than ten samples. Then an interaction function in R was applied. We kept genes that have CPM < 1 and were expressed in more than 50% of the samples (20/40 samples) for clustering analysis with factoextra R package (PCA = Principal Component Analysis) (Kassambara and Mundt, 2017) and Clust v1.12.0 (Abu-Jamous and Kelly, 2018). The same parameters were applied for finding differentially expressed genes (DEGs) with limma (Ritchie et al., 2015) and for identifying enriched gene sets using GSEA (Gene Set Enrichment Analysis) (Subramanian et al., 2005).

Using the GSEA v4.1.0 (Subramanian et al., 2005) desktop application, we performed enrichment analysis on the normalised expression data. We used a customised gene sets database (*P. ovata* GO term database), selected 1000 as the number of permutations without

collapsing the dataset and used the gene set as permutation type and selected a paired phenotype. Cytoscape v3.8.0 application (Shannon et al., 2003) with EnrichmentMap installed was used to visualise the GSEA results.

P. ovata sequences and annotations are available at DDBJ/ENA/GenBank under JAHHQI010000000. Gene Ontology (GO) terms and KEGG ortholog (KO) identifiers were obtained by submitting *P. ovata* protein sequences to the eggNOG-mapper website (Cantalapiedra et al., 2021).

An extensive set of comparisons, each designated by a Roman numeral, was made between samples according to genotype, tissue type and developmental stage. For ease of reference all the comparisons are compiled in Table 1.

Table 1. A list of each comparison used for DEG and gene set enrichment analyses with assigned Roman numeral, the symbols used to define it, the samples included in each comparison and the total number of samples in the set.

Comparison		Symbols	Comparison		Number of samples (@ 4 replicates)
			Group 1	Group 2	
I	Genotype	WTvsMT	WT_INT_2 WT_INT_4 WT_MSC_2 WT_MSC_4	MT_INT_2 MT_INT_4 MT_MSC_2 MT_MSC_4	32
II		WTvsMT_2	WT_INT_2 WT_MSC_2	MT_INT_2 MT_MSC_2	16
III		WTvsMT_4	WT_INT_4 WT_MSC_4	MT_INT_4 MT_INT_4	16
IV		WTvsMT_INT	WT_INT_2 WT_INT_4	MT_INT_2 MT_INT_4	16
V		WTvsMT_INT_2	WT_INT_2	MT_INT_2	8
VI		WTvsMT_INT_4	WT_INT_2	MT_INT_2	8
VII		WTvsMT_MSC	WT_MSC_2 WT_MSC_4	MT_MSC_2 MT_MSC_4	16
VIII		WTvsMT_MSC_2	WT_MSC_2	MT_MSC_2	8
IX		WTvsMT_MSC_4	WT_MSC_4	MT_MSC_4	8
X	Tissue	INTvsMSC	WT_INT_2 WT_INT_4 MT_INT_2 MT_INT_4	WT_MSC_2 WT_MSC_4 MT_MSC_2 MT_MSC_4	32
XI		INTvsMSC_2	WT_INT_2 MT_INT_2	MT_MSC_2 MT_MSC_4	16
XII		INTvsMSC_4	WT_INT_4 MT_INT_4	WT_MSC_4 MT_MSC_4	16

XIII		WT_INTvsMSC	WT_INT_2 WT_INT_4	WT_MSC_2 WT_MSC_4	16
XIV		WT_INTvsMSC_2	WT_INT_2	WT_MSC_2	8
XV		WT_INTvsMSC_4	WT_INT_4	WT_MSC_4	8
XVI		MT_INTvsMSC	MT_INT_2 MT_INT_4	MT_MSC_2 MT_MSC_4	16
XVII		MT_INTvsMSC_2	MT_INT_2	MT_MSC_2	8
XVIII		MT_INTvsMSC_4	MT_INT_4	MT_MSC_4	8
XIX	Time-point	2vs4	WT_INT_2 WT_MSC_2 MT_INT_2 MT_MSC_2	WT_INT_4 WT_MSC_4 MT_INT_4 MT_MSC_4	32
XX		2vs4_INT	WT_INT_2 MT_INT_2	WT_INT_4 MT_INT_4	16
XXI		2vs4_MSC	WT_MSC_2 MT_MSC_2	WT_MSC_4 MT_MSC_4	16
XXII		WT_2vs4	WT_INT_2 WT_MSC_2	WT_INT_4 WT_MSC_4	16
XXIII		WT_INT_2vs4	WT_INT_2	WT_INT_4	8
XXIV		WT_MSC_2vs4	WT_MSC_2	WT_MSC_4	8
XXV		MT_2vs4	MT_INT_2 MT_MSC_2	MT_INT_4 MT_MSC_4	16
XXVI		MT_INT_2vs4	MT_INT_2	MT_INT_4	8
XXVII		MT_MSC_2vs4	MT_MSC_2	MT_MSC_4	8

WT = wild type, MT = *ray* mutant, MSC = mucilage secretory cells, INT = integument, 2 = 2 days post anthesis, 4 = 4 days post anthesis

Results

Transcript analysis on INT and MSC tissues was performed to identify pathways and genes involved in mucilage biosynthesis.

Clustering gene expression data using Principal Component Analysis (PCA)

The principal component analysis (PCA) plot explained variation in sample groups by 36.95%. PC1 (20.85%) differentiated the sample groups based on tissue type (INT versus MSC), while PC2 (16.1%) mainly separated 4 DPA from 2 DPA samples (Figure 4). WT and *ray* samples overlapped. Integument 6 DPA samples overlapped with samples 2 and 4 DPA, but they all sit in the integument cluster.

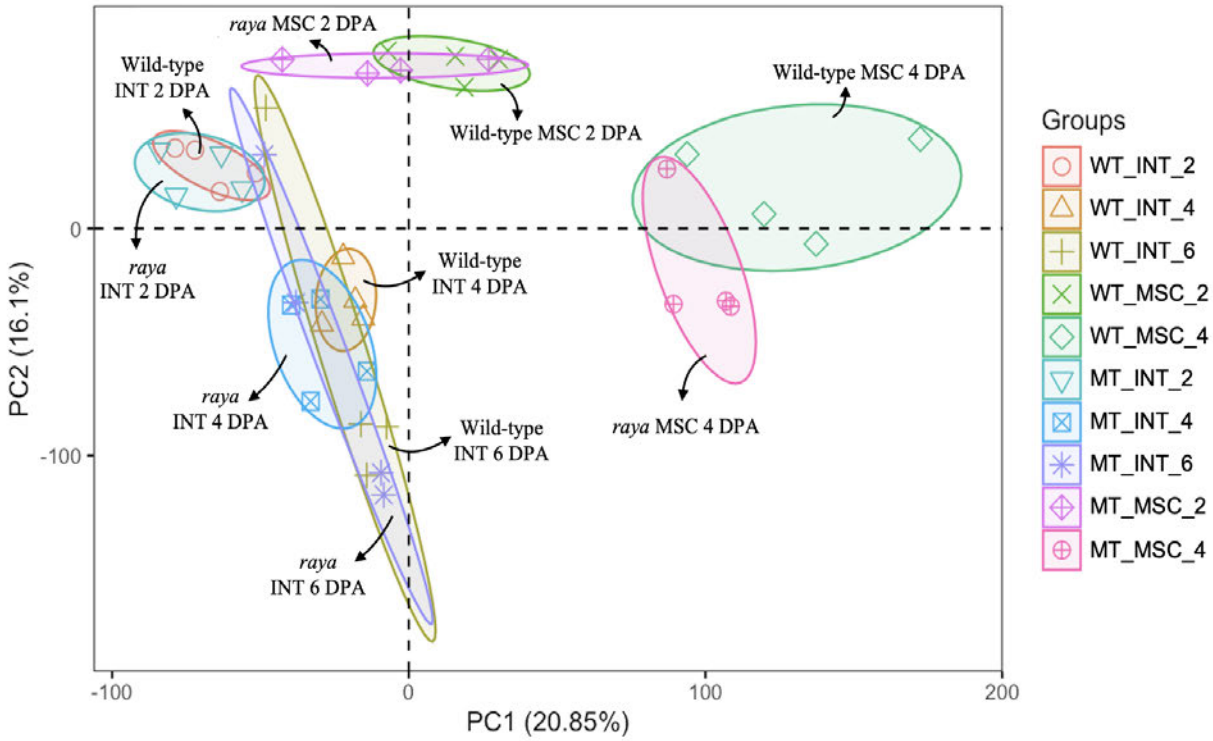


Figure 4. Principal Component Analysis (PCA) on *P. ovata* developing seed RNAseq data.

The *P. ovata* genome has 25,045 genes (see Chapter 3). Of this total, 15,951 or 63.7% were expressed in the developing seed samples (Figure 5). About 15,060 genes were found in both WT and *ray* samples, while 9,094 (36.3%) were not present in these tissues at all (Figure 5). More genes were specific to *ray* samples (611 genes) than to the WT (280 genes) (Figure 5).

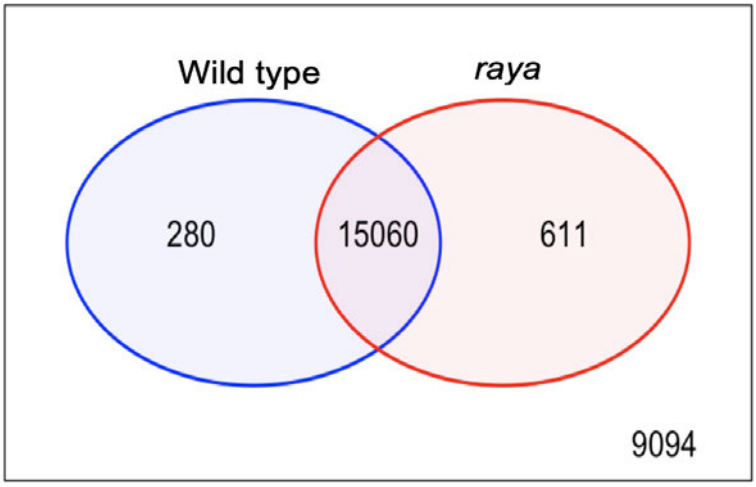


Figure 5. Numbers of genes, from the total identified in the genome (Chapter 3), expressed exclusively in WT and *ray* samples, that overlap, or that are not present in any of these tissues.

Identification of transcription factors during seed development in WT and *ray*

Genes encoding transcription factors (ko03000) were identified. At the genome level, there were 50 KO identifiers (327 genes) grouped into six eukaryotic types: basic leucine zipper (bZIP), basic helix-loop-helix (bHLH), zinc finger, helix-turn-helix, beta Scaffold factors with minor groove contacts, and other transcription factors with one other transcription factor matching a prokaryotic type. There were 48 KO assignments identified in WT (191 genes) or WT and *ray* (188 genes), with about 49 KO numbers (202 genes) for *ray*. About 18 KO terms were not expressed in these samples. Only two KO numbers (3 genes) were specific to WT, while seven (14 genes) were specific to *ray*. A list of identified transcription factors can be found in Supplementary Table 2 with TFs specific to WT in Table 2 and those for *ray* in Table 3.

Differential expression of *P. ovata* genes and gene set enrichment analysis between the two genotypes (Comparisons I–IX)

Sets of samples from WT versus *ray* were compared as shown in Table 1. The highest total number of DEGs between WT and *ray* were eight in Comparison I (where all samples were included) and VII (MSC samples only) (Table 1). In Comparison I, five upregulated and three downregulated genes were found. All upregulated genes were related to resistance against *Phytophthora infestans*. Two genes encode putative late blight resistance protein R1A-6, two encode putative serine/threonine-protein kinases, with one gene for molybdopterin biosynthesis protein CNX2. The three downregulated genes were linked to control of programmed cell death and were cysteine proteinase inhibitor 3, molybdopterin biosynthesis protein CNX2 and nicotinate phosphoribosyltransferase 1. Only cysteine proteinase inhibitor three was downregulated regardless of tissue type and time point.

Gene set enrichment across comparisons I to IX produced variable results. No enriched gene sets were detected when comparing all samples (Comparison I), similarly for all integument samples (Comparison IV) or for integument samples at 4 DPA (Comparison VI). However, there were 6 clusters of significance in other Comparison I–IX (Supplementary Table 3 and Figure 6). The highest number of gene sets and clusters were found in 2 DPA samples (Comparison II, V, and VIII) (Table 4). There was a downregulated glucan catabolic cluster in *ray* MSC layer at 2 DPA (Comparison VIII) and an upregulated cell wall metabolic cluster in the integument at 2 DPA (Comparison V). Cell wall metabolic processes that were upregulated consist of two gene sets. They were GO:0044036 (Biological Process (BP) cell wall macromolecule metabolic process) and GO:0010383 (BP cell wall polysaccharide metabolic process). Sixteen out of 27 genes in the cell wall metabolic cluster are in the leading-edge subset based on the GSEA enrichment score (Table 5).

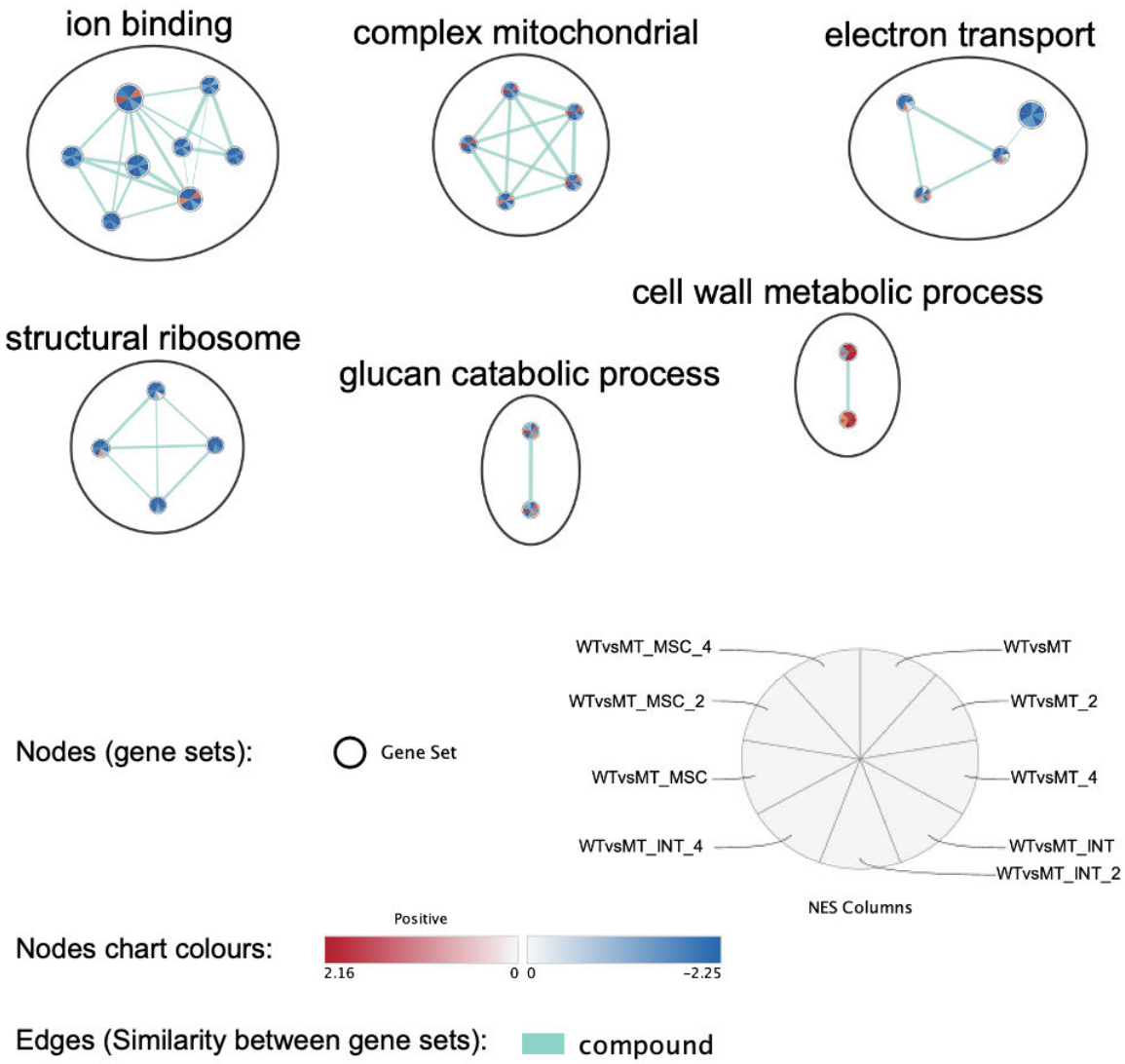


Figure 6. Gene ontology (GO) term enrichment analysis on gene sets from comparisons I-IX, comparing two genotypes. Network enrichment analysis was built using all expressed genes in all tissues processed using GSEA and then visualised using Cytoscape v3.8.0 (FDR q value > 0.1). Big circles represent clusters. Each cluster consists of two or more nodes (gene sets). Each node slice represents a gene set (GO term) from each comparison. Red indicates enriched upregulated genes, and blue shows enriched downregulated genes; the circle size represents gene number, nodes with shared genes are connected by blue lines (edges), and all nodes are connected, annotated and clustered using the MCL algorithm. NES = normalised enrichment score, WT = wild type, MT = *ray* mutant, MSC = mucilage secretory cells, INT = integument, 2 = 2 days post anthesis, 4 = 4 days post anthesis.

Differential expression of *P. ovata* genes and gene set enrichment analysis between two tissue types (Comparisons X-XVIII)

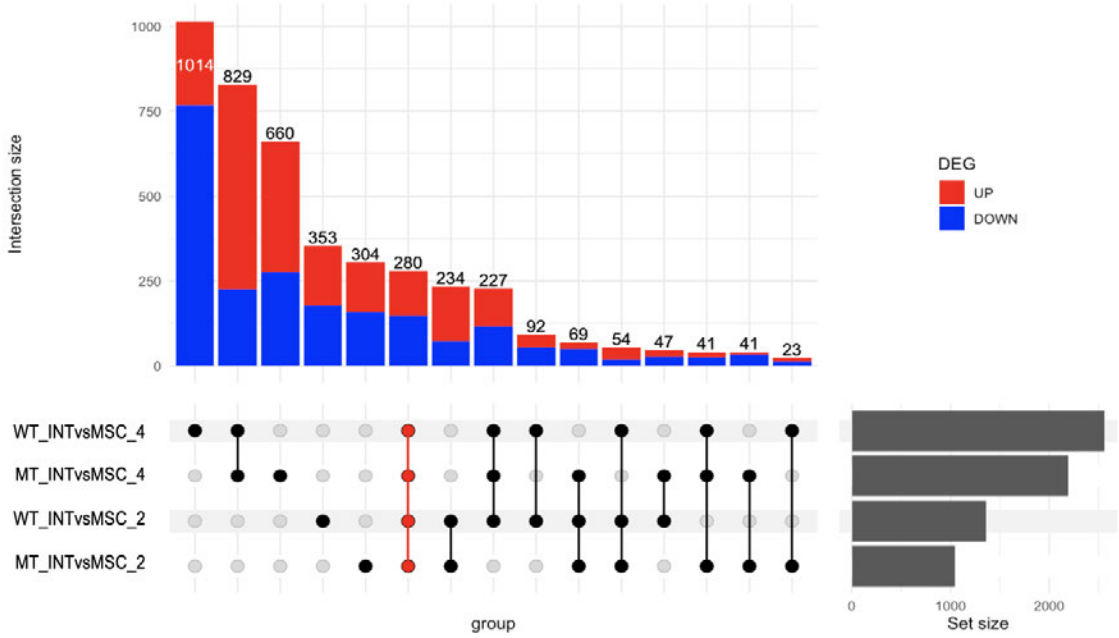


Figure 7. ComplexUpset plot visualises the number of DEGs in each comparison between INT and MSC tissues (Comparisons X – XVIII).

Overall, the total number of DEGs between the tissue types range from 1,046 (Comparison XVII) to 3,442 (Comparison XII) (Table 6). In Figure 7, WT at 4 DPA ("WT_INTvsMSC_4", Comparison XV) has 354 specific DEGs more than *ray* at the same age ("MT_INTvsMSC_4", Comparison XVIII), 661 more DEGs than samples at 2 DPA from WT ("WT_INTvsMSC_2", Comparison XIV), and 734 more DEGs than samples at 2 DPA from *ray* ("MT_INTvsMSC_2", Comparison XVII). More down-regulated DEGs were found in each comparison, including DEGs from WT at 4 DPA (Comparison XV) (Table 6). For example, 76% of the 1,014 DEGs specific to Comparison XV ("WT_INTvsMSC_4") were down-regulated (Figure 7). In contrast, *ray* at 4 DPA (Comparison XVIII) had a greater number of up-regulated DEGs at over 50% (Table 6 and Figure 7).

There were 280 common genes whose expression differs between INT and MSC tissues (Figure 7). Five of them were identified as transcription factors. They encoded homeobox-leucine zipper protein HDG5 (KN361_1g002634), homeobox-leucine zipper protein MERISTEM L1 (KN361_4g005596), heat stress transcription factor A-1e (KN361_1g004967), transcription factor MYB61 (KN361_4g000850), and floral homeotic protein DEFICIENT (KN361_1g002122). Two pathways were identified as being related to these common genes. Two genes are involved in CAM (Crassulacean acid metabolism, M00168) and two in beta-Oxidation, acyl-CoA synthesis (M00086).

The expression levels of these five transcription factors were compared in Figure 8. KN361_1g002122 is the only TF with higher expression in INT than MSC in WT and *ray*. The other four TFs were significantly higher in MSC than INT (P=0.029), especially for KN361_1g004967. This TF (KN361_1g004967) was predicted to encode heat stress transcription factor A-1e. The transcript levels of KN361_1g004967 in CPM (count per million) were 38.60 (WT INT 2 DPA), 59.17 CPM (WT INT 4 DPA), 556.25 (WT MSC 2 DPA), and 795.60 (WT MSC 4 DPA). The expression of KN361_1g004967 was upregulated in MSC tissue and increased from 2 to 4 DPA. The same pattern was observed in *ray*. 44.51 (*ray* INT 2 DPA), 63.39 CPM (*ray* INT 4 DPA), 393.50 (*ray* MSC 2 DPA), and 748.66 (*ray* MSC 4 DPA). This gene is listed in the top ten upregulated genes between INT and MSC. Overall, the LogFC of this gene was 4.08 in WT while 3.52 in *ray*.

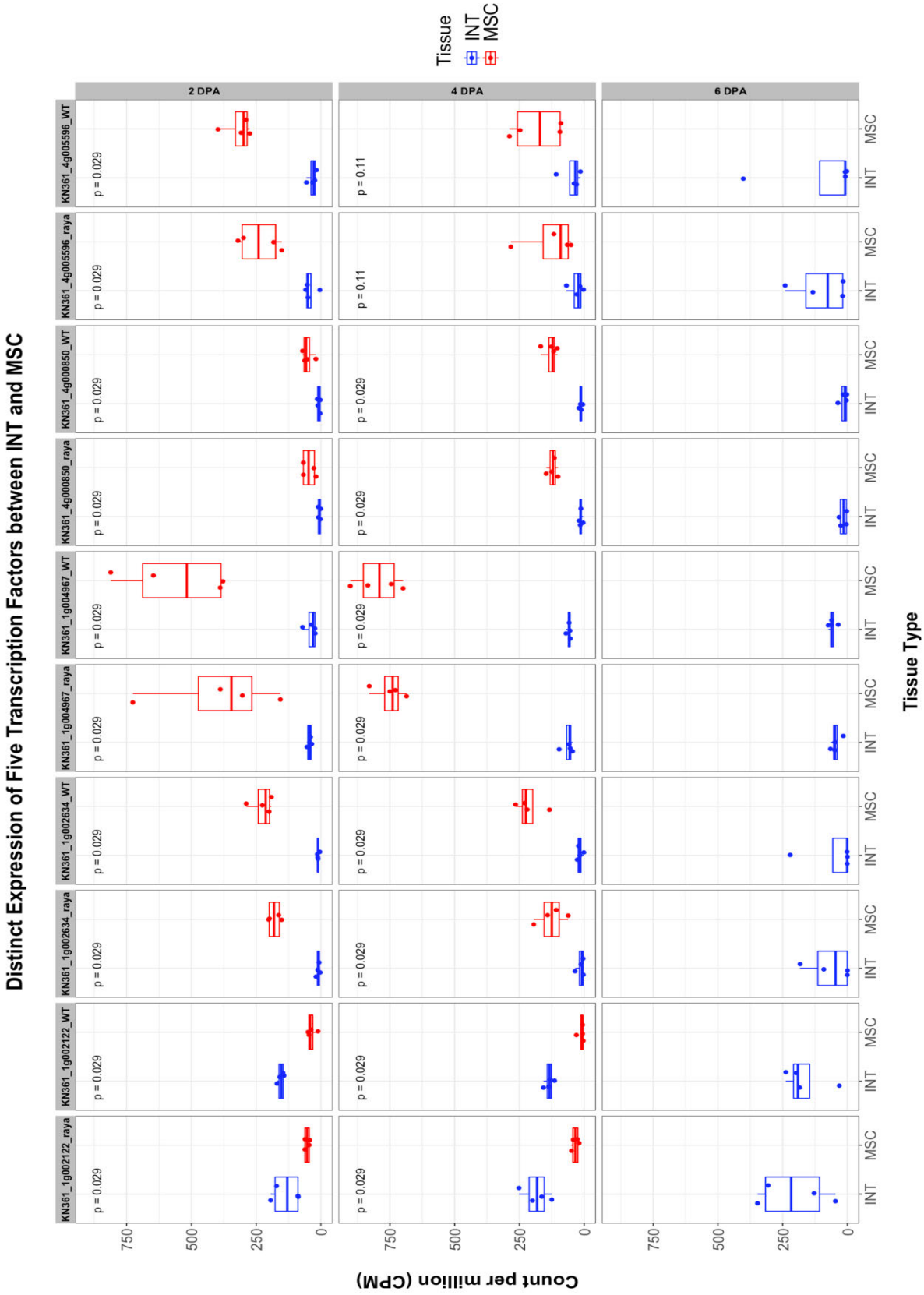


Figure 8 Expression levels of five transcription factors differ between integument (INT) and MSC tissues in WT and *ray* levels of five transcription factors differ between integument (INT) and MSC tissues in WT and *raya*.

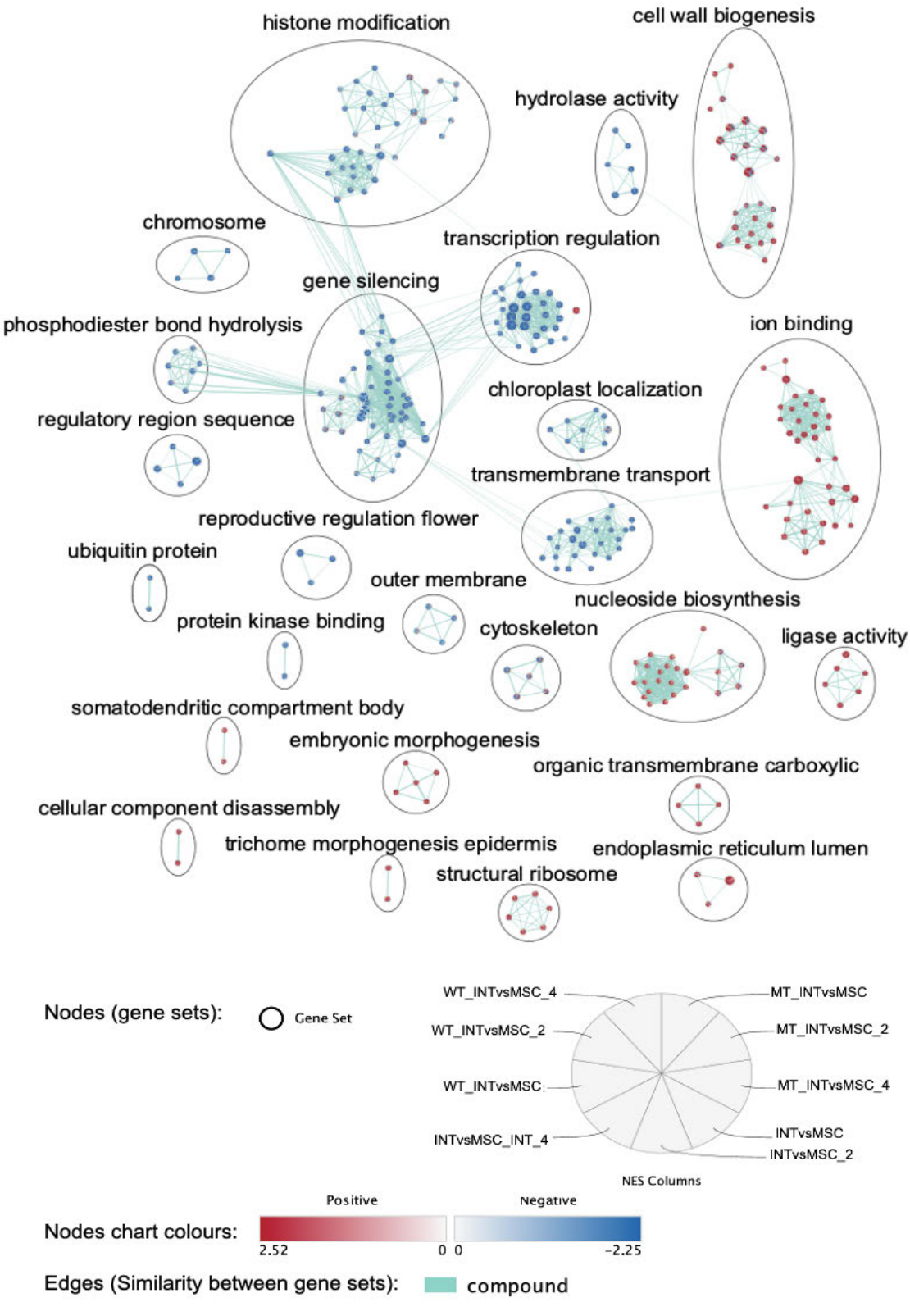

histone modification
cell wall biogenesis
hydrolase activity
chromosome
transcription regulation
gene silencing
phosphodiester bond hydrolysis
ion binding
chloroplast localization
regulatory region sequence
transmembrane transport
reproductive regulation flower
ubiquitin protein
outer membrane
nucleoside biosynthesis
protein kinase binding
cytoskeleton
ligase activity
somatodendritic compartment body
embryonic morphogenesis
organic transmembrane carboxylic
cellular component disassembly
trichome morphogenesis epidermis
endoplasmic reticulum lumen
structural ribosome
Nodes (gene sets):
Gene Set
WT_INTvsMSC_4
MT_INTvsMSC
WT_INTvsMSC_2
MT_INTvsMSC_2
WT_INTvsMSC
MT_INTvsMSC_4
INTvsMSC_INT_4
INTvsMSC
INTvsMSC_2
NES Columns
Positive
Negative
Nodes chart colours:
2.52
0 0
-2.25
Edges (Similarity between gene sets):
compound

Figure 9. Gene ontology (GO) term enrichment analysis on gene sets from comparisons X-XVIII, comparing two tissues. Network enrichment analysis was built using all expressed genes in all tissues processed using GSEA and then visualised using Cytoscape v3.8.0 (FDR q value > 0.1). Big circles represent clusters. Each cluster consists of two or more nodes (gene sets). Each node slice represents a gene set (GO term) from each comparison. Red indicates enriched upregulated genes, and blue shows enriched downregulated genes; the circle size represents gene number, nodes with shared genes are connected by blue lines (edges), and all nodes are connected, annotated and clustered using the Community cluster (GLay) algorithm. NES = normalised enrichment score, WT = wild type, MT = *ray* mutant, MSC = mucilage secretory cells, INT = integument, 2 = 2 days post anthesis, 4 = 4 days post anthesis.

There were 25 clusters of gene sets or GO terms (Figure 9 and Supplementary Table 4) identified in this set of comparisons. The WT (Comparisons XIII, XIV, and XV) had more gene sets (up to 4 times as many) than *ray* (Comparisons XVI, XVII, and XVIII) regardless of time point. Both WT and *ray* had a smaller number of gene sets at 2 DPA (Comparisons XIV and XVII) than at 4 DPA (Comparisons XV and XVIII), and *ray* at 2 DPA (Comparison XIV) had the smallest number, at only 18 gene sets. The cell wall biogenesis cluster was upregulated significantly except in 2 DPA samples (Comparisons XI, XIV, and XVII). This cluster consists of 27 gene sets (GO terms) (Table 7). The highest number of genes in this cluster was classified in GO:0005794 (CC Golgi apparatus) with 215 genes (Table 7). The leading-edge subset members for the cell wall biogenesis cluster are listed in Table 8.

Differential expression of *P. ovata* genes and gene set enrichment analysis between two-time points (Comparisons XIX–XXVII)

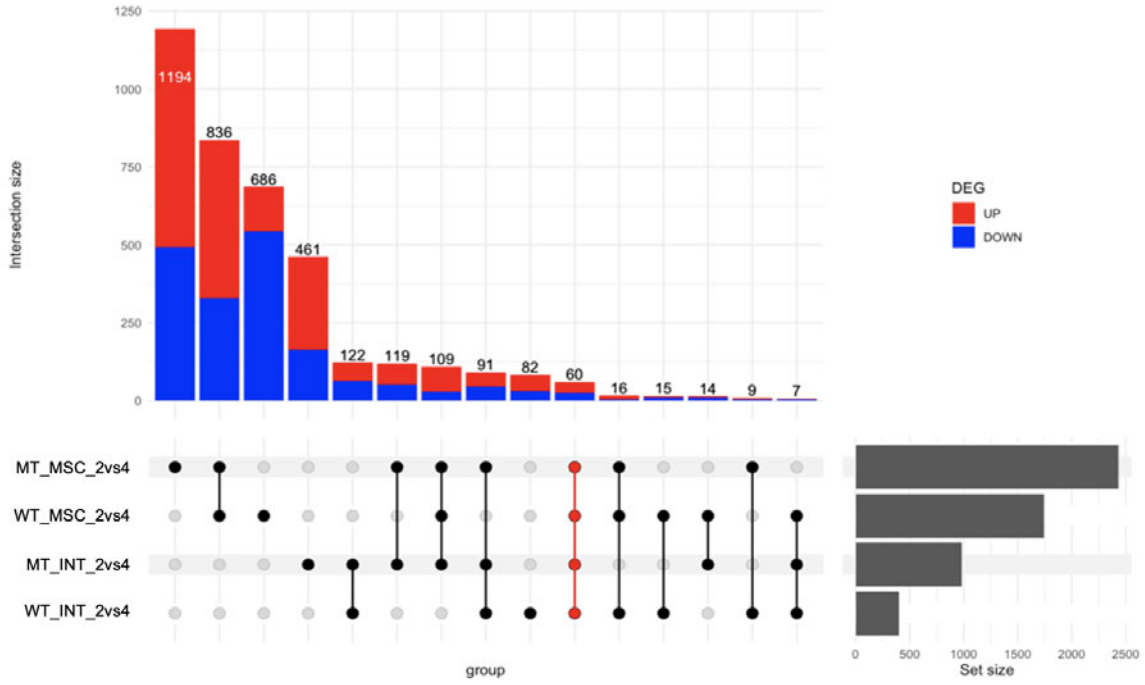


Figure 10. ComplexUpset plot visualises the number of DEGs between 2 and 4 DPA samples (Comparisons XIX–XXVII).

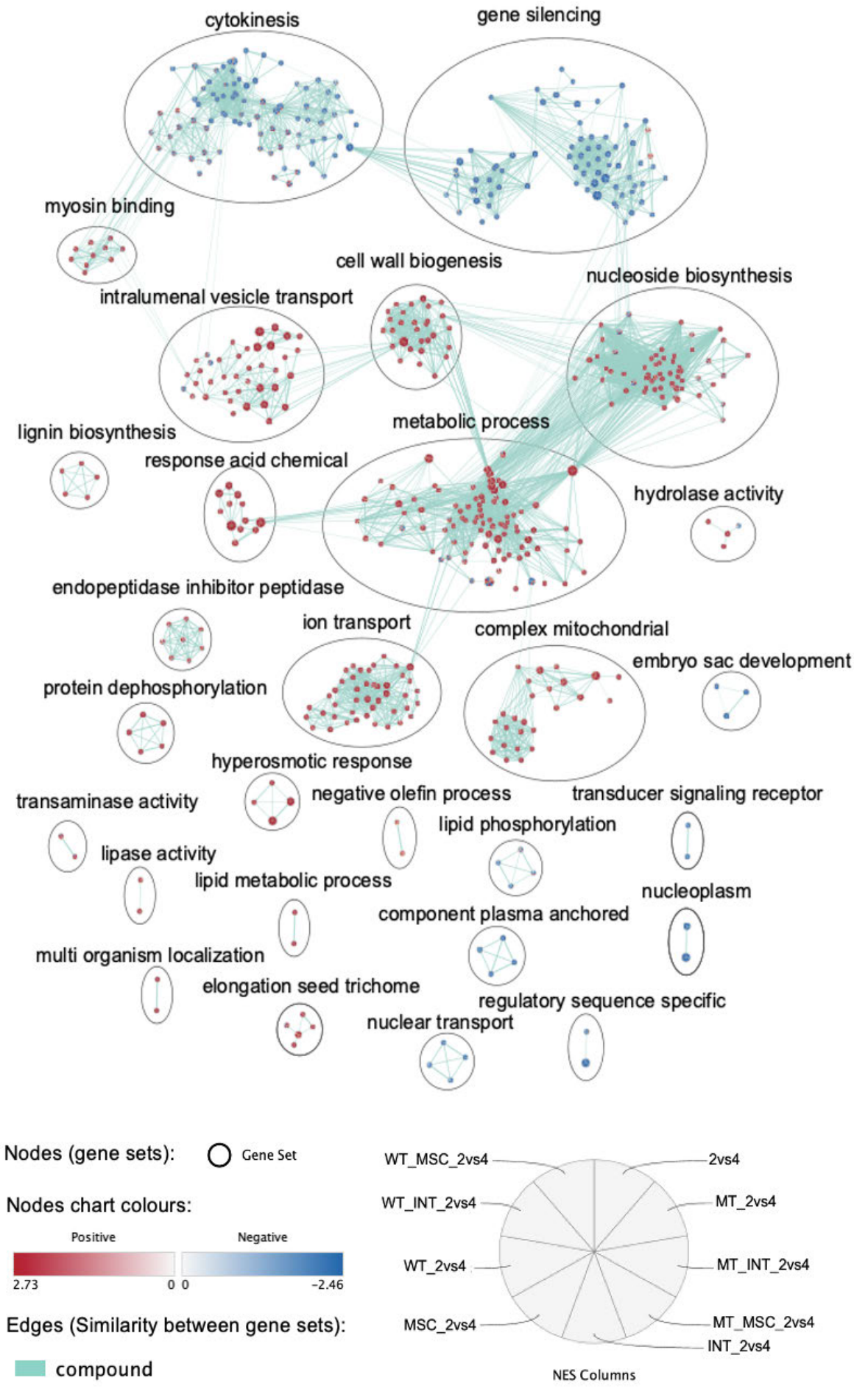

cytokinesis
gene silencing
myosin binding
cell wall biogenesis
nucleoside biosynthesis
intralumenal vesicle transport
lignin biosynthesis
metabolic process
response acid chemical
hydrolase activity
endopeptidase inhibitor peptidase
ion transport
complex mitochondrial
embryo sac development
protein dephosphorylation
hyperosmotic response
transaminase activity
negative olefin process
transducer signaling receptor
lipid phosphorylation
lipase activity
lipid metabolic process
nucleoplasm
component plasma anchored
multi organism localization
elongation seed trichome
regulatory sequence specific
nuclear transport
Nodes (gene sets):
Gene Set
Nodes chart colours:
Positive
Negative
2.73
0 0
-2.46
Edges (Similarity between gene sets):
compound
WT_MSC_2vs4
2vs4
WT_INT_2vs4
MT_2vs4
WT_2vs4
MT_INT_2vs4
MSC_2vs4
MT_MSC_2vs4
INT_2vs4
NES Columns

Figure 11. Gene ontology (GO) term enrichment analysis for gene sets from comparisons XIX-XXVII, comparing two time points. Network enrichment analysis was built using all expressed genes in all tissues processed using GSEA and then visualised using Cytoscape v3.8.0 (FDR q value > 0.1). Big circles represent clusters. Each cluster consists of two or more nodes (gene sets). Each node slice represents a gene set (GO term) from each comparison. Red indicates enriched upregulated genes, and blue shows enriched downregulated genes; the circle size represents gene number, nodes with shared genes are connected by blue lines (edges), and all nodes are connected, annotated and clustered using the Community cluster (GLay) algorithm. NES = normalised enrichment score, WT = wild type, MT = *ray* mutant, MSC = mucilage secretory cells, INT = integument, 2 = 2 days post anthesis, 4 = 4 days post anthesis.

Overall, the total number of DEGs in time-point comparisons range from 443 (Comparison XXIII) to 3,099 (Comparison XXI) (Table 9). Different tissues and genotypes had different patterns of up and down-regulated DEGs (Table 9). In Comparison XIX, when all samples were included, regardless of tissue and genotype, more genes (53.79%) were downregulated across this time point. This trend was followed by Comparison XXI (MSC), XXII (WT), and XXIV (WT MSC). In contrast, Comparison XX (INT), XXIII (WT INT), XXV (*ray*), XXVI (*ray* INT), and XXVII (*ray* MSC) (Table 9).

Figure 10 shows that MSC layers from *ray* ("MT_MSC_2vs4", Comparison XXVII) had 1,194 DEGs across two time points unique to this group, whereas the MSC layers from WT ("Wildtype_MSC_2vs4", Comparison XXIV) have about 686 DEGs. Downregulated genes account for 80% of the 686 DEGs in the MSC layer WT (Comparison XXIV), indicating that many of them were less transcriptionally active at 4 DPA as compared to just two days earlier. Sixty common genes were differentially expressed in all four comparisons, with 36 up and 24 down-regulated (Figure 10). No transcription factors or pathways were identified in this set of common DEGs but there were three genes identified as glycosyltransferases (ko01003). They

were UDP-arabinopyranose mutase 1 (KN361_4g000651), cellulose synthase-like protein D5 (KN361_4g003626), and cellulose synthase-like protein D3 (KN361_1g002613).

There were thirty clusters (Figure 11 and Supplementary Table 5) from Comparisons XIX-XXVII. Across these comparisons (Table 9), the number of clusters varies from 6 (Comparison XXIII) to 19 (Comparison XXII). The number of gene sets in each cluster ranges from 2 to 91 (Supplementary Table 5). The metabolic processes cluster had the highest number of gene sets (91) with at least 6 clusters directly connected (Figure 11). Only one connected cluster shows a downregulated trend (gene silencing cluster) while the other five connected clusters were upregulated. The upregulated clusters were nucleoside biosynthesis, ion transport, intralumenal vesicle transport, cell wall biogenesis and complex mitochondria (Figure 11). These five upregulated clusters were absent in the WT integument (Comparison XXIII) (Supplementary Table 5). The cell wall biogenesis cluster contains thirty gene sets (GO terms), with GO:0005794 (CC Golgi apparatus) having the highest number of genes (215 genes) (Table 10). A subset of the genes in the cell wall biogenesis cluster with their protein product is listed in Table 11.

Tissue-, stage- and genotype-specific expression of selected glycosyltransferase genes during early seed development

Selected glycosyltransferases studied by Jensen et al. (2013) and Phan et al. (2016) involved in xylan backbone biosynthesis (*IRX7*, *IRX9*, and *IRX14*) and addition of backbone substitutions (*GT61* genes) were investigated here.

Six *PoIRX10* genes have been identified in the *P. ovata* genome (Chapter 2). Two genes (KN361_2g006239 and KN361_4g000130) were not expressed in any of the samples analysed here (Figure 12). The KN361_2g006638 gene was expressed in both INT and MSC of WT and *ray* whilst both KN361_4g000131, and KN361_4g000136 were expressed only in WT and *ray* MSC tissues, where the latter had a notably higher CPM count in the mutant tissue versus the

WT. A single gene, KN361_4g000139, was specifically expressed in *ray* MSC samples only (Figure 12), even if only at a low level.

Two *PoIRX7*, three *PoIRX9* and one *PoIRX14* genes have been identified in the *P. ovata* genome (Chapter 2). As shown in Figure 13, all these genes have expression levels of below 100 CPM in INT samples from both genotypes, while levels can reach 400 CPM in MSC samples. Of the two *PoIRX7* genes, one is absent or very minimally represented in any of the samples (KN361_2g002752) whilst the other, KN361_3g003100 although present in INT at low levels, features in the MSC tissue and shows an increase of almost double in *ray* MSCs versus WT (Figure 13). Of the three *PoIRX9* genes, two (KN361_3g003370 and KN361_4g003968) were represented in all tissues at only low levels with higher levels of KN361_1g004593 in MSC samples, also elevated in *ray* versus WT. The single *PoIRX14* gene shows the highest CPM counts of any of the genes in this grouping, in the MSC samples, particularly in *ray* where levels were over 400 CPM and were again elevated over WT levels (Figure 13).

There were 19 *PoGT61* genes that have been identified in the *P. ovata* genome (Chapter 2). All *PoGT61* genes showed almost no expression in INT tissues, while eight were expressed in the MSC layers (Figure 14). The expression levels of seven of the genes were appreciable, with low CPM for only *PoGT61_17.* All seven of the highly expressed genes showed increased CPM in the *ray* MSC tissues, versus the WT samples, and two of the genes, *PoGT61_1L* and *PoGT61_6* exceed 2,000 CPM in *ray* samples (Figure 14).

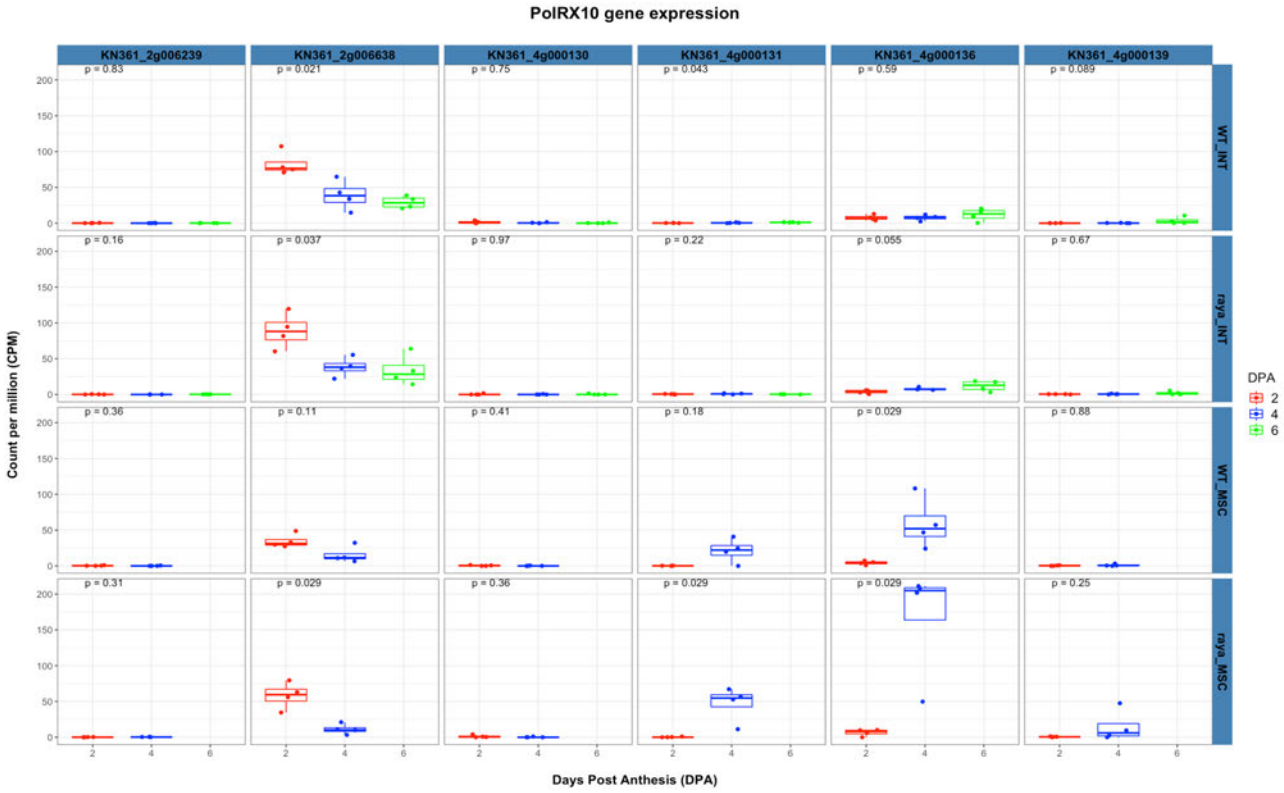


Figure 12. Expression of six *PoIRX10* genes in normalised CPM (Count Per Million) presented for the integument (INT) and mucilage secretory cells (MSC) from two genotypes from 2 to 6 DPA. No RNA was extracted from MSC samples at 6 DPA.

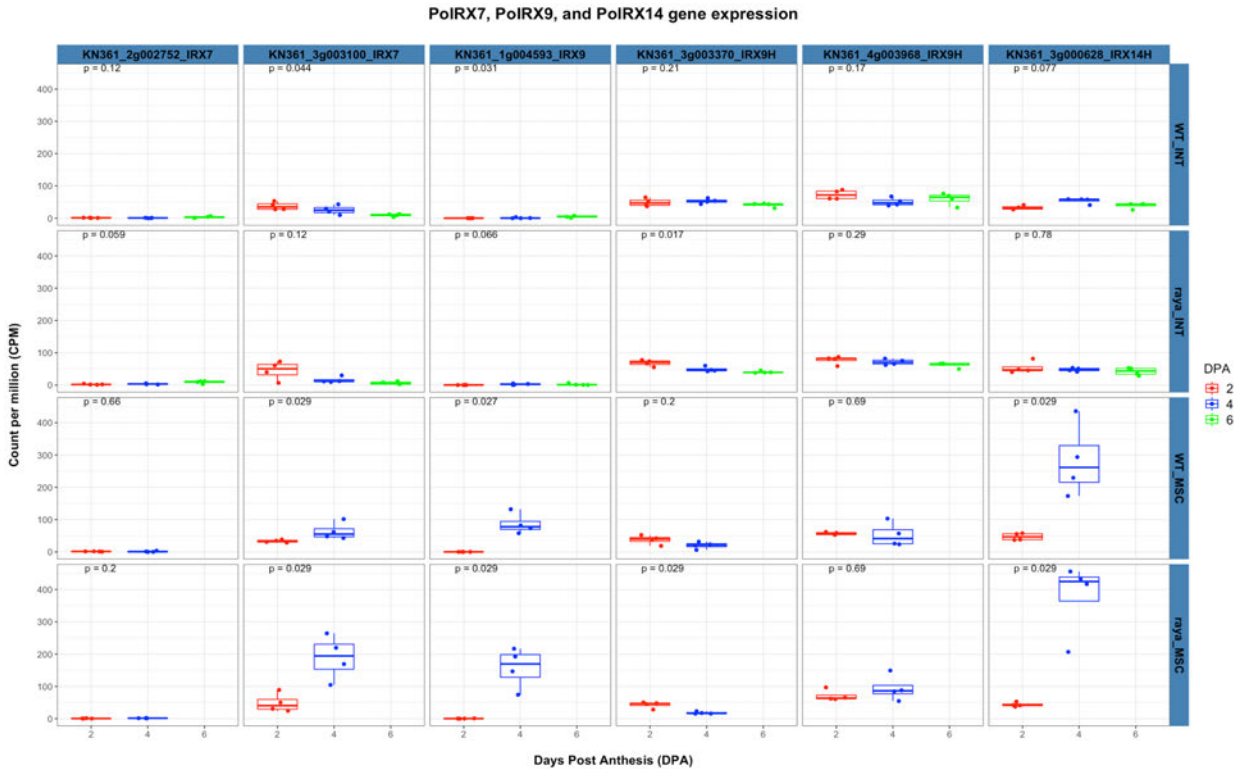


Figure 13. Expression of two *PoIRX7*, three *PoIRX9*, and one *PoIRX14* gene in normalised CPM (Count Per Million) presented for the integument (INT) and mucilage secretory cells (MSC) from two genotypes from 2 to 6 DPA. No RNA was extracted from MSC samples at 6 DPA.

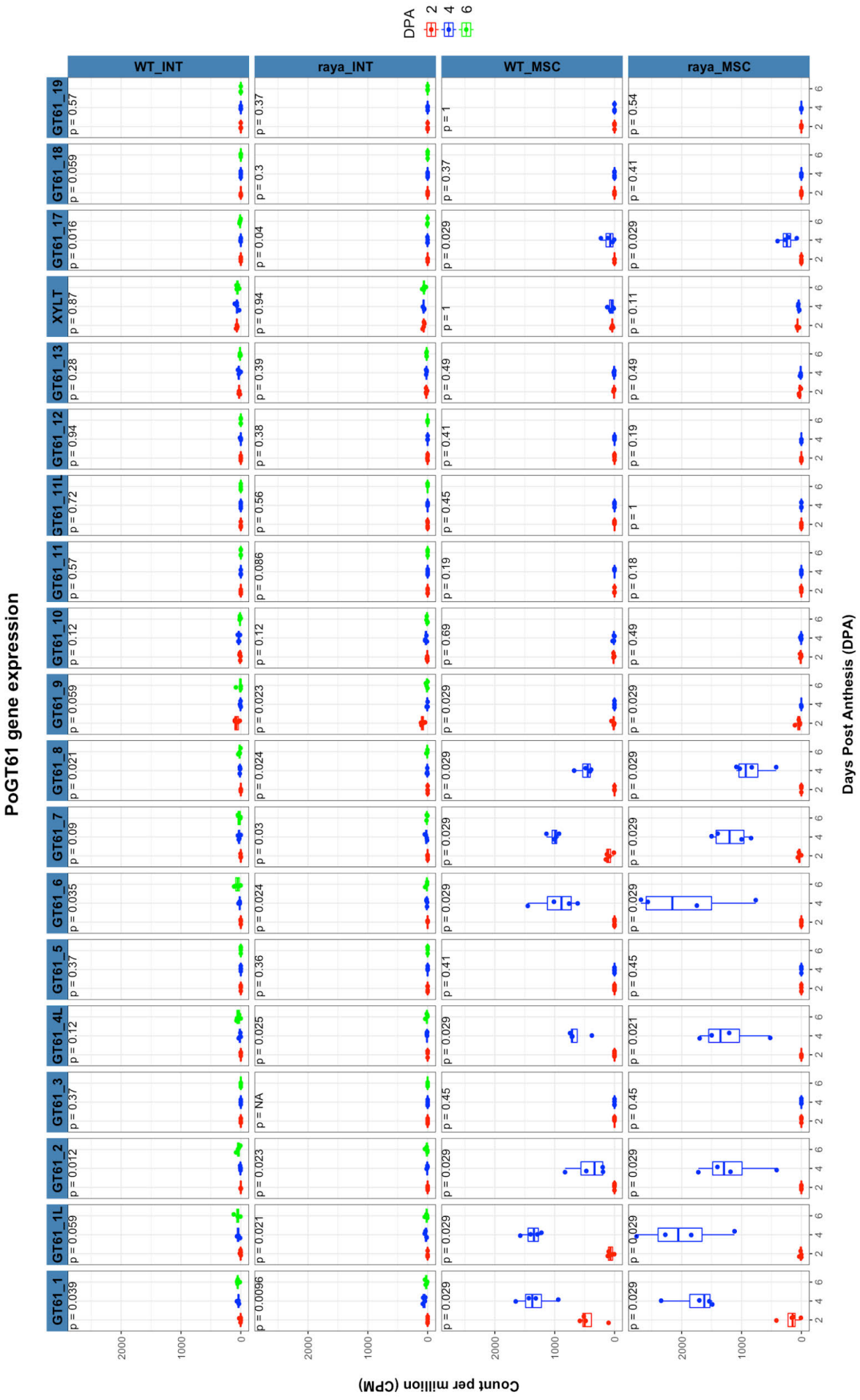

PoGT61 gene expression
DPA 2 4 6
WT_INT
raya_INT
WT_MSC
raya_MSC
GT61_19 p = 0.57 p = 0.37 p = 1 p = 0.54
GT61_18 p = 0.059 p = 0.3 p = 0.37 p = 0.41
GT61_17 p = 0.016 p = 0.04 p = 0.029 p = 0.029
XYLT p = 0.87 p = 0.94 p = 1 p = 0.11
GT61_13 p = 0.28 p = 0.39 p = 0.49 p = 0.49
GT61_12 p = 0.94 p = 0.38 p = 0.41 p = 0.19
GT61_11L p = 0.72 p = 0.56 p = 0.45 p = 1
GT61_11 p = 0.57 p = 0.086 p = 0.19 p = 0.18
GT61_10 p = 0.12 p = 0.12 p = 0.69 p = 0.49
GT61_9 p = 0.059 p = 0.023 p = 0.029 p = 0.029
GT61_8 p = 0.021 p = 0.024 p = 0.029 p = 0.029
GT61_7 p = 0.09 p = 0.03 p = 0.029 p = 0.029
GT61_6 p = 0.035 p = 0.024 p = 0.029 p = 0.029
GT61_5 p = 0.37 p = 0.36 p = 0.41 p = 0.45
GT61_4L p = 0.12 p = 0.025 p = 0.029 p = 0.021
GT61_3 p = 0.37 p = NA p = 0.45 p = 0.45
GT61_2 p = 0.012 p = 0.023 p = 0.029 p = 0.029
GT61_1L p = 0.059 p = 0.021 p = 0.029 p = 0.029
GT61_1 p = 0.039 p = 0.0096 p = 0.029 p = 0.029
0 1000 2000
2 4 6
Count per million (CPM)
Days Post Anthesis (DPA)

Figure 14. Expression of nineteen *PoGT61* genes in normalised CPM (Count Per Million) presented for the integument (INT) and mucilage secretory cells (MSC) from two genotypes from 2 to 6 DPA. No RNA was extracted from MSC samples at 6 DPA

Discussion

A comprehensive transcriptome profile of developing seed tissues, specifically the MSC layer and INT tissues, obtained using a combination of cryo-sectioning, LCM, and RNA sequencing, reveals enormous diversity in gene expression associated with tissue type and developmental stage. We showed that a robust and comprehensive RNA-seq experiment with a tiny amount of starting material with optimised steps could produce meaningful data.

Transcription factors may regulate mucilage production

A number of transcription factors involved in controlling mucilage polysaccharide production have been identified in Arabidopsis (Johnson et al., 2002; Debeaujon et al., 2003; Shi et al., 2012; Wang et al., 2014). However, the proposed transcriptional network regulating MSC differentiation and mucilage production in Arabidopsis (Golz et al., 2018) is not universal or conserved among mucilage-producing plants, especially for plants that do not have a columella system in the seed coat nor a pectin-rich mucilage. In fact, only one out of the five transcription factors enriched in *P. ovata* MSC tissues at two time points and for both genotypes are listed by Golz et al. (2018). The TF common to both plants is *MYB61*, while the putative mucilage-related TFs found only in *P. ovata* are the homeobox-leucine zipper protein MERISTEM L1 (*ML1*), HOMEDOMAIN GLABROUS5 (*HDG5*), heat stress transcription factor A-1e (*HSFA1E*), and *DEFICIENS*. *MYB61* is listed as a tier 2 regulator according to Golz et al. (2018), which means that its disruption impacts MSC differentiation but this is not as severe as tier 3 regulators that show significant defects in mucilage production, modification, and

adherence (Golz et al., 2018). *MYB61* is required for mucilage deposition and extrusion. A reverse genetic approach to defining *AtMYB61* function by Penfield et al. (2001) shows that *MYB61* is required for seed mucilage release during imbibition. Ruthenium red staining of *myb61* seeds shows a deficiency in mucilage extrusion upon hydration and significantly reduced levels of soluble polysaccharides such as rhamnose and galacturonic acid compared to the WT (Penfield et al., 2001). The manipulation of *MYB61* in *P. ovata* is therefore likely to yield valuable information regarding its role and regulation of heteroxylan biosynthesis in particular. This gene is a prime candidate for further studies, either for editing via CRISPR-Cas9 to downregulate it, or for overexpression under the control of a suitable promoter. A global knockout may well be lethal so the use of a seed specific promoter, such as the ones described by McGee et al. (2019), might be the best approach.

The remaining four TF's are less well characterised but are also good choices to manipulate to understand their potential role in mucilage polysaccharide biosynthesis. *ML1* is an HD-ZIP IV family transcription factor required for epidermis specification (Rombolá-Caldentey et al., 2014). *ATML1* and *PDF2* bind to an L1 box sequence of *LIP1* promoters. Expression of *LIP1* is highly induced by gibberellic acid (GA) and repressed by DELLA proteins (Rombolá-Caldentey et al., 2014). Silencing of both HD-ZIP transcription factors inhibits epidermal gene expression and delays germination. In the presence of GA, DELLA was destabilised to release *ATML1/PDF2* to activate L1 box gene expression (Rombolá-Caldentey et al., 2014). Mutation in the DELLA proteins increases the size of the mucilage halo surrounding the seed of *della*. Since mutation of DELLA does not affect transcription factors assigned in tiers 1-3, Golz et al. (2018) classified them as gibberellic acid (GA) pathways independent of all three-tier regulators. The work of Bueso et al. (2014) showed positive correlations between GA, seed longevity, and seed mucilage formation. No report shows the direct impact of *ATML1* on mucilage production. However, the interaction between this TF with DELLA and the GA pathway indicates it may have a role in mucilage production.

A sequence similar to *GLABRA2* is one of 14 TFs identified specifically in the higher mucilage yielding *ray* and has been well studied in Arabidopsis. Plants carrying mutations in this TF produce seeds with many mucilage deficiencies, such as no mucilage on *gl2* seeds (Shi et al., 2012; Wang et al., 2014). Xu et al. (2022) provided evidence of physical interaction between *GL2* and *DF1* to activate *MUM4* and *GATL5* for pectin biosynthesis. No study has reported direct links between the other 13 TF identified here (WT - Table2, *ray* - Table 3) with mucilage production, but they may be more indirect. For example, bZIP63 has been implicated with sugar transport in response to starvation and *CSA* was involved in sugar partitioning from leaves to anthers during male reproduction. The expression of these two genes may increase the sugar supply to the seed coat for mucilage biosynthesis. From results obtained using H^1-NMR analyses of metabolites, Miart et al. (2021) suggested that the main sugar supplies in early to mid-stage flax seed development were used for mucilage biosynthesis in the seed coat. Thus, these TFs are promising candidate genes for controlling mucilage biosynthesis.

Distinct differences between WT and *ray* in the cell wall metabolic cluster may link to increased mucilage polysaccharide production

The gene expression profile of *ray* was distinct from WT especially for integument tissue at 2 DPA. In the comparison between the two genotypes, the cell wall metabolic cluster was upregulated significantly for *ray* INT compared to WT INT at 2 DPA (Comparison V). The upregulation of this cluster is not present in other comparisons (I-IV and IX) but it shows a down regulated trend in the MSC layer especially at 2 DPA (Comparison VII and IX). Comparison between two tissues (Comparison X-XVIII) shows that the cell wall biogenesis cluster is not enriched in Comparison XI (2 DPA), XIV (WT 2 DPA), and XVI (*ray* 2 DPA). In addition, comparison XVII (*ray* at 2 DPA) has the smallest number of DEGs, gene sets, and clusters (Table 5). Comparisons between two time points (XIX–XXVII) show that only integument WT (Comparison XXIII) does not have an enriched cell wall biogenesis cluster. In addition, Comparison XXIII has the smallest number of DEGs gene set and clusters. This means

the cell wall cluster in *ray* was already high in the integument tissue at 2 DPA compared to the WT, but not higher than the integument at 4 DPA and relatively similar to the MSC layer at 2 DPA of *ray* itself. In contrast, the cell wall cluster in WT was not enriched in the integument. These results support a central role for some of the genes in the cell wall biogenesis cluster in driving mucilage polysaccharide yield.

Four gene families are identified in the leading gene set of the cell wall cluster that might drive differences between WT and *ray*. They are glycosyltransferases, methyltransferases, acetyltransferases and glycoside hydrolases (Table 5). It has previously been suggested that Glycosyltransferase family 47 (IRX10 and IRX10L) proteins are responsible for xylan backbone elongation in *P. ovata* seed mucilage without interacting with GT43 members *(IRX9*, IRX9H, IRX14, and IRX14H) (Jensen et al., 2013). Our results indicate this may be true in the integument layer but is not the case in MSCs as at least four out of six GT43 genes found in the genome (Chapter 3) were expressed in these tissues (Figure 13). Maximum expression of GT43 (400 CPM) (Figure 13) was up to two times the expression of GT47 (200 CPM) (Figure 10). Previous studies by Jensen et al. (2013) using hand sectioning to isolate the mucilaginous layer and Kotwal et al. (2016) who used whole ovaries, reported either no or only low expression of GT43 genes in their samples. This is likely to be due to the dilution of the tissue-specific expression of genes from the single layer of MSCs by all the transcripts in the multiple tissue types present in the starting tissue. This also applies to glycosyltransferase 61 (*GT61*) genes in *P. ovata,* where high expression was found in the MSC layer up to 3,000 CPM but not in integument layers. Overall, expression of *GT43*, *GT47*, and *GT61* genes were higher in *ray* compared to WT and this may be due to the generation of more mucilage polysaccharides in *ray*. Overexpression of combinations of these genes in transgenic plants, potentially on a multi-gene plasmid, may be an option to phenocopy the *ray* mutant.

QUASIMODO2 (*QUA2*) is the only gene in the leading-edge subset of the cell wall cluster (Table 5) that has methyltransferase activity. It has been reported to be expressed in leaves,

seedlings, and root tips (Du et al., 2020; Kohorn et al., 2021). Recent work from Du et al. (2022) has also detected the expression of QUA2 in the seed coat and shown it affects mucilage production. It is reported that mucilage of *qua2* has a reduced pectin content of 50% and adhesion defects from leaf samples (Kohorn et al., 2021). Du et al. (2020) observed a reduction in crystalline cellulose, abnormal cellulose organization and movement of cellulose synthase complexes in *qua2*. Homogalacturonan (HG), which is the most abundant pectin component in Arabidopsis mucilage, is methylated in the Golgi apparatus before being secreted into the cell wall and interacting with cellulose (Du et al., 2020). Therefore, mutation of *QUA2* not only affects pectin content but also cellulose content. *Trichome birefringence-like 35 TBL35* is also the only gene in the leading-edge subset of the cell wall cluster (Table 5) that has acetyltransferase activity. Even though *TBL35* belongs to Domain Unknown Function (DUF) 231 protein family, it has been proven to have roles in xylan acetylation as mutation of this gene caused a specific reduction in xylan 3-O-monoacetylation and 2,3-di-O-acetylation (Yuan et al., 2016). Two genes in the leading-edge subset of the cell wall cluster (Table 4) have glycoside hydrolase activity, namely *alpha-L-arabinofuranosidase 1 ASD1* and *mannan endo-1,4-beta-mannosidase 2 MAN2*. Heterologous expression of *ASD1* and *ASD2* was able to rescue mucilage extrusion from *sh1 shp2* seeds (David 2008). Even though there was evidence of mannan biosynthesis in the *P. ovata* mucilage layer studied by Jensen et al. (2013), they speculated that the results were due to endosperm contamination in their samples. Mannan is found in Arabidopsis mucilage in galactoglucomannan and studies of mutants have shown that it plays an important scaffolding role (Yu et al. 2014; Voiniciuc et al., 2015). Varying amounts of mannan were also detected in whole mucilage of *Plantago* species (Phan et al., 2016) and the pure nature of our starting tissues suggests that the presence of genes for mannan biosynthesis may not in fact be due to contamination at all. Further research is needed to confirm and explore the roles of *TBL35*, *ASD1* and *MAN2* in seed mucilage.

Cellular components enriched in cell wall and intralumenal vesicle transport clusters could link to mucilage biosynthesis and trafficking

Eight cellular components in the cell wall cluster (Table 7) were enriched in MSCs, especially at 4 DPA. They are Golgi apparatus (GO:0005794, 215 genes), Golgi apparatus subcompartment (GO:0098791, 95), trans-Golgi network (TGN) (GO:0005802, 85), vesicle (GO:0031982, 148), intracellular vesicle (GO:0097708, 146), cytoplasmic vesicle (GO:0031410, 146), endosome (GO:0005768, 128), and Golgi membrane (GO:0000139, 12). This list indicates that the carbohydrate or polysaccharide biosynthetic process involves several cellular compartments: the Golgi apparatus, vesicles, and vacuoles. Hyde (1970) observed mucilage deposition inside vacuoles and between the plasma membrane and cell wall accompanied by an increased number and size of Golgi vesicles in *P. ovata* MSCs. The deposition of pectinaceous mucilage in Arabidopsis has been well described by Young et al. (2008) and Meents et al. (2019) used fluorescently tagged IRX9 as a marker to track xylan production in Golgi of the developing tracheary elements of Arabidopsis. Cell wall polysaccharides, not only pectin, are synthesized in the Golgi and there is an increasing number of Golgi, with more stacks, in the developing seed coat cells, and Golgi were randomly distributed in the cell (Young et al., 2008; Meents et al., 2019). The trans-Golgi network (TGN) is the membrane compartment of the Golgi that is responsible for sorting and packaging cargo molecules targeted to the plasma membrane or vacuoles (Sinclair et al., 2018) and the plant TGN also acts as an early endosome (Sinclair et al., 2018). McFarlane et al. (2013) found out that mutation of ECHIDNA, encoding a TGN-localized protein, led to the mislocated accumulation of cell wall polysaccharides in the vacuole rather than in the apoplast, and so no mucilage was secreted from the *echidna* seeds. However, in *P. ovata*, this natural process was observed by Hyde (1970) as shown by these cellular components accumulating in the vacuole. This difference could explain why these two species have a different mechanism for releasing

mucilage, one being cell-free and the other relying on cellular rupture, as pointed out by Cowley and Burton (2021).

Conclusions and Future Directions

This study aimed to better understand the mechanisms controlling mucilage polysaccharide production during early seed development of *P. ovata*. Here, we provide valuable *P. ovata* RNA-seq datasets and comprehensive analyses based on them, with the potential to extend these analyses to focus on other gene families and processes occurring in these two cell types. There are many things to explore, for example the processes underlying transport from the integument to feed other developing seed tissues, followed by the degradation of the integument tissue which might, or might not, be driven by programmed cell death. Even though the data were generated from a tiny amount of tissue, the additional care taken to source high quality starting material by combining the cryosectioning procedure, laser capture microdissection, pico-scale RNA isolation, and unique molecular identifiers (UMIs) to eliminate PCR duplication, has made an enormous difference. In addition, we customised a ribodepletion kit to remove ribosomal RNAs in our samples and have used a number of free bioinformatics tools to reveal the biological meaning of our datasets and show that the expression of many genes varies across time points and tissue types. There is now a whole list of candidate transcription factors and genes from this study that will need to be validated using qPCR and *in situ* hybridisation, followed by transgenic analyses, to illuminate their role in polysaccharide synthesis and ultimately to provide breeding targets for greater mucilage yield.

Acknowledgments

The authors thank Dr Fabien Voisin for his technical support in utilising Phoenix-HPC. We acknowledge Dr Gwen Mayo and Melissa Pickering for their assistance in microscopy work and plant-growing facility. This study was supported by the Australian Research Council (ARC) Centres of Excellence in Plant Cell Walls (CE110001007), Plant Energy Biology (CE140100008) and Linkage Grant (LP180100971). This work was also supported by supercomputing resources provided by the Phoenix HPC service, Undercroft Glasshouse University of Adelaide, and Adelaide Microscopy at the University of Adelaide. We acknowledge the South Australian Genomics Centre (SAGC), which provides RNA sequencing. The SAGC is supported by the National Collaborative Research Infrastructure Strategy (NCRIS) via BioPlatforms Australia and the SAGC partner institutes. LH is supported by the University of Adelaide's Adelaide Graduate Research Scholarship (AGRS) and The National Research and Innovation Agency (BRIN-Indonesia).

Table 2. Sequences similar to known transcription factors expressed in WT, but not *ray*, at 2 and 4 DPA.

Types			Gene id	Protein product
Helix-turn-helix	Homeo domain other	K09338 HD-ZIP; homeobox-leucine zipper protein	KN361_2g000160	homeobox-leucine zipper protein ATHB-9;homeobox-leucine zipper protein HOX32
Helix-turn-helix	Tryptophan clusters Myb, Myb-factors	K09422 MYBP; transcription factor MYB, plant	KN361_2g001124	transcription factor MYB7
Helix-turn-helix	Tryptophan clusters Myb, Myb-factors	K09422 MYBP; transcription factor MYB, plant	KN361_2g003979	transcription factor MYB60

Table 3. Sequences similar to known transcription factors expressed in *ray*, but not WT, at 2 and 4 DPA.

Types			Gene id	Product
Basic leucine zipper (bZIP)	Plant G-box-binding factors	K14431 TGA; transcription factor TGA	KN361_1g000018	bZIP transcription factor 63
Basic leucine zipper (bZIP)	Plant G-box-binding factors	K14431 TGA; transcription factor TGA	KN361_3g003198	transcription factor TGA9
Helix-turn-helix	Homeo domain other	K09338 HD-ZIP; homeobox-leucine zipper protein	KN361_3g002323	homeobox-leucine zipper protein GLABRA 2
Helix-turn-helix	Tryptophan clusters Myb, Myb-factors	K09422 MYBP; transcription factor MYB, plant	KN361_1g000238	transcription factor MYB105;transcription factor CSA
Helix-turn-helix	Tryptophan clusters Myb, Myb-factors	K09422 MYBP; transcription factor MYB, plant	KN361_1g003087	transcription factor MYB73
Helix-turn-helix	Tryptophan clusters Myb, Myb-factors	K09422 MYBP; transcription factor MYB, plant	KN361_1g004499	transcription factor MYB36;transcription factor RAX2
Helix-turn-helix	Tryptophan clusters Myb, Myb-factors	K09422 MYBP; transcription factor MYB, plant	KN361_2g002144	transcription factor MYB26
Helix-turn-helix	Tryptophan clusters Myb, Myb-factors	K09422 MYBP; transcription factor MYB, plant	KN361_2g005689	Myb-related protein 306
Beta-Scaffold factors with minor groove contacts	MADS-box regulators of differentiation, Homeotic genes	K09264 K09264; MADS-box transcription factor, plant	KN361_3g002179	MADS-box protein SVP;MADS-box protein JOINTLESS

Beta-Scaffold factors with minor groove contacts	Heteromeric CCAAT factors	K08064 NFYA, HAP2; nuclear transcription factor Y, alpha	KN361_1g003582	nuclear transcription factor Y subunit A-8;nuclear transcription factor Y subunit A-5;nuclear transcription factor Y subunit A-3
Other transcription factors	EREBP	K09286 EREBP; EREBP-like factor	KN361_1g007592	ethylene-responsive transcription factor ERF053
Other transcription factors	EREBP	K09286 EREBP; EREBP-like factor	KN361_4g002229	ethylene-responsive transcription factor 5
Other transcription factors	EREBP	K09286 EREBP; EREBP-like factor	KN361_4g004986	ethylene-responsive transcription factor 12
Other transcription factors	EREBP	K14516 ERF1; ethylene-responsive transcription factor 1	KN361_2g009196	ethylene-responsive transcription factor 1B

Table 4. Pairwise RNA-seq comparisons between two genotypes.

Comparison (WT vs *ray*)	Up	Down	Total DEGs	Gene sets*	Cluster*
I (All samples)	5	3	8	0	0
II (2 DPA)	2	2	4	16	3
III (4 DPA)	2	2	4	2	1
IV (INT)	2	2	4	0	0
V (INT 2 DPA)	1	1	2	5	2
VI (INT 4 DPA)	5	1	6	0	0
VII (MSC)	6	2	8	1	1
VIII (MSC 2 DPA)	0	1	1	11	4
IX (MSC 4 DPA)	2	1	3	4	1

* Gene sets and clusters are summarised from Supplementary Table 3 and Figure 5.

Table 5. A leading-edge subset of genes in cell wall metabolic cluster.

Gene ID	Protein product	Protein family/activity
KN361_4G003968	putative beta-1,4-xylosyltransferase IRX9H	Glycosyltransferase family 43 (GT43)
KN361_3G003100	putative glucuronoxylan glucuronosyltransferase IRX7	GT47
KN361_1G008021	putative galacturonosyltransferase 12 GAUT12/IRX8	GT8
KN361_1G004062	4-beta-mannosidase 2 MAN2	Glycoside Hydrolase Family 5 (GH5)
KN361_4G004071	putative pectin methyltransferase QUA2/TSD2	Methyltransferase
KN361_4G007423	beta-1,3-galactosyltransferase 11 HPGT1/B3GALT11	GT31
KN361_1G002262	protein trichome birefringence-like 35 TBL35	Acetyltransferase
KN361_3G004832	beta-1,3-galactosyltransferase 11 HPGT1/B3GALT11	GT31
KN361_2G000201	beta-1,3-galactosyltransferase 20 GALT2/B3GALT20	GT31
KN361_3G004508	putative galacturonosyltransferase 14 GAUT14	GT8
KN361_4G005186	beta-1,3-galactosyltransferase 20 GALT2/B3GALT20	GT31
KN361_3G003370	putative beta-1,4-xylosyltransferase IRX9H	GT43
KN361_1G007839	alpha-L-arabinofuranosidase 1 ASD1	GH5
KN361_1G011348	xyloglucan galactosyltransferase XLT2	GT47
KN361_2G006638	putative beta-1,4-xylosyltransferase IRX10L	GT47
KN361_1G000109	alpha-L-arabinofuranosidase 1 ASD1	GH51

Table 6. Pairwise RNA-seq comparisons between two tissue types.

Comparison (INT vs MSC)	Up	Down	Total DEGs	Gene sets	Cluster*
X (All samples)	934 (38.25%)	1,508 (61.75%)	2,442	163	18
XI (2 DPA)	877 (45.39%)	1,055 (54.60%)	1,932	64	9
XII (4 DPA)	1,442 (41.88%)	2,001 (58.12%)	3,443	119	15
XIII (WT)	707 (43%)	941 (57%)	1,648	149	16
XIV (WT 2 DPA)	688 (50.7%)	668 (49.3%)	1,356	70	10
XV (WT 4 DPA)	1,190 (46.5%)	1,370 (53.5%)	2,560	109	15
XVI (*ray*)	534 (42%)	742 (58%)	1,276	38	7
XVII (*ray* 2 DPA)	527 (50.4%)	519 (49.61%)	1,046	18	4
XVIII (*ray* 4 DPA)	1,293 (59%)	901 (41%)	2,194	94	11

* Gene sets and clusters were summarised from Supplementary Table 4 and Figure 7

Table 7. Twenty-seven gene sets of the cell wall biogenesis cluster upregulated in MSC tissue.

GO terms	Description	Number of genes
GO:0005794	CC Golgi apparatus	215
GO:0031982	CC vesicle	148
GO:0097708	CC intracellular vesicle	146
GO:0031410	CC cytoplasmic vesicle	146
GO:0005768	CC endosome	128
GO:0044431	none	97
GO:0098791	CC Golgi apparatus subcompartment	95
GO:0005975	BP carbohydrate metabolic process	90
GO:0005802	CC trans-Golgi network	85
GO:0044262	BP cellular carbohydrate metabolic process	66
GO:0016192	BP vesicle-mediated transport	56
GO:0044264	BP cellular polysaccharide metabolic process	50
GO:0071669	BP plant-type cell wall organization or biogenesis	42
GO:0006073	BP cellular glucan metabolic process	35
GO:0044042	BP glucan metabolic process	35
GO:0042546	BP cell wall biogenesis	34
GO:0000271	BP polysaccharide biosynthetic process	33
GO:0009832	BP plant-type cell wall biogenesis	30
GO:0098657	BP import into cell	20
GO:0006897	BP endocytosis	16
GO:0030244	BP cellulose biosynthetic process	12
GO:0000139	CC Golgi membrane	12
GO:0030243	BP cellulose metabolic process	12
GO:0009834	BP plant-type secondary cell wall biogenesis	12
GO:0051274	BP beta-glucan biosynthetic process	12
GO:0051273	BP beta-glucan metabolic process	12
GO:0006471	BP protein ADP-ribosylation	4

CC = Cellular Compartment; BP = Biological Process

Table 8. A subset of genes in the cell wall biogenesis cluster using GSEA ranking.

Gene id	Protein product
KN361_1G006655	serine/threonine-protein kinase-like protein ACR4
KN361_1G003551	O-fucosyltransferase 35 OFUT35
KN361_4G004070	putative pectin methyltransferase QUA2
KN361_4G006180	fatty acid amide hydrolase
KN361_2G000873	V-type proton ATPase subunit F
KN361_1G006340	monosaccharide-sensing protein 2
KN361_1G006406	ADP-ribosylation factor 2
KN361_1G006329	CMP-sialic acid transporter 2;CMP-sialic acid transporter 4
KN361_2G000799	BTB/POZ domain-containing protein NPY1
KN361_1G010850	NADH-cytochrome b5 reductase-like protein
KN361_3G005174	transcriptional corepressor LEUNIG
KN361_3G001738	putative galacturonosyltransferase 7;hypothetical protein
KN361_2G001203	putative methyltransferase PMT26
KN361_4G004298	UDP-rhamnose/UDP-galactose transporter 2 (URGT2)
KN361_4G004196	UDP-rhamnose/UDP-galactose transporter 5;UDP-rhamnose/UDP-galactose transporter 6
KN361_3G002964	protein reduced wall acetylation RWA1
KN361_2G003655	Ras-related protein RABH1b
KN361_3G008685	heat shock cognate protein 70
KN361_1G005601	phosphoglycerate kinase;phosphoglycerate kinase 3
KN361_4G005195	1-phosphatidylinositol-3-phosphate 5-kinase FAB1B
KN361_3G005173	GTP-binding nuclear protein Ran-3
KN361_3G005346	NAC domain-containing protein 43
KN361_3G007884	trafficking protein particle complex 2-specific subunit 130
KN361_3G003488	phosphoglucomutase
KN361_0G000351	transmembrane 9 superfamily protein member 3
KN361_1G003655	Ras-related protein RABD2c;Ras-related protein RABD2b
KN361_1G002840	protein candidate G-protein coupled receptor 7
KN361_1G011531	putative UDP-arabinopyranose mutase 5
KN361_2G004308	putative methyltransferase PMT7;putative methyltransferase PMT6
KN361_2G006538	V-type proton ATPase subunit C
KN361_3G003818	putative UDP-arabinopyranose mutase 2
KN361_4G004839	ADP-ribosylation factor 2
KN361_2G001542	protein-tyrosine sulfotransferase
KN361_1G003839	chloride channel protein CLC-f
KN361_4G004572	protein embryo sac development arrest EDA30
KN361_1G004055	two pore calcium channel protein 1A
KN361_3G000481	COBRA-like protein 4
KN361_2G007972	beta-amylase 1
KN361_2G001810	golgi SNAP receptor complex member 1-2
KN361_4G006009	galacturonokinase
KN361_1G003786	glucose-1-phosphate adenylyltransferase small subunit

KN361_2G000408	cellulose synthase A catalytic subunit 1 UDP-forming
KN361_2G000158	ADP-ribosylation factor GTPase-activating protein AGD12
KN361_1G003529	protein trichome birefringence-like 16
KN361_1G001941	cellulose synthase A catalytic subunit 4 UDP-forming
KN361_4G005337	protein GRIP
KN361_1G003564	UDP-glucuronic acid decarboxylase 6;UDP-glucuronic acid decarboxylase 5
KN361_4G003908	calcium-transporting ATPase 3
KN361_1G011367	vacuolar protein-sorting-associated protein 33
KN361_3G003100	putative glucuronoxylan glucuronosyltransferase IRX7
KN361_3G002565	pyruvate kinase isozyme A
KN361_3G001678	katanin p60 ATPase-containing subunit A1
KN361_2G000069	(24)-sterol reductase
KN361_1G001105	heat shock protein 70-17
KN361_2G001529	receptor-like serine/threonine-protein kinase NCRK
KN361_3G003744	sucrose transport protein SUC3
KN361_3G002325	putative plastidic glucose transporter 2
KN361_3G003194	6-galactosyltransferase GALT29A
KN361_2G001422	4-alpha-glucanotransferase
KN361_3G004354	phosphoglucan phosphatase LSF1
KN361_0G000558	cellulose synthase A catalytic subunit 1 UDP-forming
KN361_4G003658	magnesium/proton exchanger 2;magnesium/proton exchanger
KN361_3G003505	vacuolar-sorting protein BRO1
KN361_2G000814	V-type proton ATPase subunit a3
KN361_4G001055	7-dehydrocholesterol reductase
KN361_2G000201	3-galactosyltransferase 20
KN361_1G010072	3-galactosyltransferase 2
KN361_2G000483	V-type proton ATPase subunit D
KN361_2G001746	glycosyltransferase-like KOBITO 1
KN361_3G001622	vacuolar protein sorting-associated protein 51
KN361_2G001024	putative sphingolipid transporter spinster 3
KN361_2G003435	putative pectin methyltransferase QUA3
KN361_4G004012	glucose-6-phosphate isomerase 1
KN361_2G000060	Ras-related protein RABF2a
KN361_2G004086	4-alpha-glucanotransferase DPE2
KN361_4G005145	2-alpha-mannosidase MNS3
KN361_2G001102	protein MODIFIED TRANSPORT TO THE VACUOLE 1
KN361_2G003495	cellulose synthase A catalytic subunit 3 UDP-forming
KN361_3G000089	galactan beta-1,4-galactosyltransferase GALS1
KN361_1G004268	golgi SNAP receptor complex member 1-1
KN361_3G000517	putative galacturonosyltransferase 9
KN361_1G011320	protein PPLZ12;hypersensitive-induced response protein 4
KN361_1G002843	phosphatidate phosphatase PAH1
KN361_2G009473	vacuolar-sorting protein BRO1
KN361_2G001274	4-beta-mannosidase 7
KN361_1G008065	WAT1-related protein

KN361_3G001922	prolyl 4-hydroxylase 1
KN361_1G006407	ADP-ribosylation factor 2
KN361_2G002647	6-galactosyltransferase GALT29A
KN361_2G008834	alpha-glucan water dikinase
KN361_1G004007	putative galacturonosyltransferase 10
KN361_4G001076	kelch repeat-containing protein
KN361_2G008220	mitogen-activated protein kinase MMK1
KN361_3G007102	L-arabinokinase
KN361_3G008804	V-type proton ATPase subunit H
KN361_2G000090	vacuolar protein sorting-associated protein 54
KN361_2G006539	sialyltransferase-like protein 2
KN361_3G001892	chitinase-like protein 1
KN361_4G004676	hypothetical protein
KN361_3G004743	protein sweetIE
KN361_4G007332	tubulin beta-1 subunit
KN361_1G001356	polyprenol reductase 2
KN361_1G001284	triosephosphate isomerase
KN361_2G007106	UDP-glucuronate 4-epimerase 4
KN361_2G003205	peptidyl-prolyl isomerase CYP21-4
KN361_1G010673	peptidyl-prolyl isomerase CYP21-2
KN361_4G004835	protein ECHIDNA

Table 9. Pairwise RNA-seq comparisons between two-time points.

Comparison (2 vs 4 DPA)	Up	Down	Total DEGs	Gene sets	Cluster*
XIX (All)	780	908	1,688	230	18
XX (INT)	643	536	1,179	290	18
XXI (MSC)	1,349	1,750	3,099	172	16
XXII (Wildtype)	175	274	449	179	19
XXIII (Wildtype INT)	213	189	443	80	6
XXIV (Wildtype MSC)	788	955	1,743	101	10
XXV (*ray*)	638	436	1,074	177	16
XXVI (*ray* INT)	590	393	983	170	14
XXVII (*ray* MSC)	1452	982	2,434	117	13

* Gene sets and clusters were summarised from Supplementary Table 5 and Figure 9

Table 10. Thirty gene sets in the cell wall biogenesis cluster upregulated in the 4 DPA samples.

GO terms	Description	Number of genes
GO:0005794	CC Golgi apparatus	215
GO:0005975	BP carbohydrate metabolic process	90
GO:0071554	BP cell wall organization or biogenesis	67
GO:0044262	BP cellular carbohydrate metabolic process	66
GO:0005976	BP polysaccharide metabolic process	60
GO:0044264	BP cellular polysaccharide metabolic process	50
GO:0071669	BP plant-type cell wall organization or biogenesis	42
GO:0016051	BP carbohydrate biosynthetic process	40
GO:0006073	BP cellular glucan metabolic process	35
GO:0044042	BP glucan metabolic process	35
GO:0042546	BP cell wall biogenesis	34
GO:0000271	BP polysaccharide biosynthetic process	33
GO:0034637	BP cellular carbohydrate biosynthetic process	33
GO:0033692	BP cellular polysaccharide biosynthetic process	32
GO:0009832	BP plant-type cell wall biogenesis	30
GO:0016052	BP carbohydrate catabolic process	26
GO:0005982	BP starch metabolic process	20
GO:0044275	BP cellular carbohydrate catabolic process	18
GO:0009250	BP glucan biosynthetic process	17
GO:0009834	BP plant-type secondary cell wall biogenesis	12
GO:0009251	BP glucan catabolic process	12
GO:0030243	BP cellulose metabolic process	12
GO:0044247	BP cellular polysaccharide catabolic process	12
GO:0005983	BP starch catabolic process	12
GO:0051274	BP beta-glucan biosynthetic process	12
GO:0030244	BP cellulose biosynthetic process	12
GO:0051273	BP beta-glucan metabolic process	12
GO:0005996	BP monosaccharide metabolic process	10
GO:0046835	BP carbohydrate phosphorylation	9
GO:0019200	MF carbohydrate kinase activity	8

CC = cellular component; BP = Biological process; MF = molecular function

Table 11. A subset of genes in the cell wall biogenesis cluster using GSEA ranking.

Gene id	Protein product
KN361_4G000651	UDP-arabinopyranose mutase 1
KN361_4G004806	phosphoglycerate kinase 3
KN361_1G007892	3-galactosyltransferase 2
KN361_3G004812	reticulon-like protein B2
KN361_3G004387	xyloglucan glycosyltransferase 4;putative xyloglucan glycosyltransferase 8
KN361_3G000481	COBRA-like protein 4
KN361_3G008877	Ras-related protein RABH1b
KN361_3G001892	chitinase-like protein 1
KN361_2G008220	mitogen-activated protein kinase MMK1
KN361_2G008242	phosphoglucanwater dikinase GWD3
KN361_4G005544	endoglucanase 25
KN361_1G006740	V-type proton ATPase subunit a3
KN361_3G000219	xyloglucan glycosyltransferase 4
KN361_1G010880	L-galactose dehydrogenase
KN361_4G003908	calcium-transporting ATPase 3
KN361_2G000090	vacuolar protein sorting-associated protein 54
KN361_1G003564	UDP-glucuronic acid decarboxylase 6;UDP-glucuronic acid decarboxylase 5
KN361_2G003205	peptidyl-prolyl isomerase CYP21-4
KN361_3G000251	metal tolerance protein 8;metal tolerance protein 12
KN361_1G003529	protein trichome birefringence-like 16
KN361_2G003495	cellulose synthase A catalytic subunit 3 UDP-forming
KN361_2G001422	4-alpha-glucanotransferase
KN361_4G005145	2-alpha-mannosidase MNS3
KN361_2G002115	putative methyltransferase PMT5;putative methyltransferase PMT4
KN361_1G001941	cellulose synthase A catalytic subunit 4 UDP-forming
KN361_1G001105	heat shock protein 70-17
KN361_3G009077	putative protein S-acyltransferase 14
KN361_2G008834	alpha-glucan water dikinase
KN361_3G004396	alpha-mannosidase 2
KN361_3G000517	putative galacturonosyltransferase 9
KN361_2G000419	copper transporter 5
KN361_3G008315	long chain acyl-CoA synthetase 8
KN361_1G004007	putative galacturonosyltransferase 10
KN361_1G011320	protein PPLZ12;hypersensitive-induced response protein 4
KN361_2G001021	dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit
KN361_2G002647	6-galactosyltransferase GALT29A
KN361_4G007028	putative methyltransferase PMT23
KN361_2G000201	3-galactosyltransferase 20
KN361_1G000107	alpha-L-arabinofuranosidase 1
KN361_2G001746	glycosyltransferase-like KOBITO 1
KN361_3G003744	sucrose transport protein SUC3
KN361_3G003194	6-galactosyltransferase GALT29A

KN361_3G000304	V-type proton ATPase subunit d2
KN361_2G001274	4-beta-mannosidase 7
KN361_2G007972	beta-amylase 1
KN361_3G003648	vesicle transport v-SNARE 12
KN361_2G004071	hypersensitive-induced response protein 1
KN361_4G004196	UDP-rhamnose/UDP-galactose transporter 5;UDP-rhamnose/UDP-galactose transporter 6
KN361_3G002325	putative plastidic glucose transporter 2
KN361_1G001698	transmembrane 9 superfamily protein member 1
KN361_1G005769	phosphatidate phosphatase PAH1
KN361_3G003488	phosphoglucomutase
KN361_2G001102	protein MODIFIED TRANSPORT TO THE VACUOLE 1
KN361_4G006009	galacturonokinase
KN361_2G001203	putative methyltransferase PMT26
KN361_1G003655	Ras-related protein RABD2c;Ras-related protein RABD2b
KN361_1G006152	protein trichome birefringence-like 1;protein trichome birefringence
KN361_1G011531	putative UDP-arabinopyranose mutase 5
KN361_4G000358	golgin candidate 2
KN361_1G004983	cellulose synthase A catalytic subunit 8 UDP-forming
KN361_2G003435	putative pectin methyltransferase QUA3
KN361_3G002964	protein reduced wall acetylation RWA1
KN361_3G005346	NAC domain-containing protein 43
KN361_2G004086	4-alpha-glucanotransferase DPE2
KN361_1G002155	hexokinase-2
KN361_3G008644	phosphoglucomutase
KN361_2G000408	cellulose synthase A catalytic subunit 1 UDP-forming
KN361_2G000158	ADP-ribosylation factor GTPase-activating protein AGD12
KN361_3G007102	L-arabinokinase
KN361_2G006538	V-type proton ATPase subunit C
KN361_2G001542	protein-tyrosine sulfotransferase
KN361_2G000814	V-type proton ATPase subunit a3;V-type proton ATPase subunit a2
KN361_2G007106	UDP-glucuronate 4-epimerase 4
KN361_3G008496	phosphomannomutase
KN361_1G003839	chloride channel protein CLC-f
KN361_1G004277	trafficking protein particle complex 2-specific subunit 120
KN361_4G004012	glucose-6-phosphate isomerase 1
KN361_1G010673	peptidyl-prolyl isomerase CYP21-2
KN361_4G004835	protein ECHIDNA
KN361_3G007884	trafficking protein particle complex 2-specific subunit 130
KN361_3G004187	glycosyltransferase family protein 92 protein
KN361_2G000483	V-type proton ATPase subunit D
KN361_4G005458	putative UDP-arabinopyranose mutase 5
KN361_0G000558	cellulose synthase A catalytic subunit 1 UDP-forming
KN361_4G003949	xylulose kinase 2
KN361_4G004676	hypothetical protein
KN361_3G001622	vacuolar protein sorting-associated protein 51

KN361_2G003655	Ras-related protein RABH1b
KN361_4G005200	cation-chloride cotransporter 1
KN361_4G004298	UDP-rhamnose/UDP-galactose transporter 2
KN361_2G000699	BURP domain-containing protein RD22
KN361_2G003441	putative methyltransferase PMT15
KN361_4G001055	7-dehydrocholesterol reductase
KN361_3G001183	neutral ceramidase 1
KN361_1G006345	3-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase
KN361_4G001076	kelch repeat-containing protein
KN361_1G004268	golgi SNAP receptor complex member 1-1
KN361_2G008164	sulfhydryl oxidase 2
KN361_2G000340	3-galactosyltransferase 9
KN361_1G005601	phosphoglycerate kinase;phosphoglycerate kinase 3
KN361_2G009022	galactokinase
KN361_2G004308	putative methyltransferase PMT7;putative methyltransferase PMT6
KN361_3G008804	V-type proton ATPase subunit H
KN361_3G003784	putative methyltransferase PMT4;putative methyltransferase PMT5
KN361_2G000873	V-type proton ATPase subunit F
KN361_4G006526	vesicle transport protein
KN361_1G010072	3-galactosyltransferase 2
KN361_2G007499	protein CSI1
KN361_1G004466	phosphatidylinositol 4-phosphate 5-kinase 9
KN361_4G004558	aldehyde dehydrogenase
KN361_3G000385	pectinesterase 31
KN361_2G006163	O-fucosyltransferase 1
KN361_3G003100	putative glucuronoxylan glucuronosyltransferase IRX7
KN361_2G001792	putative protein S-acyltransferase 14
KN361_4G003971	sucrose transport protein SUC4
KN361_4G004024	isoamylase 3
KN361_3G004354	phosphoglucan phosphatase LSF1
KN361_3G005076	L-arabinokinase
KN361_3G001678	katanin p60 ATPase-containing subunit A1
KN361_2G000080	UDP-arabinose 4-epimerase 1
KN361_4G006119	diacylglycerol O-acyltransferase 1;diacylglycerol O-acyltransferase 1B
KN361_1G001284	triosephosphate isomerase
KN361_1G000338	glutaredoxin-C4
KN361_1G006354	putative galacturonosyltransferase 3
KN361_1G003609	tobamovirus multiplication protein 2A
KN361_4G006886	7-bisphosphatase
KN361_2G001810	golgi SNAP receptor complex member 1-2
KN361_1G006340	monosaccharide-sensing protein 2
KN361_4G006153	AP-4 complex subunit sigma
KN361_1G010597	phosphoglucan phosphatase LSF2
KN361_3G004867	long chain acyl-CoA synthetase 4
KN361_1G007614	callose synthase 12

KN361_4G002411	dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit 2
KN361_3G000594	metal transporter Nramp3
KN361_4G004572	protein embryo sac development arrest EDA30
KN361_1G010850	NADH-cytochrome b5 reductase-like protein
KN361_2G003816	putative anion transporter 5;putative anion transporter 6
KN361_3G004743	protein SWEETIE
KN361_3G005174	transcriptional corepressor LEUNIG
KN361_2G000547	protein reversion-to-ethylene sensitivity RTE11
KN361_4G005337	protein GRIP
KN361_3G000089	galactan beta-1,4-galactosyltransferase GALS1
KN361_3G001724	sialyltransferase-like protein 1
KN361_4G006980	mannan synthesis protein MNN2

Supplementary Table 1. Quality report of total RNAseq.

Sample Name	Conc. (ng/ul)	RIN	Index	Mean Size Avg (bp)	Total Clusters Passing Filter (Million)
WT_INT_2_S1	0.266	7.5	GTTACACC + AGAAGTGC	425	34.4
WT_MSC_2_S2	0.266	6.6	TGAGACAG + CTGGTATC	414	39.9
WT_INT_2_S3	0.266	7.5	GCGTAGAT + AATCGCAC	426	43.6
WT_MSC_2_S4	0.266	7.6	CCTGTATG + GCTCTATG	443	37.4
WT_INT_2_S5	0.266	7.4	GCAGCATA + GTCATCGA	423	41.3
WT_MSC_2_S6	0.266	7.3	CCTTGTAC + CTCCTACA	417	33.4
WT_INT_2_S7	0.266	7.4	GTCAGGAA + TCTTCGGA	425	34.6
WT_MSC_2_S8	0.266	6.9	CAATGTGG + TGTCAACG	411	41.1
MT_INT_2_S9	0.266	6.9	ACCGAGAT + CACCACAA	419	39.1
MT_MSC_2_S10	0.266	7.4	CACAGTCA + TCAACGCT	405	34
MT_INT_2_S11	0.266	7.8	ATACGAGC + GCCAATCA	410	37.2
MT_MSC_2_S12	0.266	8.3	TTCACGCA + ATCTCGCA	404	33.2
MT_INT_2_S13	0.266	7.5	AAGATGCC + ACTCCAAG	419	34.5
MT_MSC_2_S14	0.266	7.1	AACCACGT + ACTTCCAC	419	31.6
MT_INT_2_S15	0.266	8.1	ACAGAACG + CTCCTAAG	427	34.9
MT_MSC_2_S16	0.266	8.2	ACGCACAA + CACTGCAA	426	37.7
WT_INT_4_S17	0.266	7.2	CCTTCACA + GTAGAGCA	433	20.8
WT_MSC_4_S18	0.266	7.7	ATGTCGTG + AGGATGTG	407	24.2
WT_INT_4_S19	0.266	8.1	GCAGACTT + TCCTCTGA	440	36.9
WT_MSC_4_S20	0.266	7.4	GAACAGAC + GTCTACCT	417	24.9
WT_INT_4_S21	0.266	7.5	GTCAAGTC + AAGAGGAG	427	38.7
WT_MSC_4_S22	0.266	7.8	TGTTAGCG + AGACGGTT	425	37.8
WT_INT_4_S23	0.266	8.3	AGCAGCAA + GTAGAAGC	417	35.4
WT_MSC_4_S24	0.266	7.7	GAAGGATC + AGGCTCTT	402	37.2
MT_INT_4_S25	0.266	7.2	ATCCTGTC + ACGAGATG	416	43.2
MT_MSC_4_S26	0.266	7.5	GCACTACA + GAAGCACA	404	42.9
MT_INT_4_S27	0.266	7.7	AACAGTCC + CTTCTCAG	428	40.1
MT_MSC_4_S28	0.266	7.5	ACGTGACT + ACTCGTTC	454	36.2
MT_INT_4_S29	0.266	7.5	ACTCGACA + GCACAAGA	430	40.1
MT_MSC_4_S30	0.266	7.6	CACCGATA + GAGTAACC	410	33.2
MT_INT_4_S31	0.266	7.4	GAACAAGC + ACGACCAT	414	35.3
MT_MSC_4_S32	0.266	7.2	AGGAAGCT + CAATCACC	399	40.9
WT_INT_6_S33	0.266	8.1	CTGTACTC + CTAGATGC	415	43.7
WT_INT_6_S34	0.266	8.3	TACCGCTA + CGTAAGGT	421	41.4
WT_INT_6_S35	0.266	6.8	TGTATGGC + CTACCACT	424	38.7
WT_INT_6_S36	0.266	7.8	GAGCCTAT + CTCTGTAC	420	42.1
MT_INT_6_S37	0.266	7.4	TCACTCTG + CTGTTACG	413	41.3
MT_INT_6_S38	0.266	7.4	ACGCGTTA + CCACCATA	406	42.8
MT_INT_6_S39	0.266	7.1	ACATTCGG + TCACTGTC	413	38.2
MT_INT_6_S40	0.266	7	TAGCACGT + AAGAGTCG	413	42.1

Supplementary Table 2. A list of identified transcription factors in WT (A), WT and *ray* (B), WT only (C), *ray* (D), *ray* only (E), not in WT and *ray* (F). Number 1 or blank represents true or false for those conditions.

Transcription factor	Gene id	A	B	C	D	E	F
K14431 TGA; transcription factor TGA	KN361_1g000018				1	1	
	KN361_1g006214						1
	KN361_1g010194	1	1		1		
	KN361_2g002631						1
	KN361_2g007478	1	1		1		
	KN361_3g001069						1
	KN361_3g003198				1	1	
	KN361_4g007574	1	1		1		
K09060 GBF; plant G-box-binding factor	KN361_1g003364	1	1		1		
	KN361_2g003185	1	1		1		
	KN361_2g003334	1	1		1		
	KN361_2g006101						1
	KN361_3g000516						1
	KN361_3g003977	1	1		1		
	KN361_3g004663	1	1		1		
	KN361_3g004898	1	1		1		
K16241 HY5; transcription factor HY5	KN361_2g000411	1	1		1		
K14432 ABF; ABA responsive element binding factor	KN361_1g000226	1	1		1		
	KN361_1g004953						1
	KN361_1g011181	1	1		1		
	KN361_2g003794						1
	KN361_2g004039	1	1		1		
	KN361_3g004093	1	1		1		
	KN361_4g003738	1	1		1		
	KN361_4g006346	1	1		1		
K20557 VIP1; transcription factor VIP1	KN361_1g001355	1	1		1		
	KN361_1g002145	1	1		1		
	KN361_1g006673	1	1		1		
	KN361_2g006191	1	1		1		
	KN361_2g009403	1	1		1		
	KN361_3g001866	1	1		1		
	KN361_4g003717	1	1		1		
	KN361_4g007106	1	1		1		
K21626 CCNDBP1, DIP1, GCIP; cyclin-D1-binding protein 1	KN361_3g001965	1	1		1		
K12126 PIF3; phytochrome-interacting factor 3	KN361_1g001540	1	1		1		

K16189 PIF4; phytochrome-interacting factor 4	KN361_3g008186						1
K13422 MYC2; transcription factor MYC2	KN361_1g001846	1	1		1		
	KN361_1g002590	1	1		1		
	KN361_1g009135	1	1		1		
	KN361_2g003280	1	1		1		
	KN361_2g004620	1	1		1		
K10779 ATRX; transcriptional regulator ATRX	KN361_1g011503	1	1		1		
K07466 RFA1, RPA1, rpa; replication factor A1	KN361_1g009580	1	1		1		
	KN361_3g004947	1	1		1		
	KN361_4g000232	1	1		1		
K11308 MYST1, MOF, KAT8; histone acetyltransferase MYST1	KN361_4g004497	1	1		1		
K15263 LYER; cell growth-regulating nucleolar protein	KN361_1g003596	1	1		1		
K10643 CNOT4, NOT4, MOT2; CCR4-NOT transcription complex	KN361_1g011477	1	1		1		
	KN361_2g002853	1	1		1		
K00558 DNMT1, dcm; DNA (cytosine-5)-methyltransferase 1	KN361_2g006513	1	1		1		
	KN361_2g007437	1	1		1		
	KN361_2g009172	1	1		1		
	KN361_3g004923	1	1		1		
	KN361_4g003776	1	1		1		
K12236 NFX1; transcriptional repressor NF-X1	KN361_2g006545	1	1		1		
K09250 CNBP; cellular nucleic acid-binding protein	KN361_1g006732	1	1		1		
	KN361_1g011127	1	1		1		
	KN361_2g002531	1	1		1		
K09377 CSRP; cysteine and glycine-rich protein	KN361_1g002751				1		1
	KN361_1g010956						1
	KN361_2g003432						1
	KN361_3g008487						1
	KN361_4g003903						1
	KN361_4g006883	1	1				
K09313 CUTL; homeobox protein cut-like	KN361_2g009143	1	1		1		
K09338 HD-ZIP; homeobox-leucine zipper protein	KN361_1g001521						1
	KN361_1g002634	1	1		1		
	KN361_1g002636	1	1		1		
	KN361_1g003678	1	1		1		
	KN361_1g004416						1
	KN361_1g004633	1	1		1		
	KN361_1g005495						1
	KN361_1g006153						1

	KN361_1g010076	1	1		1		
	KN361_2g000160	1		1			
	KN361_2g000953						1
	KN361_2g001026						1
	KN361_2g003169						1
	KN361_2g003552	1	1		1		
	KN361_2g006387	1	1		1		
	KN361_2g007578	1	1		1		
	KN361_2g007713	1	1		1		
	KN361_3g000203						1
	KN361_3g000318	1	1		1		
	KN361_3g002323				1	1	
	KN361_3g002364	1	1		1		
	KN361_3g003745	1	1		1		
	KN361_3g003788	1	1		1		
	KN361_3g004447	1	1		1		
	KN361_3g004666						1
	KN361_3g006211	1	1		1		
	KN361_3g007853						1
	KN361_3g007867						1
	KN361_3g008622	1	1		1		
	KN361_3g008758	1	1		1		
	KN361_4g000102						1
	KN361_4g000717						1
	KN361_4g000936	1	1		1		
	KN361_4g005005	1	1		1		
	KN361_4g005596	1	1		1		
K06620 E2F3; transcription factor E2F3	KN361_3g003253	1	1		1		
K09391 E2F7_8; transcription factor E2F7/8	KN361_3g000256	1	1		1		
K04683 TFDP1; transcription factor Dp-1	KN361_2g004010	1	1		1		
K09419 HSFF; heat shock transcription factor	KN361_1g002553	1	1		1		
	KN361_1g002650	1	1		1		
	KN361_1g002739	1	1		1		
	KN361_1g004967	1	1		1		
	KN361_1g006019						1
	KN361_1g010242	1	1		1		
	KN361_2g000521	1	1		1		
	KN361_3g000676						1
	KN361_3g004083						1
	KN361_4g007310	1	1		1		
K09422 MYBP; transcription factor MYB, plant	KN361_1g000238				1	1	
	KN361_1g001020						1
	KN361_1g001412	1	1		1		
	KN361_1g001543	1	1		1		
	KN361_1g001861	1	1		1		
	KN361_1g002429	1	1		1		
	KN361_1g002610						1

KN361_1g002801	1	1		1		
KN361_1g003087				1	1	
KN361_1g003736						1
KN361_1g003982						1
KN361_1g004499				1	1	
KN361_1g005234						1
KN361_1g005609						1
KN361_1g006640	1	1		1		
KN361_1g009349						1
KN361_1g009451						1
KN361_1g011558	1	1		1		
KN361_2g000579						1
KN361_2g000688	1	1		1		
KN361_2g001017	1	1		1		
KN361_2g001124	1		1			
KN361_2g001663	1	1		1		
KN361_2g002144				1	1	
KN361_2g002578	1	1		1		
KN361_2g002763						1
KN361_2g002901						1
KN361_2g003165						1
KN361_2g003425	1	1		1		
KN361_2g003979	1		1			
KN361_2g005689				1	1	
KN361_2g006374						1
KN361_2g007519						1
KN361_2g007667						1
KN361_2g007746						1
KN361_2g008351	1	1		1		
KN361_2g008587	1	1		1		
KN361_2g008640	1	1		1		
KN361_2g009006						1
KN361_2g009007						1
KN361_2g009047						1
KN361_2g009048						1
KN361_2g009187	1	1		1		
KN361_2g009440						1
KN361_3g000158						1
KN361_3g000341	1	1		1		
KN361_3g000381						1
KN361_3g000845	1	1		1		
KN361_3g001171						1
KN361_3g002716						1
KN361_3g002982						1
KN361_3g003747	1	1		1		
KN361_3g004141	1	1		1		
KN361_3g004351						1
KN361_3g004487	1	1		1		
KN361_3g004788						1
KN361_3g005464	1	1		1		

	KN361_3g005729						1
	KN361_3g007382						1
	KN361_4g000339	1	1		1		
	KN361_4g000850	1	1		1		
	KN361_4g002816	1	1		1		
	KN361_4g003935						1
	KN361_4g004697						1
	KN361_4g004901						1
	KN361_4g005304						1
	KN361_4g005445	1	1		1		
	KN361_4g005583						1
	KN361_4g006683	1	1		1		
	KN361_4g006983	1	1		1		
	KN361_4g007324	1	1		1		
	KN361_0g000123	1	1		1		
K12133 LHY; MYB-related transcription factor LHY	KN361_1g005716	1	1		1		
K09264 K09264; MADS-box transcription factor, plant	KN361_1g002122	1	1		1		
	KN361_1g004174	1	1		1		
	KN361_1g004313	1	1		1		
	KN361_1g006022	1	1		1		
	KN361_2g000770						1
	KN361_2g000832						1
	KN361_2g002819	1	1		1		
	KN361_2g003489						1
	KN361_2g006301	1	1		1		
	KN361_2g007624						1
	KN361_2g008174	1	1		1		
	KN361_2g008397						1
	KN361_3g000487	1	1		1		
	KN361_3g000489						1
	KN361_3g002179				1	1	
	KN361_3g003825						1
	KN361_3g005375						1
	KN361_3g008636						1
	KN361_3g008637	1	1		1		
	KN361_4g000637	1	1		1		
	KN361_4g004353	1	1		1		
	KN361_4g005875	1	1		1		
	KN361_4g006229						1
K03120 TBP, tbp; transcription initiation factor TFIID	KN361_1g003332						1
	KN361_1g010352	1	1		1		
	KN361_3g000185						1
	KN361_3g001386	1	1		1		
K09272 SSRP1, POB3; FACT complex subunit SSRP1/POB3	KN361_3g000298	1	1		1		
K08064 NFYA, HAP2; nuclear transcription factor Y, alpha	KN361_1g003582				1	1	
	KN361_2g000469	1	1		1		

	KN361_2g003772	1	1		1		
	KN361_3g007263						1
	KN361_3g007396	1	1		1		
	KN361_4g005206	1	1		1		
K08065 NFYB, HAP3; nuclear transcription Y subunit beta	KN361_1g005358						1
	KN361_1g005497						1
	KN361_1g005753	1	1		1		
	KN361_1g006560						1
	KN361_1g007636	1	1		1		
	KN361_4g004627						1
	KN361_4g006830	1	1		1		
K08066 NFYC, HAP5; nuclear transcription factor Y, gamma	KN361_2g007769	1	1		1		
	KN361_3g002633	1	1		1		
	KN361_4g000628	1	1		1		
	KN361_4g004810	1	1		1		
K04681 RBL1; retinoblastoma-like protein 1	KN361_1g004743	1	1		1		
K04498 EP300, CREBBP, KAT3; E1A/CREB-binding protein	KN361_1g001095	1	1		1		
	KN361_3g008133	1	1		1		
K16221 TCP21, CHE; transcription factor TCP21	KN361_1g001680	1	1		1		
	KN361_1g004541						1
	KN361_2g006285						1
	KN361_2g006459	1	1		1		
	KN361_2g008341	1	1		1		
	KN361_3g001032						1
	KN361_3g005324						1
	KN361_3g008532	1	1		1		
	KN361_4g007276	1	1		1		
K09284 AP2; AP2-like factor, euAP2 lineage	KN361_1g002836						1
	KN361_3g000640	1	1		1		
K09285 OVM, ANT; AP2-like factor, ANT lineage	KN361_2g006670	1	1		1		
K09286 EREBP; EREBP-like factor	KN361_1g000253						1
	KN361_1g001868						1
	KN361_1g003415						1
	KN361_1g003513						1
	KN361_1g004259						1
	KN361_1g004798	1	1		1		
	KN361_1g004933						1
	KN361_1g006872						1
	KN361_1g007592				1	1	
	KN361_1g007650	1	1		1		
	KN361_1g007847	1	1		1		
	KN361_2g000651	1	1		1		
	KN361_2g000817						1
	KN361_2g001372						1
	KN361_2g001843						1
	KN361_2g002179						1
	KN361_2g002269	1	1		1		

	KN361_2g005992	1	1		1		
	KN361_2g007336						1
	KN361_2g007516	1	1		1		
	KN361_2g009066	1	1		1		
	KN361_2g009361						1
	KN361_3g000388						
	KN361_3g000638	1	1		1		
	KN361_3g001087	1	1		1		
	KN361_3g001602						1
	KN361_3g004062	1	1		1		
	KN361_3g004863						1
	KN361_3g005203	1	1		1		
	KN361_3g005353						1
	KN361_3g005487	1	1		1		
	KN361_3g007469						1
	KN361_3g007982						1
	KN361_3g008750	1	1		1		
	KN361_3g008790						1
	KN361_3g008791						1
	KN361_4g000438						1
	KN361_4g001149	1	1		1		1
	KN361_4g001272						1
	KN361_4g002228						1
	KN361_4g002229				1	1	
	KN361_4g002231	1	1		1		
	KN361_4g002273						1
	KN361_4g004537	1	1		1		
	KN361_4g004636						1
	KN361_4g004986				1	1	
	KN361_4g004991						1
	KN361_4g007045	1	1		1		
	KN361_4g007177						1
	KN361_4g007178						1
	KN361_4g007284						1
K14516 ERF1; ethylene-responsive transcription factor 1	KN361_2g007705						1
	KN361_2g007706						1
	KN361_2g009196				1	1	
	KN361_3g001346						1
K13434 PTI6; pathogenesis-related genes transcriptional	KN361_1g004742	1	1		1		
	KN361_3g004835	1	1		1		
	KN361_4g003596	1	1		1		
	KN361_4g007445	1	1		1		
K09287 RAV; RAV-like factor	KN361_1g002697	1	1		1		
	KN361_3g003912						1
	KN361_4g006345						1
K18834 WRKY1; WRKY transcription factor 1	KN361_3g008251	1	1		1		
K18835 WRKY2; WRKY transcription factor 2	KN361_4g004728	1	1		1		

K13425 WRKY22; WRKY transcription factor 22	KN361_2g003370	1	1		1		
K13424 WRKY33; WRKY transcription factor 33	KN361_2g003123	1	1		1		
K21994 LBD18; LOB domain-containing protein 18	KN361_3g009098	1	1		1		
K21596 CAMTA; calmodulin-binding transcription activator	KN361_1g001783	1	1		1		
	KN361_2g007524	1	1		1		
	KN361_2g009052	1	1		1		
	KN361_4g000737	1	1		1		
K14514 EIN3; ethylene-insensitive protein 3	KN361_2g001652	1	1		1		
	KN361_2g001654	1	1		1		
	KN361_2g002555	1	1		1		
	KN361_3g001206	1	1		1		
	KN361_3g003290	1	1		1		
	KN361_3g004821	1	1		1		
K05527 bolA; BolA family transcriptional regulator	KN361_2g007694	1	1		1		
K07735 algH; putative transcriptional regulator	KN361_2g008408	1	1		1		
	KN361_3g000410						1
	KN361_4g004330	1	1		1		

Supplementary Table 3. Presence and absence of six clusters between WT and *ray*.

Cluster name	Gene sets	I	II	III	IV	V	VI	VII	VIII	IX
Ion binding	8	0	1	0	0	0	0	0	1	0
Complex mitochondria	5	0	1	0	0	0	0	0	1	0
Electron transport	4	0	0	1	0	0	0	1	0	1
Structural ribosome	4	0	1	0	0	1	0	0	1	0
Glucan catabolic process	2	0	0	0	0	0	0	0	1	0
Cell wall metabolic process	2	0	0	0	0	1	0	0	0	0

0 = absence; 1= presence

Supplementary Table 4. Presence and absence of twenty-five clusters between samples from INT and MSC tissues.

Cluster gene set name	Gene sets	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	Total
Gene silencing	49	1	0	1	1	0	1	0	0	1	5
Ion binding	37	1	1	1	1	1	1	1	1	1	9
Histone modification	36	1	1	1	1	1	1	0	0	1	7
Transcription regulation	30	1	1	1	1	1	1	0	1	1	8
Cell wall biogenesis	27	1	0	1	1	0	1	1	0	1	6
Transmembrane transport	26	1	1	1	1	1	0	1	0	1	7
Nucleoside biosynthesis	24	0	0	1	0	0	1	0	0	0	2
Chloroplast localization	9	1	0	1	1	1	1	0	0	0	5
Phosphodiester bond hydrolysis	7	1	0	1	1	0	1	0	0	0	4
Hydrolase activity	6	1	1	1	1	1	0	0	0	1	6
Ligase activity	6	1	1	0	1	0	1	1	1	0	6
Structural ribosome	6	1	1	0	1	1	0	1	1	0	6
Cytoskeleton	5	0	1	0	0	1	0	0	0	0	2
Embryonic morphogenesis	5	1	0	0	1	0	1	0	0	0	3
Chromosome	4	1	0	1	1	0	1	0	0	0	4
Organic transmembrane carboxylic	4	0	1	0	0	1	0	0	0	0	2
Outer membrane	4	0	0	1	0	0	1	0	0	0	2
Regulatory region sequence	4	1	0	1	1	1	1	0	0	1	6
Endoplasmic reticulum lumen	3	1	0	1	1	0	1	0	0	1	5
Reproductive regulation flower	3	1	0	1	1	0	1	0	0	1	5
Cellular component disassembly	2	1	0	0	0	0	0	0	0	0	1
Protein kinase binding	2	1	0	0	0	0	0	0	0	0	1
Somatodendritic compartment body	2	0	0	0	0	0	0	1	0	0	1
Trichome morphogenesis epidermis	2	0	0	0	0	0	0	1	0	0	1
Ubiquitin protein	2	0	0	0	0	0	0	0	0	1	1

0 = absence; 1= presence

Supplementary Table 5. Presence and absence of twenty-eight clusters between samples from 2 and 4 DPA.

Cluster name	Gene set	XIX	XX	XXI	XXII	XXIII	XXIV	XXV	XXVI	XXVII	Total
Metabolic processes	91	1	1	1	1	1	1	1	1	1	9
Cytokinesis	85	1	1	1	1	1	1	1	1	1	9
Gene silencing	74	1	1	1	1	1	1	1	1	1	9
Nucleoside biosynthesis	62	1	1	1	1	0	1	1	1	1	8
Ion transport	44	1	1	1	1	0	1	1	1	1	8
Intralumenal vesicle transport	39	1	1	1	1	0	1	1	1	1	8
Cell wall biogenesis	30	1	1	1	1	0	1	1	1	1	8
Complex mitochondrial	26	1	1	1	1	0	1	1	1	1	8
Response acid chemical	12	1	1	0	1	1	0	1	1	0	6
Myosin binding	10	1	0	1	0	0	1	1	0	1	5
Endopeptidase inhibitor peptidase	8	0	1	0	0	0	0	0	0	0	1
Elongation seed trichome	5	1	0	1	1	0	1	1	0	0	5
Lignin biosynthesis	5	1	1	0	1	0	0	1	0	0	4
Protein dephosphorylation	5	1	1	1	1	0	0	1	1	1	7
Component plasma anchored	4	0	1	1	0	0	0	0	1	0	3
Nuclear transport	4	0	0	0	0	0	0	0	0	1	1
Hyperosmotic response	4	1	1	1	1	0	0	1	1	0	6
Lipid phosphorylation	4	0	1	0	0	0	0	0	1	0	2
Hydrolase activity	4	1	1	1	1	0	0	0	0	0	4
Embryo sac development	3	1	1	1	1	1	0	1	0	0	6
Lipase activity	2	0	0	0	1	0	0	0	0	0	1
Multi organism localization	2	0	0	0	1	0	0	0	0	0	1
Negative olefin process	2	0	0	0	1	0	0	0	0	0	1
Lipid metabolic process	2	0	0	0	1	0	0	0	0	0	1
Nucleoplasm	2	1	0	1	0	0	0	0	0	0	2
Regulatory sequence specific	2	0	0	0	0	0	0	0	0	1	1
Transducer signaling receptor	2	0	0	0	0	0	0	0	0	1	1
Transaminase activity	2	1	1	0	0	1	0	1	1	0	5

0 = absence; 1= presence

References

Abu-Jamous B, Kelly SA-O (2018) Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data.

Anderson E, Fireman M (1935) The mucilage from psyllium seed, *Plantago* psyllium, L. Journal of Biological Chemistry **109:** 437-442

Andrews S (2017) FastQC: a quality control tool for high throughput sequence data. 2010.

Arsovski AA, Haughn GW, Western TL (2010) Seed coat mucilage cells of *Arabidopsis thaliana* as a model for plant cell wall research. Plant Signaling & Behavior **5:** 796-801

Bansal R (2018) Cell wall invertase and sucrose synthase regulate sugar metabolism during seed development in Isabgol (*Plantago ovata* Forsk.). Proceedings of the National Academy of Sciences, India Section B: Biological Sciences **88:** 73-78

Beers EP, Woffenden BJ, Zhao C (2000) Plant proteolytic enzymes: possible roles during programmed cell death. Plant molecular biology **44:** 399-415

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30:** 2114-2120

Bushnell B (2014) BBMap: a fast, accurate, splice-aware aligner. *In*. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States)

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J (2021) eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Molecular Biology and Evolution **38:** 5825-5829

Cowley JM, Burton RA (2021) The goo-d stuff: *Plantago* as a myxospermous model with modern utility. New Phytol **229:** 1917-1923

Cowley JM, Herliana L, Neumann KA, Ciani S, Cerne V, Burton RA (2020) A small-scale fractionation pipeline for rapid analysis of seed mucilage characteristics. Plant Methods **16:** 1-12

Cowley JM, O'Donovan LA, Burton RA (2021) The composition of Australian Plantago seeds highlights their potential as nutritionally-rich functional food ingredients. Sci Rep **11:** 12692

Debeaujon I, Nesi N, Perez P, Devic M, Grandjean O, Caboche M, Lepiniec L (2003) Proanthocyanidin-accumulating cells in Arabidopsis testa: regulation of differentiation and role in seed development. Plant Cell **15:** 2514-2531

Dhar M, Kaul S, Sareen S, Koul A (2005) *Plantago ovata*: Genetic diversity, cultivation, utilization and chemistry. Plant Genetic Resources: characterization and utilization **3:** 252-263

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29:** 15-21

Du J, Kirui A, Huang S, Wang L, Barnes WJ, Kiemle SN, Zheng Y, Rui Y, Ruan M, Qi S, Kim SH, Wang T, Cosgrove DJ, Anderson CT, Xiao C (2020) Mutations in the Pectin Methyltransferase QUASIMODO2 Influence Cellulose Biosynthesis and Wall Integrity in Arabidopsis. The Plant Cell **32:** 3576-3597

Du J, Ruan M, Li X, Lan Q, Zhang Q, Hao S, Gou X, Anderson CT, Xiao C (2022) Pectin methyltransferase QUASIMODO2 functions in the formation of seed coat mucilage in Arabidopsis. Journal of Plant Physiology: 153709.

Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics **32:** 3047-3048

Fischer MH, Yu N, Gray GR, Ralph J, Anderson L, Marlett JA (2004) The gel-forming polysaccharide of psyllium husk (*Plantago ovata* Forsk). Carbohydrate Research **339:** 2009-2017

Franco EAN, Sanches-Silva A, Ribeiro-Santos R, de Melo NR (2020) Psyllium (*Plantago ovata* Forsk): From evidence of health benefits to its food application. Trends in Food Science & Technology **96:** 166-175

Ghaderi-Far F, Alimagham S, Kameli A, Jamali M (2012) Isabgol (*Plantago ovata* Forsk) seed germination and emergence as affected by environmental factors and planting depth. Journal of Agricultural Science and Technology **6:** 185-194

Golz JF, Allen PJ, Li SF, Parish RW, Jayawardana NU, Bacic A, Doblin MS (2018) Layers of regulation - Insights into the role of transcription factors controlling mucilage production in the Arabidopsis seed coat. Plant Sci **272:** 179-192

Gonzalez A, Mendenhall J, Huo Y, Lloyd A (2009) TTG1 complex MYBs, MYB5 and TT2, control outer seed coat differentiation. Dev Biol **325:** 412-421

Guo Q, Cui SW, Wang Q, Christopher Young J (2008) Fractionation and physicochemical characterization of psyllium gum. Carbohydrate Polymers **73:** 35-43

Gupta M, Kaul S, Dhar MK (2018) Identification and characterization of some putative genes involved in arabinoxylan biosynthesis in *Plantago ovata*. 3 Biotech **8:** 266

Haughn G, Chaudhury A (2005) Genetic analysis of seed coat development in *Arabidopsis*. Trends in Plant Science **10:** 472-477

Hierl G, Vothknecht U, Gietl C (2012) Programmed cell death in Ricinus and Arabidopsis: the function of KDEL cysteine peptidases in development. Physiol Plant **145:** 103-113

Hyde BB (1970) Mucilage-producing cells in the seed coat of *Plantago ovata*: developmental fine structure. American Journal of Botany**:** 1197-1206

Jensen JK, Johnson N, Wilkerson CG (2013) Discovery of diversity in xylan biosynthetic genes by transcriptional profiling of a heteroxylan containing mucilaginous tissue. Frontiers in plant science **4:** 183-183

Johnson CS, Kolevski B, Smyth DR (2002) TRANSPARENT TESTA GLABRA2 a Trichome and Seed Coat Development Gene of Arabidopsis, Encodes a WRKY Transcription Factor. The Plant Cell **14:** 1359

Karimzadeh G, Omidbaigi R (2004) Growth and seed characteristics of isabgol (*Plantago ovata* Forsk) as influenced by some environmental factors. Journal of Agricultural Science and Technology **6:** 103-110

Kassambara A, Mundt F (2017) Factoextra R Package: Easy Multivariate Data Analyses and Elegant Visualization. *In*,

Kohorn BD, Dexter-Meldrum J, Zorensky FDH, Chabout S, Mouille G, Kohorn S (2021) Pectin Dependent Cell Adhesion Restored by a Mutant Microtubule Organizing Membrane Protein. Plants **10**

Kotwal S, Kaul S, Sharma P, Gupta M, Shankar R, Jain M, Dhar MK (2016) De novo transcriptome analysis of medicinally important *Plantago ovata* using RNA-Seq. PloS one **11:** e0150273

Kumar J (2015) Good agricultural practices for Isabgol. ICAR – Directorate of Medicinal and Aromatic Plants Research, Gujarat, India

Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30:** 923-930

Lima NB, Trindade FG, da Cunha M, Oliveira AE, Topping J, Lindsey K, Fernandes KV (2015) Programmed cell death during development of cowpea (*Vigna unguiculata* (L.) Walp.) seed coat. Plant Cell Environ **38:** 718-728

Lopez-Fernandez MP, Maldonado S (2015) Programmed cell death in seeds of angiosperms. J Integr Plant Biol **57:** 996-1002

McFarlane HE, Watanabe Y Fau - Gendre D, Gendre D Fau - Carruthers K, Carruthers K Fau - Levesque-Tremblay G, Levesque-Tremblay G Fau - Haughn GW, Haughn Gw Fau - Bhalerao RP, Bhalerao Rp Fau - Samuels L, Samuels L (2013) Cell wall polysaccharides are mislocalized to the Vacuole in echidna mutants. Plant & Cell Physiology **54:** 1867-1880

Meents MJ, Motani S, Mansfield SD, Samuels AL (2019) Organization of Xylan Production in the Golgi During Secondary Cell Wall Biosynthesis1[OPEN]. Plant Physiology **181:** 527 - 546

Miart F, Fontaine J-X, Mongelard G, Wattier C, Lequart M, Bouton S, Molinié R, Dubrulle N, Fournet F, Demailly H (2021) Integument-Specific Transcriptional Regulation in the Mid-Stage of Flax Seed Development Influences the Release of Mucilage and the Seed Oil Content. Cells **10:** 2677

Naran R, Chen G, Carpita NC (2008) Novel rhamnogalacturonan I and arabinoxylan polysaccharides of flax seed mucilage. Plant Physiology **148:** 132-141

North HM, Berger A, Saez-Aguayo S, Ralet M-C (2014) Understanding polysaccharide production and properties using seed coat mutants: future perspectives for the exploitation of natural variants. Annals of Botany **114:** 1251-1263

Penfield S, Meissner RC, Shoue DA, Carpita NC, Bevan MW (2001) MYB61 Is Required for Mucilage Deposition and Extrusion in the Arabidopsis Seed Coat. The Plant Cell **13:** 2777-2791

Phan JL, Cowley JM, Neumann KA, Herliana L, O'Donovan LA, Burton RA (2020) The novel features of *Plantago ovata* seed mucilage accumulation, storage and release. Sci Rep **10:** 1-14

Phan JL, Tucker MR, Khor SF, Shirley N, Lahnstein J, Beahan C, Bacic A, Burton RA (2016) Differences in glycosyltransferase family 61 accompany variation in seed coat mucilage composition in *Plantago* spp. Journal of Experimental Botany **67:** 6481-6495

Rao M, Sadaphule P, Khembete M, Lunawat H, Thanki K, Gabhe N (2013) Characterization of psyllium (*Plantago ovata*) polysaccharide and its use as a binder in tablets. Indian Journal of Pharmaceutical Education and Research **47**

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research **43:** e47-e47

Rombolá-Caldentey B, Rueda-Romero P, Iglesias-Fernández R, Carbonero P, Oñate-Sánchez L (2014) Arabidopsis DELLA and two HD-ZIP transcription factors regulate GA signaling in the epidermis through the L1 box cis-element. The Plant cell **26:** 2905-2919

Saez-Aguayo S, Ralet M-C, Berger A, Botran L, Ropartz D, Marion-Poll A, North HM (2013) PECTIN METHYLESTERASE INHIBITOR6 promotes *Arabidopsis* mucilage release by limiting methylesterification of homogalacturonan in seed coat epidermal cells. The Plant Cell **25:** 308

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research **13:** 2498-2504

Shea Miller S, Bowman L-AA, Gijzen M, Miki BLA (1999) Early development of the seed coat of soybean (*Glycine max*). Annals of Botany **84:** 297-304

Shi L, Katavic V, Yu Y, Kunst L, Haughn G (2012) Arabidopsis glabra2 mutant seeds deficient in mucilage biosynthesis produce more oil. Plant J **69:** 37-46

Sinclair R, Rosquete MR, Drakakaki G (2018) Post-Golgi Trafficking and Transport of Cell Wall Components. Frontiers in Plant Science **9**

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences **102:** 15545

Tucker M, Ma C, Phan J, Neumann K, Shirley N, Hahn M, Cozzolino D, Burton R (2017) Dissecting the genetic basis for seed coat mucilage heteroxylan biosynthesis in *Plantago ovata* using gamma irradiation and infrared spectroscopy. Frontiers in Plant Science **8**

Verma A, Mogra R (2013) Psyllium (*Plantago ovata*) husk: A wonder food for good health.

Voiniciuc C, Schmidt MH-W, Berger A, Yang B, Ebert B, Scheller HV, North HM, Usadel B, Günl M (2015) MUCI10 produces galactoglucomannan that maintains pectin and cellulose architecture in *Arabidopsis* seed mucilage. Plant physiology**:** pp. 00851.02015

Voiniciuc C (2022) Modern mannan: a hemicellulose's journey. New Phytologist **234:** 1175-1184

Wang Z, Chen M, Chen T, Xuan L, Li Z, Du X, Zhou L, Zhang G, Jiang L (2014) TRANSPARENT TESTA2 regulates embryonic fatty acid biosynthesis by targeting FUSCA3 during the early developmental stage of Arabidopsis seeds. The Plant Journal **77:** 757-769

Western TL, Burn J, Tan WL, Skinner DJ, Martin-McCaffrey L, Moffatt BA, Haughn GW (2001) Isolation and characterization of mutants defective in seed coat mucilage secretory cell development in *Arabidopsis*. Plant Physiology **127:** 998-1011

Western TL, Young DS, Dean GH, Tan WL, Samuels AL, Haughn GW (2004) MUCILAGE-MODIFIED4 Encodes a Putative Pectin Biosynthetic Enzyme Developmentally Regulated by APETALA2, TRANSPARENT TESTA GLABRA1& GLABRA2 in the Arabidopsis Seed Coat. Plant Physiology **134:** 296

Xu Y, Wang Y, Du J, Pei S, Guo S, Hao R, Wang D, Zhou G, Li S, O'Neill M, Hu R, Kong Y (2022) A DE1 BINDING FACTOR 1–GLABRA2 module regulates rhamnogalacturonan I biosynthesis in Arabidopsis seed coat mucilage. The Plant Cell **34:** 1396-1414

Young RE, McFarlane HE, Hahn MG, Western TL, Haughn GW, Samuels AL (2008) Analysis of the Golgi apparatus in Arabidopsis seed coat cells during polarized secretion of pectin-rich mucilage The Plant Cell **20:** 1623-1638

Yu L, Shi D, Li J, Kong Y, Yu Y, Chai G, Hu R, Wang J, Hahn MG, Zhou G (2014) CELLULOSE SYNTHASE-LIKE A2, a glucomannan synthase, is involved in maintaining adherent mucilage structure in Arabidopsis seed. Plant Physiology **164**(4):1842-56

Yu L, Yakubov GE, Zeng W, Xing X, Stenson J, Bulone V, Stokes JR (2017) Multi-layer mucilage of *Plantago ovata* seeds: Rheological differences arise from variations in arabinoxylan side chains. Carbohydr Polym **165:** 132-141

Yuan Y, Teng Q, Zhong R, Ye Z-H (2016) Roles of Arabidopsis TBL34 and TBL35 in xylan acetylation and plant growth. Plant Science **243:** 120-130

Chapter 7

Summary and future directions

ꕥ•ꕥ

Thesis Summary

Plantago ovata husk, commonly called psyllium, is a valuable commodity due to its applications in industrial and health products. However, production can be affected by yield loss during unseasonal weather due to sensitivity to environmental changes. This study demonstrates that the use of advanced technologies like sequencing, bioinformatic analyses, microscopy techniques, and immunolabeling can provide new fundamental information about capsule-associated seed shattering and early seed development impacting mucilage production that may be useful in addressing some of these production issues by providing sensible targets in breeding programs.

Having a genome reference is highly beneficial for most crops, especially for *P. ovata*, as conventional breeding approaches have not been encouraging, and low diversity is likely to continue to hinder future genetic improvement via breeding. In Chapter 3, mining or collecting previous molecular data was performed. Ideally, a researcher will use genomic and transcriptomic data from the same plant to assemble the genome and create gene models to prevent a low mapping rate. However, valuable *P. ovata* data available since 2010 from public and private repositories have not been utilised so in Chapter 3 a combination of historical data and newly generated data was used to build a genome. The available data were evaluated or assessed carefully and mapping rates of the RNAseq data from a number of different cultivars matching to the *de novo* assembled genome were consistently high (up to 96%). Two sequencing technologies were used for genomic data, PacBio CLR long reads and Hi-C short reads, to achieve chromosome level resolution. Published short Illumina data could not be used to generate this because *P. ovata* contains a significant proportion of repeats (61.90%), meaning the sequences are relatively similar and remain too fragmented. Without a high-performance computer (HPC) facility and bioinformatics pipelines like Snakemake, the process will take forever or even fail. Most studies of non-model dicot or eudicot species would use the Arabidopsis protein database to annotate the transcriptomic assemblies. In this

study, the genome was annotated using plant protein (Viridiplantae) entries from UniProt consisting of 2,108 species, including Arabidopsis. Utilising advanced technology, computer resources, and the latest databases, we have thus generated a good quality *Plantago ovata* chromosome level assembly.

This study also used different microscopy techniques and immunolabelling in Chapters 4, 5, and 6. Dissecting microscopes were used to track seed and capsule developmental changes. Scanning electron microscopy (SEM) was used to study the surface of the capsule, especially in the dehiscence zone area where shattering occurs (Chapter 4). We used a standard microtome to prepare sections for staining and immunolabelling to compare wild-type and putative mutant developments (Chapters 4 and 5). A cryostat proved to be essential to prepare frozen sections for laser capture microdissection (LCM) of regions of interest from developing seeds for targeted RNA sequencing (Chapter 6). A customised ribodepletion kit was designed to deplete rRNA in *P. ovata* material so only informative parts of the transcriptome were left. Partnering with a sequencing facility, we have therefore used total RNA as low as 0.266 ng/mL to generate valuable data.

Once protocols were optimised to obtain and generate high-quality data, this study focused on answering fundamental questions related to two biological events: shattering (Chapter 4) and mucilage polysaccharide production (Chapter 6). Bioinformatic analyses for Chapters 4 and 6 included presence/absence gene expression analysis, Principal Component Analysis (PCA), differential gene expression analysis using EdgeR and gene set enrichment analysis (GSEA) using GO-terms as biological themes.

In Chapter 4, a combination of toluidine blue staining, immunolabelling and SEM show that the area of the capsule around the dehiscence zone where shattering happens is not as simple as previously described in 1981 (Lamba and Gupta). The upper and lower valves meet and form a region reminiscent of a knee joint with the operculum hook forming a connecting ligament. Secondary thickening of the walls of the cell layer forming the operculum hook is

visible where xylan is a key polysaccharide. Abscission between the two valves occurs across the two outer cell layers but it is not until the operculum hook detaches from the base that the lid and seeds are released as the capsule shatters. Morphological screening identified a putative capsule mutant called *accelerato* (*ace*) for which developmental events progressed faster than wild type and shattering was earlier. However, cell wall morphology and composition appeared to be similar to wild type. Transcriptome profiles were generated by RNA sequencing for both genotypes, wild type and *ace,* and for two developmental stages – young (8 – 10 DPA) and older (12 -14 DPA), all of which show unique patterns. Plant cell wall genes were enriched in older capsules, with a significant number of them linked to xylan biosynthesis and remodelling. The more rapid developmental program of *ace* capsules is hypothesised to be due to a mutation affecting oxidative phosphorylation (OXPHOS) genes which have downstream effects on photosynthesis-related, *PoADGP1* and *PoSTK* genes.

In Chapter 5, one putative mutant with a higher yield of mucilage per seed, named *raya* (*ray*) was selected from a screen of 201 gamma-irradiated mutants. The mutant population available at the University of Adelaide contains many lines for which mucilage amount or behaviour is different to wild-type but *ray* was selected very carefully so that only mucilage quantity is different from the wild-type, with no other pleiotropic phenotypes. This mutant was then used in Chapter 6 to identify genes involved in mucilage polysaccharide production in the early stages of seed development. We profiled 40 samples from the two different tissue types, mucilage secreting cells (MSC) and the integument (INT) tissue that comprise the husks of mature seeds, at three different time points. Twenty-seven pairwise comparisons were made to detect DEGs and enrichment clusters. The DEG and enrichment cluster profiles were diverse among comparisons. The number of DEGs varied among comparisons from zero to 3,443 genes. A cell wall cluster was enriched in *ray* INT at 2 DPA but was not present in the wild-type samples at the same age. A total of 48 sequences with similarity to known transcription factors were identified from wild-type and *ray* samples, with two specific to

wild-type and seven specific to *ray*. One out of these seven sequences was similar to *GL2* (*GLABRA 2),* which is known to control mucilage polysaccharide production. Other identified TFs may be involved in mucilage polysaccharide biosynthesis and will need further investigation. *GT43* genes (*IRX9* and *IRX14*) previously hypothesised to be important for xylan biosynthesis (Jensen et al., 2013; Voiniciuc et al., 2015) but absent from previous *P. ovata* studies (Jensen et al., 2014), were confirmed to be involved in mucilage biosynthesis in *P. ovata*. In this study, these genes, and others in the xylan pathway, were expressed in early development but mainly in MSC, not in INT tissues and at a higher level in *ray* samples. These results indicate that these two tissues have unique transcriptomic profiles. It also indicates that this approach has enabled the identification of expressed genes specific to the MSC layer or INT tissues, whereas they were both diluted and comingled in whole developing seed samples

In summary, we have generated a high-quality reference genome (Chapter 3) which was used as a reference assembly for RNAseq experiments exploring capsule development (Chapter 4) and polysaccharide biosynthesis in early seed development (Chapter 6). Fundamental questions regarding the *P. ovata* shattering mechanism have been explained (Chapter 4). A candidate mutant with a higher yield of mucilage has been selected from the gamma-irradiated population and specific tissues were profiled and compared to the wild type (Chapter 5). Generating spatial and temporal specific transcriptomic data in both cases was rewarding as much information was gained and remains to be explored. Understanding the development of the capsule (fruits) and seeds is essential, especially where concepts that have not been updated in the past four decades can be expanded using modern technology and analysis techniques. All the techniques used in this study can be applied to other crops, especially non-model species, and as demonstrated here can provide valuable information for breeding programs and for improving our understanding of seed developmental biology.

Future Directions

The *Plantago ovata* genome and its thorough annotation are expected to open the door for modern genetic improvement of this plant. Such improvements could include eliminating shattering, increasing mucilage production and engineering cell wall polysaccharides, especially for different end-uses of xylan. Fundamental knowledge about *P. ovata* domestication, the role of epigenetic control and programmed cell death during fruit and seed development can be explored from the results generated from this study.

Eliminating or reducing shattering

Even though the mutant *accelerato* (*ace*) has a more rapidly developing capsule it still has a shattering problem. Ideally, we need to compare shattering and non-shattering lines, but a non-shattering mutant has never been found, although it may exist in the mutant population at the University of Adelaide and be found by more extensive screening. However, we have identified some of the candidate genes important for capsule development by tracing the underlying mutation in *ace*, including *SEEDSTICK* (*STK*) and *ARABIDOPSIS DEHISCENCE ZONE POLYGALACTURONASE1 (ADPG1)*. *STK* is induced in *ace*, and *ADPG1* was upregulated in both WT and the mutant, but the increase was higher in *ace*. Since mutated genes were found in the OXPHOS pathway, the connection between these genes needs to be explored further. While a non-shattering gamma-irradiated mutant has not been found, we could start comparing the expression of candidate genes between two related species. The Burton group at the University of Adelaide has a collection of Australian *Plantago* wild relatives which display reduced or no shattering, such as *Plantago cunninghamii*, *Plantago turrifera* and *Plantago drummondii* (personal communication Dr James Cowley). This could be done by a combination of gene expression quantification using qPCR, by performing RNAseq on capsule material or by genome sequencing. For option one, primers based on alignment of the *P. ovata* sequences with those of other plants to define the conserved regions would have a good chance to amplify target cDNAs in its wild relatives. The second approach

to perform RNAseq on the wild-relative is a better option to provide information about more broadly affected gene networks, especially when compared to the *P. ovata* data. These molecular analyses could be combined with morphological data relating to capsule development, particularly concerning the identification and fate of the specific cell layers in the capsule valves, where there may be no equivalent DZ.

Increasing mucilage production

The putative mutant *ray* has increased mucilage without defects observed during seed development. Ideally, to understand mucilage polysaccharide production, we need to compare wild type with a mutant with a lower quantity or even no mucilage. However, from the screen where mutants with low mucilage were selected these plants did not grow well and died during flowering. Instead, lines with greater amounts of mucilage were chosen. The benefit of using *ray* is that we do not have to edit the genome as the plants already produce higher quantities of mucilage, but this mutant still needs backcrossing with wild type to fix the favourable change and remove other background mutations that are likely to still be present even in the M5 generation. Until this is done, and the mutation identified, *ray* will not be of use in the breeding program.

Reverse genetics still needs to be performed for candidate genes to validate their function. We can start with the known transcription factor *GL2*; will knocking out this gene in *P. ovata* give the same phenotype as in Arabidopsis, causing no seed mucilage. Then, we can look at the other six transcription factors (Tables 2 and 3 in Chapter 6) and genes amongst the leading-edge subset in the cell wall metabolic cluster (Table 5 in Chapter 6). To perform reverse genetics, we need to use a system for designing a CRISPR/Cas9 construct to target genes of interest, a transformation system to deliver the construct to the explant (plant tissues) and a tissue culture system to effectively regenerate plants. A new functional transformation system (unpublished), developed by researchers in the Burton group is now currently available at the

University of Adelaide and a number of the candidate genes described in this thesis will be some of the first targets for functional analysis *in planta.*

Engineering cell wall polysaccharides

Heteroxylan is the main component of *P. ovata* mucilage. This polysaccharide has been used in many applications, including health supplements, to lower blood cholesterol and control weight gain and is a key ingredient in gluten-free foods because of its viscosity and water-holding properties. However, psyllium is poorly fermented by microbes in the human gut, probably because it is highly branched and hydrolase enzymes can get only limited access. This prevents the polysaccharide from being converted during fermentation into beneficial compounds that protect the gut such as short-chain fatty acids, although psyllium is still useful as a bulk dietary fibre. Having the genome (Chapter 3) has facilitated the identification of six xylanases that are likely to be useful in hydrolysing *Plantago* heteroxylan as well as xylans from other species. This will improve our analytical capability as well as being potentially useful in industrial applications. The GT61 enzymes profiled in Chapter 6 may also be useful in this way and they could be assessed directly using a technique such as TILLING (targeting induced local lesion IN genomes), making use of the mutant population or altered via CRISPR-Cas9 and plant transformation. Given that xylan is the second-largest polysaccharide component of cell walls in plant biomass, manipulating its content in plant tissues other than seed mucilage could be damaging while lines without mucilage can still develop into new plants. *P. ovata* seed mucilage is therefore an ideal system for exploring and engineering xylan composition, allowing the transfer of knowledge to other crops such as wheat and trees grown for timber, and benefitting industrial uses in food and biofuel production. In this way the resources and research in this thesis can be of use in a much broader context.

P. ovata domestication

Several resources from the genome (Chapter 3) can be used to study domestication in *P. ovata*. Three regions in the *P. ovata* nuclear genome were predicted as nuclear mitochondrial and plastid DNA during data curation while submitting the genome to the NCBI database. We could investigate gene transfer from organelles to the nuclear genome. Evaluating the GC content of the genome sequence confirms that *P. ovata* has a higher AT content, which is hypothesised to be common in domesticated plants. Comparisons of GC content between *P. ovata* and its wild relatives could confirm this hypothesis. A list of LTR Copia and Gypsy retrotransposons may help track *Plantago* domestication as well, since domestication and improvement in crops affect transposable element (TE) content. Lastly, telomere sequences can help study evolution by comparing *P. ovata* sequence with other plants.

Epigenetic control and programmed cell death

Epigenetic effects and programmed cell death have not been deeply explored here. However, some of the datasets from this study can be used to explore these pathways, especially those involved in fruit and seed development. In Chapter 3, we generated a list of lncRNA and mRNA candidates for future functional analysis. We also identified the locations and sequences of genes linked to histone modifications and DNA methylation. In Chapter 4, the expression of genes in the plant organ senescence cluster was found to be enriched. Information about genes in this cluster could be connected to capsule shattering and how programmed cell death influences shattering events. In Chapter 6, the gene silencing cluster was downregulated when comparing two-time points. Histone modification, gene silencing and transcription regulation clusters were found to be downregulated by comparing two tissues of the developing seed, the MSC and the INT, that have very different functions and fates. There is a rich vein of information in the RNAseq datasets that can be mined by future researchers interested in unravelling important processes in seed development and mucilage

polysaccharide production, relevant not only to *Plantago*, but also to the many other myxospermous plant species we are only now starting to be interested in.

References

Jensen JK, Johnson N, Wilkerson CG (2013) Discovery of diversity in xylan biosynthetic genes by transcriptional profiling of a heteroxylan containing mucilaginous tissue. Frontiers in plant science **4:** 183-183

Jensen JK, Johnson NR, Wilkerson CG (2014) *Arabidopsis thaliana* IRX10 and two related proteins from psyllium and *Physcomitrella patens* are xylan xylosyltransferases. The Plant Journal **80:** 207-215

Voiniciuc C, Günl M, Schmidt MH-W, Usadel B (2015) Highly branched xylan made by IRREGULAR XYLEM14 and MUCILAGE-RELATED21 links mucilage to *Arabidopsis* seeds. Plant physiology **169:** 2481-2495

Appendix I

A small-scale fractionation pipeline for rapid analysis of seed mucilage characteristics

ꕥ•ꕥ

Statement of Authorship

Title of Paper	A small-scale fractionation pipeline for rapid analysis of seed mucilage characteristics
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Cowley, J.M., Herliana, L., Neumann, K.A. *et al.* A small-scale fractionation pipeline for rapid analysis of seed mucilage characteristics. *Plant Methods* **16,** 20 (2020). https://doi.org/10.1186/s13007-020-00569-6

Principal Author

Name of Principal Author (Candidate)	James M. Cowley		
Contribution to the Paper	Conceived the study, designed and tested the method, analysed data, wrote the manuscript		
Overall percentage (%)	75%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	19/5/2020

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);
ii. permission is granted for the candidate in include the publication in the thesis; and
iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Lina Herliana		
Contribution to the Paper	Designed and conducted the mutant screen experiment. Assisted with data analysis and manuscript preparation.		
Signature		Date	9/6/2020

Name of Co-Author	Kylie A. Neumann		
Contribution to the Paper	Collected and assisted in data analysis for field trial quality testing		
Signature		Date	19/5/2020

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Silvano Ciani		
Contribution to the Paper	Provided field trial materials for testing		
Signature		Date	19/5/2020

Name of Co-Author	Virna Cerne		
Contribution to the Paper	Provided field trial materials for testing		
Signature		Date	19/5/2020

Name of Co-Author	Rachel A. Burton		
Contribution to the Paper	Conceived the study and contributed to writing the manuscript		
Signature		Date	19/5/2020

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Cowley *et al. Plant Methods* (2020) 16:20
https://doi.org/10.1186/s13007-020-00569-6

Plant Methods

RESEARCH Open Access

A small-scale fractionation pipeline for rapid analysis of seed mucilage characteristics

James M. Cowley[1,2], Lina Herliana[2], Kylie A. Neumann[1,2], Silvano Ciani[3], Virna Cerne[3] and Rachel A. Burton[1,2*]

Abstract

Background: Myxospermy is a process by which the external surfaces of seeds of many plant species produce mucilage—a polysaccharide-rich gel with numerous fundamental research and industrial applications. Due to its functional properties the mucilage can be difficult to remove from the seed and established methods for mucilage extraction are often incomplete, time-consuming and unnecessarily wasteful of precious seed stocks.

Results: Here we tested the efficacy of several established protocols for seed mucilage extraction and then downsized and adapted the most effective elements into a rapid, small-scale extraction and analysis pipeline. Within 4 h, three chemically- and functionally-distinct mucilage fractions were obtained from myxospermous seeds. These fractions were used to study natural variation and demonstrate structure–function links, to screen for known mucilage quality markers in a field trial, and to identify research and industry-relevant lines from a large mutant population.

Conclusion: The use of this pipeline allows rapid analysis of mucilage characteristics from diverse myxospermous germplasm which can contribute to fundamental research into mucilage production and properties, quality testing for industrial manufacturing, and progressing breeding efforts in myxospermous crops.

Keywords: Mucilage, Myxospermy, Extraction, Polysaccharide, *Plantago ovata*, Flax, Chia, Psyllium

Background

In a process called myxospermy, seeds of many plants produce viscous polysaccharide gels called mucilage when imbibed in water. The mucilage of *Arabidopsis thaliana* has often been used as a proxy for studying cell wall biosynthesis [1–8]. More recently other myxospermous species like *Linum usitatissimum* and *Plantago ovata* have also been adopted as genetic models [9–14] revealing the utility that novel systems can have in unravelling complex synthetic pathways. Furthermore, these novel model systems have the added benefit of being directly commercially-relevant. *P. ovata* (psyllium) and *L. usitatissimum* (flaxseed) mucilage are used as gums with varied applications in the food and health industries. Both are used as natural food structuring ingredients and gluten replacements [15–18] and are rich sources of dietary fibre shown to prevent various gastrointestinal diseases [19–21]. A comprehensive myxospermous model system would allow gene-structure–function links to be made but there remains a technical disconnect between these facets. The functional study of myxospermous species preceded their use as genetic models and the scale and precision of the extraction techniques have generally not been updated since. A significant number of researchers use the methods of Sharma and Koul [22], Balke and Diosady [23], or similar. These methods are simple, effective and robust, using a magnetic stirrer to heat and agitate a seed/water mixture followed by straining to isolate released mucilage from seeds. However, there are several technical issues that limit the use

*Correspondence: rachel.burton@adelaide.edu.au
[1] Australian Research Council Centre of Excellence in Plant Cell Walls, School of Agriculture, Food and Wine, University of Adelaide, Waite Campus, Urrbrae, SA, Australia
Full list of author information is available at the end of the article

of these techniques in screening applications. Firstly, the techniques are not high-throughput, generally requiring 3–4 h to process a single sample (per magnetic stirrer). Secondly, the quantity of mucilage produced is excessive for downstream chromatographic and yield analyses which require milligram-scale quantities or less. Thirdly, the techniques often offer incomplete extraction, leaving a significant amount of mucilage adhered to the seed. It is also important to note that seed mucilage is not homogenous. Its multi-layered nature is evident simply by visual inspection of stained expanded mucilage in nearly all species [24] The dual-layered nature of *Arabidopsis thaliana* mucilage has been the basis of many studies on cell wall polysaccharide biosynthesis [25] and Yu et al. [26–28] have recently highlighted the importance of fractionating mucilage to effectively unravel structural differences that underlie polysaccharide functionality.

Here we describe a pipeline suitable for the rapid extraction and fractionation of quantities of seed mucilage suitable for yield and chromatographic analyses. Within four hours, three chemically- and functionally-distinct fractions can be isolated from 24 samples per shaking incubator and the use of a shaking incubator allows adjustment in time, temperature and fractionation profile. The utility of this pipeline is demonstrated through its ability to: identify intergeneric and interspecific variation in seed mucilage extractability, yield and composition, screen for known quality parameters in field-grown myxospermous samples, and identify lines of interest from germplasm collections.

Methods

Materials

Arabidopsis thaliana seeds (ecotype *Columbia-0*) were grown as per Tucker et al. [29]. Flax (*Linum usitatissimum*) and chia (*Salvia hispanica*) seeds were purchased from Woolworths (Frewville, South Australia). *Plantago ovata* and *Plantago cunninghamii* seeds were obtained and bulked from sources listed in Phan et al. [30]. *P. ovata* varieties were grown in field trials conducted in 2017 and 2018 in Kununurra, Western Australia. Gamma-irradiated *P. ovata* mutants used for germplasm screening were obtained from a glasshouse-grown population described previously [10].

Once harvested or purchased, all seeds were dried at 37 °C for at least 72 h and then stored in sealed containers at room temperature until analysis.

Reagents and solutions

Ruthenium red hydrate (#C075) was purchased from ProSciTech, Australia and the staining solution was prepared at 0.01% w/v in water following Arsovski et al. [5]. To prevent bubble formation on the seed surfaces during imaging, the staining solution was sonicated for 5 min under vacuum to remove dissolved gases. KOH and HCl (Sigma-Aldrich) were made to a 0.2 M solution in water.

Mucilage staining

Expanded seed mucilage was observed in situ following Arsovski et al. [5] to compare and validate extraction techniques. After positioning seeds on a microscope slide (Rowe GM2715, Australia) staining solution was added beneath a coverglass (ProSciTech No. 1, Australia) and images were captured on a dissecting microscope (Zeiss Stemi 2000-C, Germany) equipped with a colour digital camera (Zeiss AxioCam ERc 5s, Germany).

Conventional seed mucilage extraction techniques

The effectiveness of total mucilage extraction by the fractionation pipeline described here was validated against previously published methods by Balke and Diosady [23], Yu et al. [26], Sharma and Koul [22], and Voiniciuc et al. [4]. Balke and Diosady's method (also used previously by our group, Phan et al. [30]) is simple, stirring seeds and water heated to 80 °C on a magnetic stirrer for 90 min after which the liberated mucilage is strained through nylon mesh to remove seeds. Yu et al.'s method uses an extended extraction (4 h) with RT 0.2 M solution of KOH. Sharma and Koul's method combines seeds with dilute acid (0.2 M HCl) in a conical flask, which is stirred on a heated magnetic stirrer until the seeds have changed colour (20 min). Seeds are strained through nylon mesh and washed twice with hot water. Finally, Voiniciuc's method uses the physical force (30 Hz for 30 min) of a tissue disruptor-type mixer mill.

Rapid small-scale mucilage fractionation pipeline

The rapid small scale fractionation of quantities of seed mucilage suitable for yield and chromatographic analyses was achieved using the following protocol incorporating elements of methods by Yu et al. [26] and Voiniciuc et al. [4]. For similarly sized seeds, seeds can be counted or for variably sized seeds 30 mg (±0.5 mg) of seeds (exact mass recorded) can be weighed. Once the pre-extraction mass was recorded, seeds were added to a 2 mL microcentrifuge tube followed by 1.5 mL of RT DI H_2O (Fig. 1a). Tubes were vortexed briefly to break surface tension and ensure all seeds are immersed. To obtain the first mucilage fraction (Fig. 1b), tubes were incubated for 1.5 h at 25 °C in a shaking incubator (Eppendorf ThermoMixer® Comfort, Germany) with agitation at 1300 rpm then centrifuged (Eppendorf 5424, Germany) for 2 min at 13,000 rpm. Ensuring that the pelleted adherent mucilage and seeds are not disturbed, the tubes were removed from the centrifuge and using a 1000 µL laboratory pipette, the supernatant—the cold water extractable (CWE) mucilage fraction—was

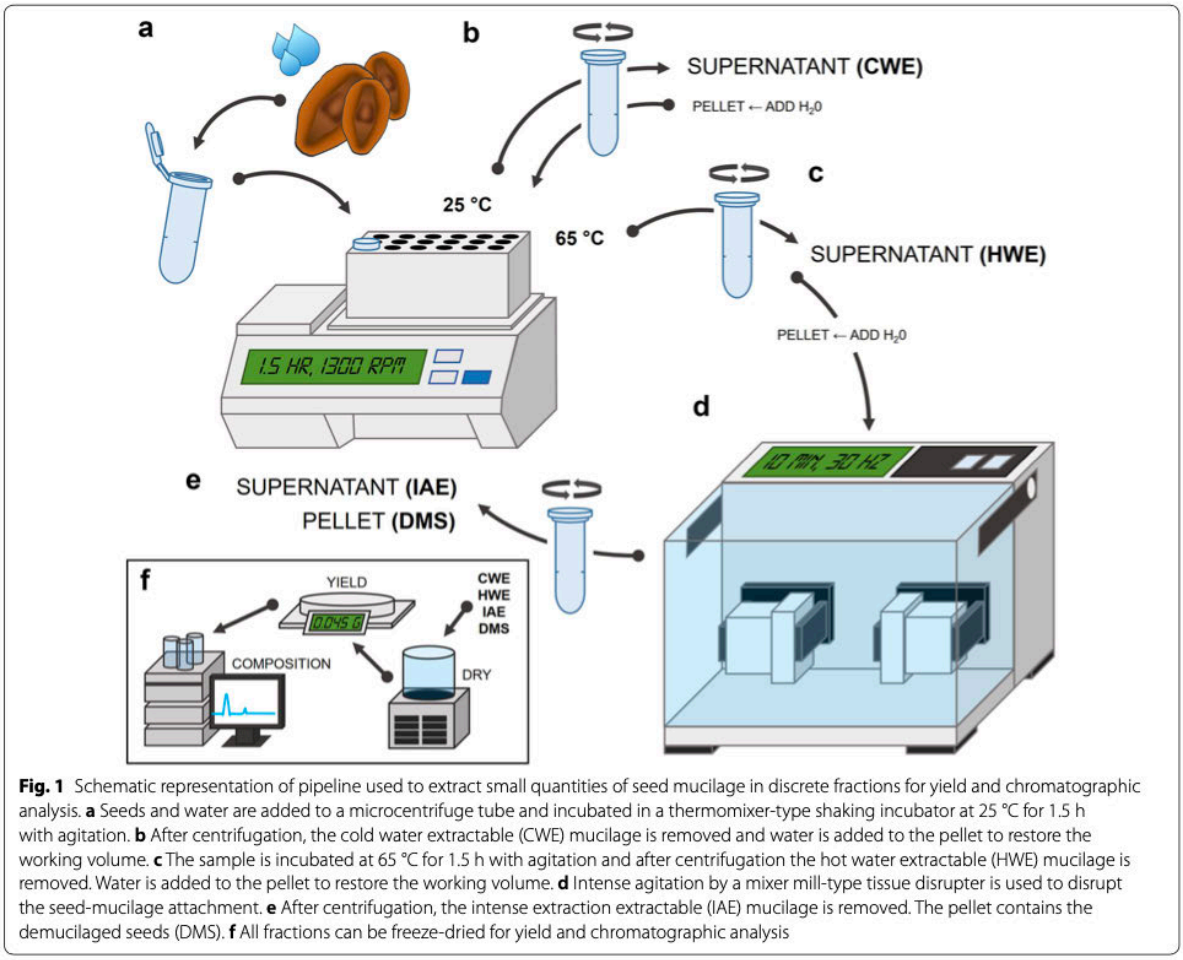


Fig. 1 Schematic representation of pipeline used to extract small quantities of seed mucilage in discrete fractions for yield and chromatographic analysis. **a** Seeds and water are added to a microcentrifuge tube and incubated in a thermomixer-type shaking incubator at 25 °C for 1.5 h with agitation. **b** After centrifugation, the cold water extractable (CWE) mucilage is removed and water is added to the pellet to restore the working volume. **c** The sample is incubated at 65 °C for 1.5 h with agitation and after centrifugation the hot water extractable (HWE) mucilage is removed. Water is added to the pellet to restore the working volume. **d** Intense agitation by a mixer mill-type tissue disrupter is used to disrupt the seed-mucilage attachment. **e** After centrifugation, the intense extraction extractable (IAE) mucilage is removed. The pellet contains the demucilaged seeds (DMS). **f** All fractions can be freeze-dried for yield and chromatographic analysis

transferred to a clean, pre-labelled microcentrifuge tube. This transfer may require multiple steps based on the volume of the supernatant. The volume of the tube contents comprising the pelleted mucilage and seeds was returned to approximately 1.5 mL with RT DI H_2O based on the tube markings (different samples may require slightly different volumes). To obtain the next mucilage fraction (Fig. 1c), a similar process was employed but at a warmer temperature: tubes were incubated for 1.5 h at 65 °C with agitation at 1300 rpm then centrifuged for 2 min at 13,000 rpm. Again the supernatant—the hot water extractable (HWE) mucilage fraction—was transferred to a clean, pre-labelled microcentrifuge tube using a 1000 µL laboratory pipette. After removal of the HWE fraction, the volume of the pellet is significantly reduced as only the most extraction-resistant mucilage remains tightly adhered to the seed. After adjusting the volume of tube contents to approximately 1.5 mL with RT DI H_2O, the tubes were agitated intensely at 30 Hz for 10 min on a tissue disruptor-type mixer mill (Retsch MM400, Germany) using a microcentrifuge tube adapter (Fig. 1d). Tubes were centrifuged for 2 min at 13,000 rpm and the supernatant—the intense agitation extractable (IAE) mucilage fraction—was transferred to a clean, pre-labelled microcentrifuge tube using a 1000 µL laboratory pipette.

From each sample, a cold water extractable (CWE), hot water extractable (HWE) and intense extraction resistant (IAE) fraction of seed mucilage has been obtained along with the corresponding demucilaged seeds (DMS) (Fig. 1e). These four fractions were frozen at − 80 °C for 24 h and then freeze-dried (Labconco Freezone 6, US) to a constant weight. Freeze-dried mucilage and demucilaged seeds were transferred to a microbalance with 0.01 mg resolution (Shimadzu AUW220D, Japan) with fine-tip tweezers to calculate mucilage yield (Fig. 1f).

Yield of mucilage fractions can be calculated using the following equation:

$$\text{Yield}(\%) = \left(\frac{\text{mass of freeze dried mucilage}}{\text{mass of seeds pre-extraction}}\right) \times 100.$$

Optional—isolated fractions may be pipetted into a 2000 µL 96 well deep well plate (Eppendorf, Germany) in place of new microcentrifuge tubes which can become unwieldy when dealing with large sample numbers. The deep well plates can accommodate many samples and several plates will fit simultaneously into a freeze-dried unit for bulk processing.

Monosaccharide profiles of fractionated seed mucilage

Freeze-dried mucilage was dispersed in water at 2 mg/mL (w/v) and an 800 µL aliquot was added to 200 µL of 5 M H_2SO_4 (final H_2SO_4 concentration of 1 M) and hydrolysed at 100 °C for 3 h as per Phan et al. [30]. Monosaccharides released by acid hydrolysis were derivatised with 1-phenyl-3-methyl-5-pyrazoline (PMP) and then separated by reversed phase high performance liquid chromatography (RP-HPLC) following Comino et al. [31] with modifications to the column and eluents listed in Hassan et al. [32]. Area under the peaks was compared to standard curves of mannose, ribose, rhamnose, glucuronic acid, galacturonic acid, glucose, galactose, xylose, arabinose and fucose [33].

Water absorption assay

After weighing 20 seeds into a 2 mL microcentrifuge tube, 1 g of water was added, and mucilage was allowed to expand undisturbed for 45 min at 25 °C. Using a 1 mL syringe without a needle, unabsorbed water was removed and weighed. Water absorption capacity can be calculated using the equation:

$$\text{Water absorption capacity}(\text{g/g}) = \frac{\text{Initial weight of water added} - \text{weight of unabsorbed water}}{\text{Initial weight of seeds added}}$$

Results and discussion

This protocol achieves total mucilage extraction

Figure 2 shows *P. ovata* seeds with stained expanded seed mucilage before and after four methods of mucilage extraction. While hot water was effective at reducing the size of the mucilage envelope (Fig. 2b), mucilage in the inner layer is more densely packed than the removed soluble fraction and thus a large amount of the mucilage remains [34]. Yu et al. [26] reported that an extended (4 h) extraction with 0.2 M KOH, a chaotropic agent, was sufficient to remove the adherent mucilage layer by disrupting hydrogen bonds in the mucilage. We confirm that this treatment is effective at removing the majority of seed mucilage (Fig. 2c) although the most strongly adherent portion remained on all treated seeds (n = 30). Mucilage removal in acid was similarly effective (Fig. 2d), however the mode of action of the acid was to hydrolyse the mucilage in situ which is not useful if any downstream functional or linkage analyses are required. An appealing alternative was the non-chemical extraction method devised by Voiniciuc et al. [4] who were able to efficiently extract all adherent mucilage from *Arabidopsis* seeds using intense agitation on a tissue disruptor-type mixer mill. The physical force of the shaking was sufficient to disrupt the mucilage-seed attachment and

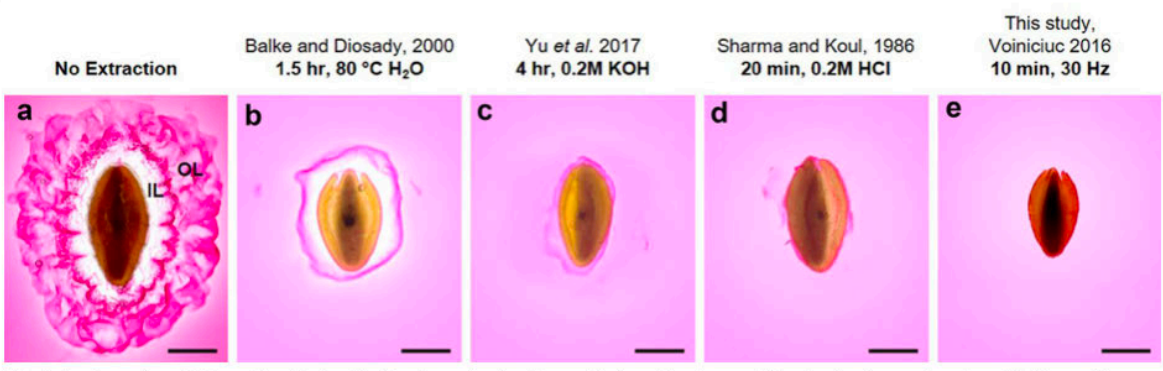


Fig. 2 Seed mucilage that remains attached to *P. ovata* seeds after the application of various published extraction protocols and staining with 0.1% w/v ruthenium red. **a** The characteristic mucilage architecture of *P. ovata* seed mucilage. **b** After extended extraction in hot water the fine structure of the outer layer (OL) of mucilage has been lost leaving an inner adherent layer (IL). **c** The use of a chaotropic agent, KOH, significantly reduces the size of the mucilage envelope leaving only a film of mucilage that is resistant to chemical extraction. **e** Mucilage has been hydrolysed in situ and only a thin feathery layer remains. **e** Extraction by intense agitation leads to total mucilage removal. Scale = 1 mm

disperse the polysaccharides. We corroborate the efficacy of this method on *P. ovata* where close to 100% of seed mucilage was removed (Fig. 2e). We also observed that mucilage staining of seeds that sequentially underwent a hot water extraction before 0.2 M KOH, 0.2 M HCl or 30 Hz agitation were no different to those that were not extracted with hot water as a first step (data not shown).

The monosaccharide profiling of fractionated *P. ovata* mucilage shows that our small-scale fractionation technique is directly comparable to the larger scale technique published by Yu et al. [26], the original study by Guo et al. [35] on which their work is based and a similar work published earlier by Marlett and Fischer [36] (Additional file 1: Table S1).

The extraction pipeline effectively provides material for identifying variation in seed mucilage characteristics

Figure 3 shows the appearance of the expanded seed mucilage envelope of *A. thaliana* (a–d), *L. usitatissimum* (flaxseed) (e–h), *S. hispanica* (chia) (i–l), *P. ovata* (psyllium) (m–p) and an Australian native relative of psyllium, *P. cunninghamii* (q–t) before and after each mucilage fractionation step in the pipeline described here, which culminates in total extraction (Fig. 4a). After CWE and HWE extraction, the mucilage envelope of *L. usitatissimum* and *A. thaliana* is significantly reduced in size compared to *S. hispanica*, *P. ovata* and *P. cunninghamii*. These changes were reflected in the differences between the ratios of extracted fractions (Fig. 4b), where *L. usitatissimum* and *A. thaliana* were most susceptible to extraction, yielding the largest proportion of water extractable (CWE + HWE) components. The easy removal of the delicate outer layer of mucilage in *A. thaliana* and *L. usitatissimum* is consistent with previous findings and established fractionation techniques [4, 34, 37–40]. Contrastingly, *Plantago* and *S. hispanica* mucilage has been reported to require more effort to efficiently extract total mucilage. For *S. hispanica*, mucilage extraction in cold water is not efficient [41] while extended hot water extractions yield only slightly greater quantities [42–44]. We corroborate these findings, reporting that only half of the mucilage is extractable by hot water (Fig. 4b). Efficient mucilage extraction from *Plantago* species is also difficult, often requiring multiple physical and/or chemical extraction steps [26, 35, 36, 45, 46]. While total yields of mucilage between *P. ovata* and *P. cunninghamii* were similar (Fig. 4a), there are clear interspecific differences in the relative proportion of each fraction (Fig. 4b). Changes in appearance of the mucilage envelope of *P. ovata* after CWE were more noticeable than for *P. cunninghamii*, which appears relatively unchanged and is reflected in CWE yield which was lower for *P. cunninghamii*. Yield of HWE mucilage was greater for *P. cunninghamii*, with less IAE mucilage than *P. ovata*.

Some, if not all, of the differences in the relative proportion of each mucilage fraction of the species studied can be ascribed to mucilage polysaccharide composition and the associated difference in properties. Monosaccharide analysis confirmed significant differences in mucilage composition between genera and species and their isolated mucilage fractions (Fig. 4c). Monosaccharide analysis also confirmed previous findings that mucilage fractions from *A. thaliana* and *L. usitatissimum* are rich in rhamnose and galacturonic acid [4, 47–50], components of pectin, a highly water-soluble polysaccharide, the presence of which may contribute to overall ease of extraction in these species. It is the presence of minor mucilage components in the HWE and IAE fractions that are known to affect the mucilage properties including fractionality. In *A. thaliana*, monosaccharide profiling of the HWE and IAE fractions (containing the adherent mucilage) confirms previous findings of a molar reduction in rhamnose and galacturonic acid residues and an increase in non-cellulosic glucose, mannose, galactose and xylose [49, 50], components of minor polysaccharides like xylan and glucomannan that are well-known to interact with and tether the adherent mucilage at the seed surface [4, 7, 51–53]. Similarly, rhamnose and galacturonic acid residues were reduced in the HWE and IAE fractions of *L. usitatissimum*, along with increases in xylose and arabinose residues associated with heteroxylan, known to significantly alter the functional properties of RG-I [47]. In *P. ovata and P. cunninghamii*, the three mucilage fractions contained high levels of xylose and arabinose (heteroxylan) with a smaller amount of rhamnose and galacturonic acid (pectin), congruent with previous findings by Phan et al. [30]. Like both *A. thaliana* and *L. usitatissimum*, pectin-associated monosaccharides are enriched in the CWE fractionation and diminish with further fractions. While the presence of pectin has been proposed to modulate the extractability of the major heteroxylan component in *Plantago* mucilage, studies have shown that heteroxylan branching has the most significant influence on the mucilage properties including the extractability [26–28]. In both *P. ovata* and *P. cunninghamii,* the ratio of arabinose to xylose residues (estimation of the degree of sidechain branching) increased with resistance to extraction, in line with those studies. However, more explicit structural characterisation will be needed to define the fine structure and its relationship to interspecific differences in extractability. The mucilage of *S. hispanica* is unique among the species studied in that its constituent polysaccharide(s) have not been found in any other genera [54]. Its unique structure containing xylose, glucose and glucuronic acid residues is consistent

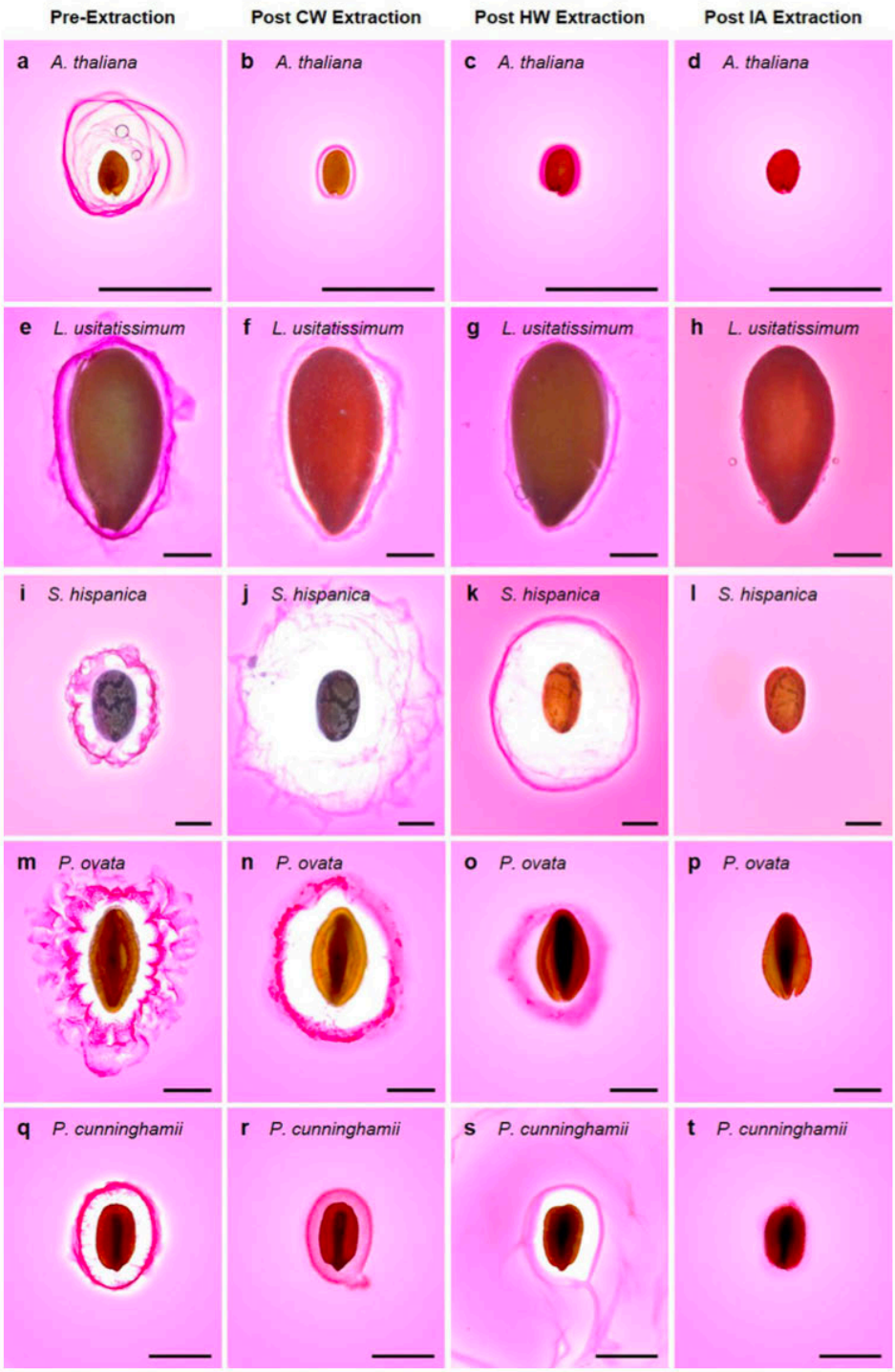


Fig. 3 Visual inspection of the ruthenium red-stained seed mucilage of five myxospermous species (*Arabidopsis thaliana*, *Linum usitatissimum*, *Salvia hispanica*, *Plantago ovata* and *Plantago cunninghamii*) after sequential fractionation shows that the size of the mucilage envelope is sequentially changed corresponding to removal/dispersion of constituent polysaccharides. Note that images are not taken of the same seed as each seed was disposed of after imaging. Scale = 1 mm

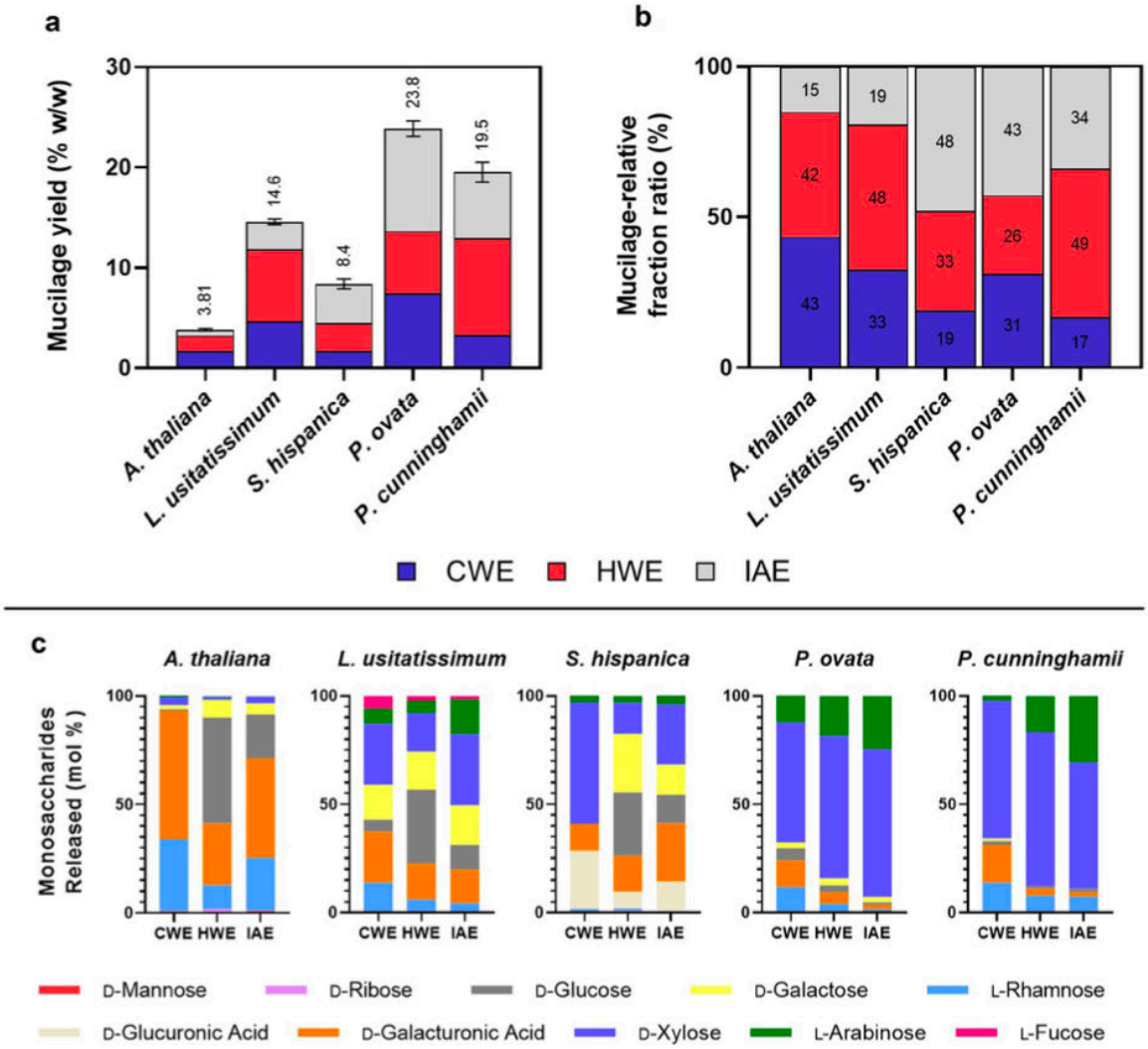


Fig. 4 The extraction pipeline is effective at providing material suitable for identifying significant intergeneric and interspecific variation in seed mucilage yield, fractionation profile and composition. **a** Yield of total seed mucilage divided into constituent fractions. Error bars refer to standard deviation in total mucilage yield of five biological replicates. **b** Extractability of seed mucilage represented as the share of each fraction within total extracted mucilage. **c** Monosaccharide analysis of fractionated mucilage showing intergeneric, interspecific and intraspecific differences in mucilage chemical composition. Values presented are in molar ratio of quantified monosaccharides released by acid hydrolysis. *CWE* cold water extractable, *HWE* hot water extractable, *IAE* intense agitation extractable

with the monosaccharide data (Fig. 4c), although the molar ratios between the constituents varies by fraction suggesting that fine structure and/or interactions with minor components (from which the other monosaccharides detected are derived) influences the extractability.

The pipeline has utility in quality testing of mucilaginous species

The production of high-quality psyllium gum from *P. ovata* seeds is hampered by agronomic issues which cause poor quality, damaged seeds [55]. Damaged seed coat allows leakage of endosperm components during extraction which alter the functional properties and cause significant discolouration due to phenolic browning which is undesirable in many applications (Cowley

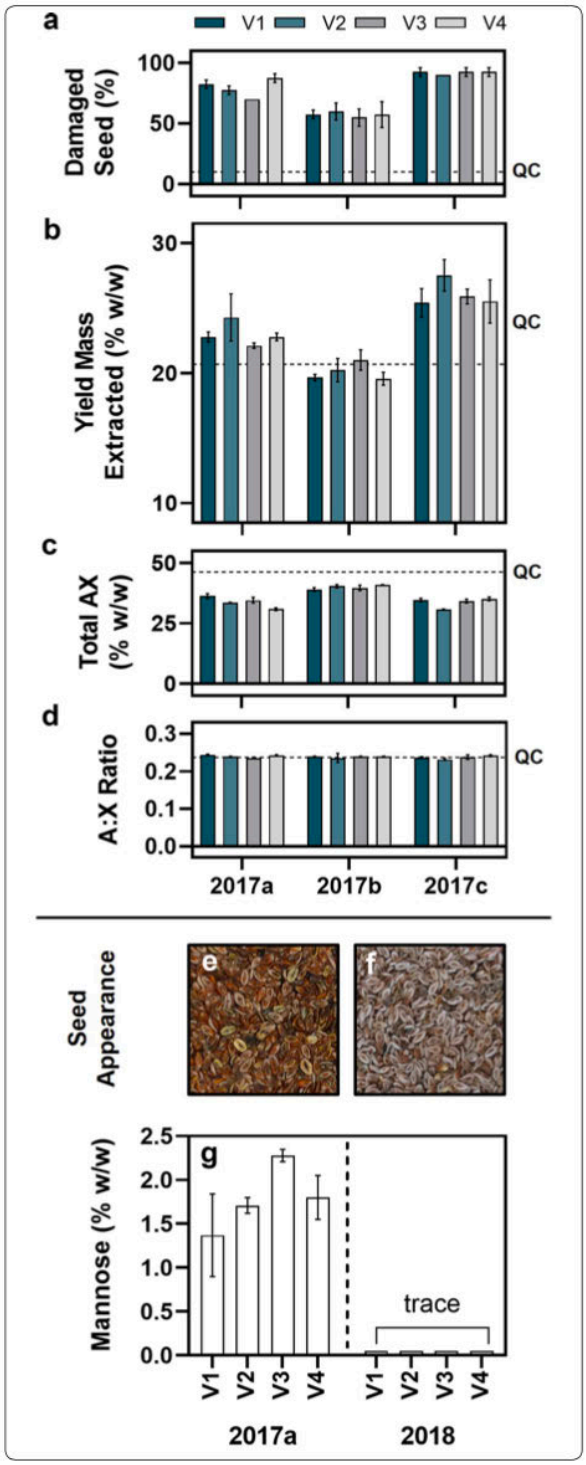


Fig. 5 In field-grown samples of *Plantago ovata*, the extraction pipeline has utility in quality testing when coupled with yield and chromatographic analysis. **a** Visual inspection of damaged seed provides a baseline quality score. **b** Yield of total mucilage varied most significantly between times of sowing with a minor intervarietal effect. **c** Heteroxylan content (the proportion of arabinose and xylose residues in extracted mucilage) inversely related to mucilage yield and seed quality. **d** The ratio of arabinose to xylose residues in extracted mucilage (an approximation of heteroxylan branching) remains unchanged indicating that seed and mucilage development were unperturbed. **e** Seed grown at trial sites in 2017 is unevenly coloured, with many blackened seeds while, **f** seeds grown at the same site in 2018 are more consistent, with the light-coloured husk material consistently visible on seeds. **g** In the 2017 field trials (analysed are samples from 2017a), significant quantities of mannose were identified by monosaccharide analysis in extracted mucilage of each variety grown indicative of seed damage-related endosperm leakage. Contrastingly, only trace amounts (well below the limit of quantitation) were found in the same varieties grown in the following year, 2018. Dotted line in **a**–**d** indicates the value of a quality control sample (QC). Error bars represent one standard deviation (some are small and not easily visible)

et al. unpublished data). As the functional component of psyllium gum, heteroxylan must be abundant and present at a consistent level to be considered good quality for industrial uses as the dilution of heteroxylan by the presence of contaminants will impact the functionality in optimised formulations. Four varieties of *P. ovata* were grown in three separate field trials with different times of sowing a factor which, due to climatic conditions, was found to significantly affect seed quality [55, 56]. A suite of quality parameters is shown in Fig. 5. Visual inspection was used to determine a baseline quality score for the four varieties at each trial (Fig. 5a). Trial 2017a had the lowest damage score followed by 2017b and then 2017c. 2017c was deemed the poorest quality with consistently high seed damage. When mucilage was extracted using our pipeline, there were clear differences in yield, with the strongest effect related to time of sowing with only minor intervarietal influence (Fig. 5b). Varieties grown at 2017b did not differ greatly in yield from the control, which was a high-quality field grown sample (QC). Conversely, varieties from 2017a and 2017c had higher mass yields after extraction. Monosaccharide profiling showed that mucilage synthesis was not disrupted as the arabinose to xylose (AX) ratio was very similar between varieties and trials and not significantly different from the QC (Fig. 5d). However, the quantity of heteroxylan in the mucilage (defined as total AX) differed between trials (Fig. 5c). 2017b, confirmed as the most successful sample, had the highest proportion of AX and was closest to the

QC. Correspondingly, 2017a and 2017c had lower proportions of AX indicating significant contamination from other components. Total AX is thereby inversely proportional to yield as a direct result of quality. No variety at any trial had AX as high as the QC, likely because seed of the QC sample was of exceptional quality.

Furthermore, known chemical markers have been defined indicating low quality or damaged seed. As one example, extreme damage of *P. ovata* seeds causes extensive leakage of endosperm components including mannose monosaccharides (Cowley et al. unpublished data). Trial 2017a was impacted by devastating unseasonable rainfall which physically damaged seed before harvesting (Fig. 5e) and subsequently led to microbial growth, leading to detectable quantities of mannose in the extracted mucilage (Fig. 5g). Mannose was found only in trace amounts in corresponding 2018 samples which were not weather damaged or microbially contaminated and more consistently high quality (Fig. 5f).

The pipeline can be used for rapid screening of myxospermous germplasm

This pipeline has utility for rapid screening of mucilage yield traits in a germplasm set, demonstrated here using gamma-irradiated *P. ovata* mutants generated in a previous study [10]. Total mucilage yield data (a pooling of CWE, HWE and IAE fractions) was obtained for a subset of 206 randomly-selected glasshouse-grown *P. ovata* mutants (Fig. 6a). In 63% of mutants, mucilage yield was within a ±10% interval of WT yield (n = 131). Only 4% of mutants yielded 10% less mucilage than WT (n = 8), while 33% yielded over 10% more (n = 68). To validate this screen, a subset of the three lowest yielding (252-7, 768-9, and 1064-5) and three highest (743-4, 1072-12, and 776-5) mutants were selected for further analysis. Expanded mucilage architecture has been used previously to visually screen for altered mucilage phenotypes in mutants of *Arabidopsis* [4, 57, 58] and *Plantago* [10]. Here we show variation in expanded mucilage architecture between the mutants (Fig. 6b) where some are distinctly different to WT (252-7, 1064-5, 743-4, and 776-5) while others are WT-like (768-9 and 1072-12). The utility of the pipeline is proven two-fold in that it can identify highly-distinctive mutants which would be identified through typical visual screening techniques (like ruthenium red staining) but also mutants with more subtle changes to yield that may appear as WT. The validation set of mutant lines was subjected to further analysis which confirmed the differences observed in total mucilage yield were statistically significant compared to the WT. Differences observed in the size of the ruthenium red-stained mucilage envelopes and the amount of mucilage extracted may be linked to alterations in polysaccharide macromolecular properties. This was examined further by comparing the relative proportion of the three mucilage fractions (Fig. 6d) with the water absorption capacity of the mucilage (Fig. 6e).

While the total yield of mucilage was significantly decreased from WT in mutant 768-9, the ruthenium red phenotype, the ratio of mucilage fractions and the water absorption capacity were not significantly different from the WT. Contrastingly, the ratio of the three mucilage fractions was significantly altered in mutants 252-7, 1064-5, and 743-4 (*mucilage extractable with water* (*mew*) mutants) where the CWE and HWE fractions comprise most or all of the mucilage and the IAE fraction is significantly diminished or totally absent. In *mew* mutants, water absorption capacity is significantly reduced from the WT presumably due to a reduction in the stronger gelling, high water-holding capacity IAE mucilage fractions [26]. The striking similarities in the phenotypes of the *mew* mutants suggests that they may contain mutant alleles. Importantly, *mew* mutant 252-7 has already been identified as a putative reduced mucilage xylan mutant [10] and the characterisation of this class of mutant is ongoing [11]. Mutant 776-5 represents a previously unseen class of *P. ovata* mucilage mutant [10, 11, 59]. While this mutant has the highest total mucilage yield in the screened population, its water absorption capacity was unchanged from WT. Its unique compact ruthenium red phenotype and shift in the ratio of the three mucilage fractions suggests intrinsically different changes to the mucilage composition, with a novel causative mutation(s) compared to the *mew* mutants. The ease of distinguishing the *mew* mutants and mutant 776-5 within the mutant population shows that the pipeline can effectively identify putative mutants with perturbed seed development and/or mucilage synthesis, ideal for forward genetic studies.

In contrast to the *mew* mutants and mutant 776-5, it was found that while the ruthenium red and mucilage fractionation phenotype of mutant 1072-12 did not differ substantially from the WT, the total yield and related water absorption capacity was significantly increased. Mutant 1071-12 may therefore represent an important genotype for use in pre-breeding efforts due to its high mucilage yield without the aberrant changes to mucilage composition and properties which makes other mutants less suitable.

Conclusions

In this study we tested the efficacy of several established protocols for seed mucilage extraction and downsized and adapted the most effective elements into a small-scale, rapid extraction and analysis pipeline. We demonstrated the utility of this pipeline for investigating intergeneric and interspecific differences in seed mucilage characteristics, as well as for quality testing and germplasm screening of myxospermous plants. This

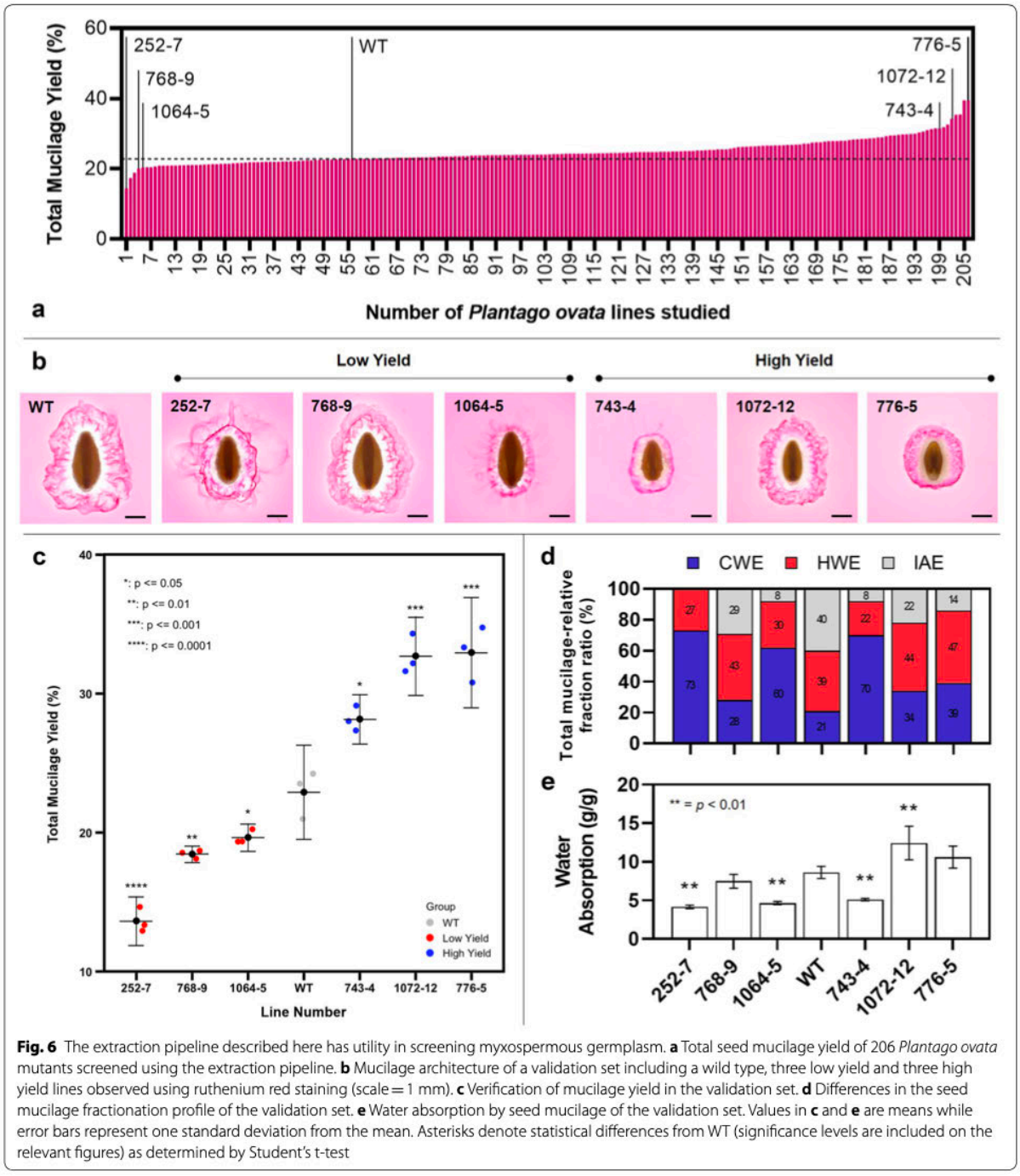


Fig. 6 The extraction pipeline described here has utility in screening myxospermous germplasm. **a** Total seed mucilage yield of 206 *Plantago ovata* mutants screened using the extraction pipeline. **b** Mucilage architecture of a validation set including a wild type, three low yield and three high yield lines observed using ruthenium red staining (scale = 1 mm). **c** Verification of mucilage yield in the validation set. **d** Differences in the seed mucilage fractionation profile of the validation set. **e** Water absorption by seed mucilage of the validation set. Values in **c** and **e** are means while error bars represent one standard deviation from the mean. Asterisks denote statistical differences from WT (significance levels are included on the relevant figures) as determined by Student's t-test

pipeline is already regularly used in our research group increasing the analysis efficiency of a range of myxospermous species. It has also been adopted by a leading food manufacturer who relies on consistently high-quality mucilage products. The use of this pipeline in fundamental research may improve our understanding of mucilage production and properties, ensure quality in food manufacturing, and aid in pre-breeding or breeding of myxospermous species—often classified as orphan crops—that could benefit from improved characterisation methods.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13007-020-00569-6.

Additional file 1: Table S1. Monosaccharide summary of *Plantago ovata* mucilage fractionated by the small-scale extraction pipeline in comparison with previously published large scale techniques.

Abbreviations

CWE: Cold water extractable; HWE: Hot water extractable; IAE: Intense agitation extractable; DI water: Deionized water; RG-I: Rhamnogalacturonan I; QC: Quality control; A:x Ratio: Arabinose to xylose ratio; AX: Arabinose + xylose (estimated heteroxylan content); WT: Wild type.

Acknowledgements

The authors thank Dr. Jana Phan and Dr. Tina Bianco-Miotto for support and guidance and Shi Fang (Sandy) Khor for assistance with HPLC. We thank Siva Sivapalan, David McNeil and Mark Warmington from the Frank Wise Institute for Tropical Agriculture, Department of Primary Industries and Regional Development, Kununurra, WA for assistance in production of the field-grown seed samples. We thank Associate Professor Matthew Tucker for his assistance in developing the gamma-irradiated *Plantago ovata* mutant population and Dayton Bird from the Tucker Lab for the kind donation of *Arabidopsis* seeds used in this study. We thank members of the Burton Lab (RABLAB), the ARC Centre of Excellence in Plant Cell Walls and the ARC Centre of Excellence in Plant Energy Biology for useful discussions and support.

Authors' contributions

JMC and RAB conceived the study. JMC designed and tested the method, analysed the data and wrote the manuscript. LH designed and conducted the mutant screen experiment, analysed the data and contributed to writing the manuscript. KAN collected and assisted in data analysis for field trial quality testing. SC and VC provided field trial material for testing the method. All authors edited the final manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the Australian Research Council Centres of Excellence in Plant Cell Walls (Grant No. 110001007) and Plant Energy Biology (Grant No. 140100008). JMC is supported by a PhD scholarship from the Australian Government's Research Training Program. LH is supported by the University of Adelaide's Adelaide Graduate Research Scholarship (AGRS).

Availability of data and materials

The datasets used and analysed during this work are available from the corresponding author upon reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors give consent for the data to be published.

Competing interests

The authors declare no competing interests.

Author details

[1] Australian Research Council Centre of Excellence in Plant Cell Walls, School of Agriculture, Food and Wine, University of Adelaide, Waite Campus, Urrbrae, SA, Australia. [2] Australian Research Council Centre of Excellence in Plant Energy Biology, School of Agriculture, Food and Wine, University of Adelaide, Waite Campus, Urrbrae, SA, Australia. [3] Dr. Schär R&D Centre, AREA Science Park, Padriciano 99, 34149 Trieste, Italy.

Received: 10 January 2020 Accepted: 14 February 2020
Published online: 24 February 2020

References

1. North HM, Berger A, Saez-Aguayo S, Ralet MC. Understanding polysaccharide production and properties using seed coat mutants: future perspectives for the exploitation of natural variants. Ann Bot. 2014;114(6):1251–63.
2. Western TL. The sticky tale of seed coat mucilages: production, genetics, and role in seed germination and dispersal. Seed Sci Res. 2012;22(01):1–25.
3. Arsovski AA, Haughn GW, Western TL. Seed coat mucilage cells of *Arabidopsis thaliana* as a model for plant cell wall research. Plant Signal Behav. 2010;5(7):796–801.
4. Voiniciuc C, Schmidt MHW, Berger A, Yang B, Ebert B, Scheller HV, et al. MUCILAGE-RELATED10 produces galactoglucomannan that maintains pectin and cellulose architecture in *Arabidopsis* seed mucilage. Plant Physiol. 2015;169(1):403–20.
5. Arsovski AA, Popma TM, Haughn GW, Carpita NC, McCann MC, Western TL. AtBXL1 encodes a bifunctional β-D-xylosidase/α-l-arabinofuranosidase required for pectic arabinan modification in *Arabidopsis* mucilage secretory cells. Plant Physiol. 2009;150(3):1219–34.
6. Western TL. Isolation and characterization of mutants defective in seed coat mucilage secretory cell development in *Arabidopsis*. Plant Physiol. 2001;127(3):998–1011.
7. Hu R, Li J, Yang X, Zhao X, Wang X, Tang Q, et al. Irregular xylem 7 (IRX7) is required for anchoring seed coat mucilage in *Arabidopsis*. Plant Mol Biol. 2016;92(1–2):25–38.
8. Ralet M-C, Crépeau M-J, Vigouroux J, Tran J, Berger A, Sallé C, et al. Xylans provide the structural driving force for mucilage adhesion to the *Arabidopsis* seed coat. Plant Physiol. 2016;171(1):165–78.
9. Jensen JK, Kim H, Cocuron J-C, Orler R, Ralph J, Wilkerson CG. The DUF579 domain containing proteins IRX15 and IRX15-L affect xylan synthesis in *Arabidopsis*. Plant J. 2011;66(3):387–400.
10. Tucker MR, Ma C, Phan J, Neumann K, Shirley NJ, Hahn MG, et al. Dissecting the genetic basis for seed coat mucilage heteroxylan biosynthesis in *Plantago ovata* using gamma irradiation and infrared spectroscopy. Front Plant Sci. 2017;8(March):326.
11. Phan J. Using Plantago ovata as a proxy to study plant cell wall polysaccharide biosynthesis. Adelaide: The University of Adelaide; 2018.
12. Venglat P, Xiang D, Qiu S, Stone SL, Tibiche C, Cram D, et al. Gene expression analysis of flax seed development. BMC Plant Biol. 2011;11(April):74.
13. Renouard S, Cyrielle C, Lopez T, Lamblin F, Lainé E, Hano C. Isolation of nuclear proteins from flax (*Linum usitatissimum* L.) seed coats for gene expression regulation studies. BMC Res Notes. 2012;5:1–7.
14. Soto-Cerda BJ, Maureira-Butler I, Muñoz G, Rupayan A, Cloutier S. SSR-based population structure, molecular diversity and linkage disequilibrium analysis of a collection of flax (*Linum usitatissimum* L.) varying for mucilage seed-coat content. Mol Breed. 2012;30(2):875–88.
15. Kumar RK, Bejkar M, Du S, Serventi L. Flax and wattle seed powders enhance volume and softness of gluten-free bread. Food Sci Technol Int. 2018. https://doi.org/10.1177/1082013218795808.
16. Haque A, Morris ER. Combined use of ispaghula and HPMC to replace or augment gluten in breadmaking. Food Res Int. 1994;27(4):379–93.
17. Mariotti M, Lucisano M, Ambrogina Pagani M, Ng PKWP, Pagani M, Ng PKWP. The role of corn starch, amaranth flour, pea isolate, and psyllium flour on the rheological properties and the ultrastructure of gluten-free doughs. Food Res Int. 2009;42(8):963–75.

18. Cappa C, Lucisano M, Mariotti M. Influence of psyllium, sugar beet fibre and water on gluten-free dough properties and bread quality. Carbohydr Polym. 2013;98(2):1657–66.
19. Anderson JW, Zettwoch N, Feldman T, Tietyen Clark J, Oeltgen P, Bishop CW. Cholesterol-lowering effects of psyllium hydrophilic mucilloid for hypercholesterolemic men. Arch Intern Med. 1988;148(2):292–6.
20. Gunness P, Gidley MJ. Mechanisms underlying the cholesterol-lowering properties of soluble dietary fibre polysaccharides. Food Funct. 2010;1(2):149.
21. Prasad K. Dietary flax seed in prevention of hypercholesterolemic atherosclerosis. Atherosclerosis. 1997;132(1):69–766.
22. Sharma PK, Koul AK. Mucilage in seeds of *Plantago ovata* and its wild allies. J Ethnopharmacol. 1986;17(3):289–95.
23. Balke DT, Diosady LL. Rapid aqueous extraction of mucilage from whole white mustard seed. Food Res Int. 2000;33(5):347–56.
24. Phan JL, Burton RA. New insights into the composition and structure of seed mucilage. Annu Plant Rev Online. 2018;1:1–41.
25. Haughn GW, Western TL. *Arabidopsis* seed coat mucilage is a specialized cell wall that can be used as a model for genetic analysis of plant cell wall structure and function. Front Plant Sci. 2012;3:64.
26. Yu L, Yakubov GGE, Zeng W, Xing X, Stenson J, Bulone V, et al. Multi-layer mucilage of Plantago ovata seeds: rheological differences arise from variations in arabinoxylan side chains. Carbohydr Polym. 2017;165:132–41.
27. Yu L, Yakubov GE, Gilbert EP, Sewell K, van de Meene AML, Stokes JR. Multi-scale assembly of hydrogels formed by highly branched arabinoxylans from *Plantago ovata* seed mucilage studied by USANS/SANS and rheology. Carbohydr Polym. 2019;207(December 2018):333–42.
28. Yu L, Yakubov GE, Martínez-Sanz M, Gilbert EP, Stokes JR. Rheological and structural properties of complex arabinoxylans from *Plantago ovata* seed mucilage under non-gelled conditions. Carbohydr Polym. 2018;193(March):179–88.
29. Tucker MR, Okada T, Hu Y, Scholefield A, Taylor JM, Koltunow AMG. Somatic small RNA pathways promote the mitotic events of megagametogenesis during female reproductive development in *Arabidopsis*. Development. 2012;139(8):1399–404.
30. Phan JL, Tucker MR, Khor SF, Shirley NJ, Lahnstein J, Beahan C, et al. Differences in glycosyltransferase family 61 accompany variation in seed coat mucilage composition in *Plantago* spp. J Exp Bot. 2016;67(22):6481–95.
31. Comino P, Shelat K, Collins H, Lahnstein J, Gidley MJ. Separation and purification of soluble polymers and cell wall fractions from wheat, rye and hull less barley endosperm flours for structure-nutrition studies. J Agric Food Chem. 2013;61(49):12111–22.
32. Hassan AS, Houston K, Lahnstein J, Shirley N, Schwerdt JG, Gidley MJ, et al. A genome wide association study of arabinoxylan content in 2-row spring barley grain. PLoS ONE. 2017;12(8):1–19.
33. Wood J, Tan H-T, Collins H, Yap K, Khor S, Lim W, et al. Genetic and environmental factors contribute to variation in cell wall composition in mature desi chickpea (*Cicer arietinum* L.) cotyledons. Plant Cell Environ. 2018;41(November 2017):2195–208.
34. Western TL, Skinner DJ, Haughn GW. Differentiation of mucilage secretory cells of the *Arabidopsis* seed coat. Plant Physiol. 2000;122(2):345–56.
35. Guo Q, Cui SW, Wang Q, Christopher YJ. Fractionation and physicochemical characterization of psyllium gum. Carbohydr Polym. 2008;73(1):35–433.
36. Marlett JA, Fischer MH. Nutrient metabolism a poorly fermented gel from psyllium seed husk increases excreta moisture and bile acid excretion in rats. J Nutr. 2002;132(April 2002):2638–43.
37. Macquet A, Ralet MC, Kronenberger J, Marion-Poll A, North HM. In situ, chemical and macromolecular study of the composition of *Arabidopsis thaliana* seed coat mucilage. Plant Cell Physiol. 2007;48(7):984–99.
38. Voiniciuc C. Quantification of the mucilage detachment from *Arabidopsis* seeds. Bio-protocol. 2016;6:1–9.
39. Mazza G, Biliaderis CG. Functional properties of flax seed mucilage. J Food Sci. 1989;54(5):1302–5.
40. Oomah BD, Kenaschuk EO, Cui W, Mazza G. Variation in the composition of water-soluble polysaccharides in flaxseed. J Agric Food Chem. 1995;43(6):1484–8.
41. Capitani MI, Corzo-Rios LJ, Chel-Guerrero LA, Betancur-Ancona DA, Nolasco SM, Tomás MC. Rheological properties of aqueous dispersions of chia (*Salvia hispanica* L.) mucilage. J Food Eng. 2015;149:70–7.
42. Fernandes SS, de las Mercedes Salas-Mellado M. Addition of chia seed mucilage for reduction of fat content in bread and cakes. Food Chem. 2017;227:237–44.
43. Muñoz LA, Cobos A, Diaz O, Aguilera JM. Chia seeds: microstructure, mucilage extraction and hydration. J Food Eng. 2012;108(1):216–24.
44. Segura-Campos M, Acosta-Chi Z, Rosado-Rubio G, Chel-Guerrero L, Betancur-Ancona D. Whole and crushed nutlets of chia (*Salvia hispanica*) from Mexico as a source of functional gums. Food Sci Technol. 2014;34(4):701–9.
45. Behbahani BA, Tabatabaei Yazdi F, Shahidi F, Hesarinejad MA, Mortazavi SA, Mohebbi M. Plantago major seed mucilage: optimization of extraction and some physicochemical and rheological aspects. Carbohydr Polym. 2017;155:68–77.
46. Benaoun F, Delattre C, Boual Z, Ursu AV, Vial C, Gardarin C, et al. Structural characterization and rheological behavior of a heteroxylan extracted from *Plantago notata* Lagasca (Plantaginaceae) seeds. Carbohydr Polym. 2017;175:96–104.
47. Naran R, Chen G, Carpita NC. Novel rhamnogalacturonan I and arabinoxylan polysaccharides of flax seed mucilage. Plant Physiol. 2008;148(1):132–41.
48. Pavlov A, Paynel F, Rihouey C, Porokhovinova E, Brutch N, Morvan C. Variability of seed traits and properties of soluble mucilages in lines of the flax genetic collection of Vavilov Institute. Plant Physiol Biochem. 2014;80:348–61.
49. Voiniciuc C, Gunl M. Analysis of monosaccharides in total mucilage extractable from *Arabidopsis* seeds. Bio-protocol. 2016;6:1–11.
50. Zhao X, Qiao L, Wu A-M. Effective extraction of *Arabidopsis* adherent seed mucilage by ultrasonic treatment. Sci Rep. 2017;7:40672.
51. Hu R, Li J, Wang X, Zhao X, Yang X, Tang Q, et al. Xylan synthesized by irregular xylem 14 (IRX14) maintains the structure of seed coat mucilage in *Arabidopsis*. J Exp Bot. 2016;67(5):1243–57.
52. Yu L, Shi D, Li J, Kong Y, Yu Y, Chai G, et al. CELLULOSE SYNTHASE-LIKE A2, a glucomannan synthase, is involved in maintaining adherent mucilage structure in *Arabidopsis* seed. Plant Physiol. 2014;164(4):1842–56.
53. Griffiths JS, Tsai AY, Xue H, Voiniciuc C, Sola K, Seifert GJ, et al. SALT-OVERLY SENSITIVE5 mediates *Arabidopsis* seed coat mucilage adherence and organization through pectins. Plant Physiol. 2014;165(July):991–1004.
54. Lin KY, Daniel JR, Whistler RL. Structure of chia seed polysaccharide exudate. Carbohydr Polym. 1994;23(1):13–8.
55. Kumar J. Good agricultural practices for isabgol. Report for the Directorate of Medicinal and Aromatic Plants; 2015.
56. McNeil DL. Growers' manual for production of *Plantago ovata* in the Ord irrigation area. ISBN Services; 2017. p. 75.
57. Macquet A, Ralet M-C, Loudet O, Kronenberger J, Mouille G, Marion-Poll A, et al. A naturally occurring mutation in an *Arabidopsis* accession affects a β-D-galactosidase that increases the hydrophilic potential of rhamnogalacturonan I in seed mucilage. Plant Cell Online. 2007;19(12):3990–4006.
58. Sullivan S, Ralet M-C, Berger A, Diatloff E, Bischoff V, Gonneau M, et al. CESA5 is required for the synthesis of cellulose with a role in structuring the adherent mucilage of *Arabidopsis* seeds. Plant Physiol. 2011;156(4):1725–39.
59. Cowley J. Analysis of Plantago mucilage mutants. Adelaide: University of Adelaide; 2016.

Publisher's Note

Appendix II

The novel features of *Plantago ovata* seed mucilage accumulation, storage and release

ꕥ•ꕥ

Appendix II – A publication

Statement of Authorship

Title of Paper	The novel features of *Plantago ovata* seed mucilage accumulation, storage and release
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Phan, J.L., Cowley, J.M, Neumann, K.A., *et al.* (2020) The novel features of *Plantago ovata* seed mucilage accumulation, storage and release. *Scientific Reports* https://doi.org/10.1038/s41598-020-68685-w

Candidate

Name of Candidate	James M. Cowley		
Contribution to the Paper	Performed compositional analyses, interpreted compositional and microscopy data, performed data analysis and formulated the model, contributed to writing the manuscript		
Overall percentage (%)	30%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am a primary author of this paper.		
Signature		Date	19/5/2020

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);
ii. permission is granted for the candidate in include the publication in the thesis; and
iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Principal Author	Jana Phan		
Contribution to the Paper	Conceived the study, performed experiments, wrote the manuscript.		
Signature		Date	9/6/2020

Name of Co-Author	Kylie A. Neumann		
Contribution to the Paper	Performed microscopy and assisted in compositional analyses		
Signature		Date	9/6/2020

Name of Co-Author	Lina Herliana		
Contribution to the Paper	Staged plants for developmental series		
Signature		Date	9/6/2020

Name of Co-Author	Lisa A. O'Donovan		
Contribution to the Paper	Performed sectioning and developmental microscopy		
Signature		Date	9/6/2020

Name of Co-Author	Rachel A. Burton		
Contribution to the Paper	Conceived the study and contributed to writing the manuscript		
Signature		Date	9/6/2020

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

www.nature.com/scientificreports

SCIENTIFIC REPORTS
nature research

OPEN

The novel features of *Plantago ovata* seed mucilage accumulation, storage and release

Jana L. Phan[1,3,5], James M. Cowley[1,2,5], Kylie A. Neumann[1,2,4], Lina Herliana[2], Lisa A. O'Donovan[2] & Rachel A. Burton[1,2]✉

Seed mucilage polysaccharide production, storage and release in *Plantago ovata* is strikingly different to that of the model plant *Arabidopsis*. We have used microscopy techniques to track the development of mucilage secretory cells and demonstrate that mature *P. ovata* seeds do not have an outer intact cell layer within which the polysaccharides surround internal columellae. Instead, dehydrated mucilage is spread in a thin homogenous layer over the entire seed surface and upon wetting expands directly outwards, away from the seed. Observing mucilage expansion in real time combined with compositional analysis allowed mucilage layer definition and the roles they play in mucilage release and architecture upon hydration to be explored. The first emergent layer of hydrated mucilage is rich in pectin, extremely hydrophilic, and forms an expansion front that functions to 'jumpstart' hydration and swelling of the second layer. This next layer, comprising the bulk of the expanded seed mucilage, is predominantly composed of heteroxylan and appears to provide much of the structural integrity. Our results indicate that the synthesis, deposition, desiccation, and final storage position of mucilage polysaccharides must be carefully orchestrated, although many of these processes are not yet fully defined and vary widely between myxospermous plant species.

Abbreviations

DPA	Days post-anthesis
ML	Mucilage layer
MSC	Mucilage secretory cell
SEM	Scanning electron microscopy

Upon exposure to aqueous environments, seeds from myxospermous species extrude a polysaccharide-rich gel from their seed surface, often called mucilage. Numerous species display myxospermy and there are a range of possible evolutionary advantages of synthesising such a carbon-rich and energy-expensive substance[1]. Of all myxospermous species, the seed mucilage system of *Arabidopsis* is the best characterised. *Arabidopsis* seed mucilage has been used extensively as a proxy for the study of plant cell wall polysaccharide biosynthesis, enabling increased molecular characterisation of pectin biosynthesis, its main polysaccharide component[2], as well as the biosynthesis of cellulose[3,4] and several hemicelluloses[5–8], which are minor but integral components. Mucilage from other species can be highly diverse[1] and while *P. ovata* mucilage is also a complex mixture of polymers, it is predominantly heteroxylan with only a minor pectin component. While the pectin component is a near-linear rhamnogalacturonan[9–11], the *P. ovata* heteroxylan (accounting for around 90% of the mucilage polysaccharides) is highly complex with the current scientific consensus defining *P. ovata* heteroxylan comprising a β-(1,4)-linked-D-xylopyranose backbone, heavily substituted at O-2 and/or O-3 positions with various mono-, di- and oligosaccharide substitutions of α-L-arabinofuranose and β-D-xylopyranose[9,11,12]. It is likely that, as with other eudicots, the β-(1,4)-linked-D-xylopyranose backbone is synthesised by several members of

[1]Australian Research Council Centre of Excellence in Plant Cell Walls, School of Agriculture, Food and Wine, University of Adelaide, Waite Campus, Urrbrae, SA 5064, Australia. [2]Australian Research Council Centre of Excellence in Plant Energy Biology, School of Agriculture, Food and Wine, University of Adelaide, Waite Campus, Urrbrae, SA 5064, Australia. [3]Present address: Australian Academy of Science, Ian Potter House, 9 Gordon St, Canberra, ACT 2601, Australia. [4]Present address: IP Australia, PO Box 200, Woden, ACT 2606, Australia. [5]These authors contributed equally: Jana L. Phan and James M. Cowley. ✉email: rachel.burton@adelaide.edu.au

glycosyltransferase (GT) families 43 and 47. There is strong evidence that GT47 protein IRX10-L, probably in concert with other GT43 and GT47 proteins, extends the backbone by adding UDP-xylose moieties[14,15]. GT61 family members have been implicated in both α-arabinosyltransferase and β-xylosyltransferase xylan backbone decoration activities in cereals[16,17], *Arabidopsis*[18,19] and *Plantago*[20], and copy number and type of GT61 was found to influence interspecific differences in *Plantago* heteroxylan fine structure[11]. The overall picture is complicated even further in that different fractions (sometimes described as layers) of *P. ovata* mucilage contain heteroxylans of varying substitution patterns showing that, like *Arabidopsis* mucilage, it is a similarly complex but orchestrated network of polysaccharides[9,10,13]. To date, many xylan synthase genes, particularly those involved in backbone decoration, are still unknown.

Plantago ovata mucilage also has economic relevance, in that in its dry state it constitutes the basis of a dietary fibre supplement, called psyllium, that is widely consumed by humans to assist with laxation, relieving constipation[21,22], and to treat metabolic disorders like hypercholesterolaemia[23]. More recently, psyllium has become a key ingredient in gluten-free food, where it provides texture and structure in the absence of gluten[24–27]. Psyllium is produced by milling the dry polysaccharides off the seed surface[28] and is often referred to as the "husk" fraction. The ratio of husk to seed is approximately 1:3, with the discarded non-husk seed components often being used for animal or fish feed[29]. From an economic standpoint, the ability to understand mucilage polysaccharide production and therefore potentially increase the valuable husk fraction is a viable breeding target for this plant species.

As well as studying the biosynthesis of mucilaginous polymers, the mechanism of mucilage extrusion from the seed coat of *Arabidopsis* has also been thoroughly characterised. In *Arabidopsis*, seed mucilage polysaccharides accumulate in the apoplast of specialised seed coat cells called 'mucilage secretory cells' (MSCs). When the mucilage polysaccharides become hydrated, they swell and rupture the primary cell wall of the MSC, releasing the mucilage[30]. The MSCs in *Arabidopsis* differentiate from the outer-most integument cell layer of the ovule. *Arabidopsis* has an outer integument composed of two cell layers and an inner integument composed of three cell layers, both of maternal origin, which grow to surround the mature ovule[31]. After pollination, at approximately 7 days post-anthesis (DPA), starch granules begin to accumulate in the MSCs and polysaccharide deposition starts in the peripheral "corners" of the cells. This pushes the protoplasm to form a central volcano-like structure in the cell. At 10 DPA this central column is reinforced by the deposition of secondary cell wall polysaccharides to form the columella. The columella is a prominent feature of the mature MSCs in *Arabidopsis* and the accumulated polysaccharides are deposited and stored around it, producing a doughnut-shaped ring. This structure results in the distinctive mature *Arabidopsis* seed coat patterning seen using SEM[32–34]. The details of MSC development, rupture and mucilage release are discussed in comprehensive reviews by Francoz et al.[30], and Voiniciuc et al.[35]. An important developmental stage during MSC development is the weakening of the radial primary cell walls of the MSCs at the end of columella formation, at approximately 13 DPA. This process enables the consequent fracturing and rupturing of the cell walls of the MSCs upon imbibition in an aqueous environment[32]. The rupturing allows the accumulated seed polysaccharides to be extruded almost instantaneously forming the distinctive mucilage envelope. Thus, MSCs of *Arabidopsis* are a highly-specialised seed coat cell with a clearly defined structure that is essential for correct seed mucilage extrusion. MSC development and mucilage release of *Linum usitatissimum* seeds, more commonly known as flax, has also recently been described, revealing an even more complex MSC structure[36]. The flax MSCs, embedded in the external surface of the seed coat were determined by Miart et al.[36], to contain four discrete, laminated layers in the apoplast, each containing chemically- and functionally-distinct polysaccharides. Each of the layers and their polysaccharide contents act in concert to effectively hydrate the polysaccharides, mechanically forcing the radial cell wall to rupture in a peeling fashion and enabling mucilage to be released. In other species such as *Salvia hispanica* (chia) and *Coleus blumei*, the seed mucilage polysaccharides are stored in the outer epidermal cell layer(s) of a nutlet that encases the true seed within[37,38], making these species myxocarpous rather than myxospermous. The events leading to the release of mucilage in these species have not been documented in detail but there appears to be great diversity in the mucilage extrusion structures between plant types[1].

The accumulation of seed mucilage polysaccharides in *P. ovata* has been investigated previously[39] and appears to be distinct from the process observed in the *Arabidopsis* MSCs. In the case of *P. ovata*, seed mucilage polysaccharides are deposited in the outer-most cell layer of a large integument. This single cell layer accumulates polysaccharides rapidly and the cells expand dramatically in size in a process that does not involve the formation of a central columella[39]. Beyond this, little is known about the precise development of these cells and so here we provide a detailed characterisation of the polysaccharide deposition and mucilage release processes of *P. ovata*, also enabling the formulation of a supporting model.

Results

Development of *P. ovata* mucilage secretory cells. *P. ovata* takes approximately 3.5 months to grow from germination through to maturity. The mature plants have long slender, straggly leaves and produce many spike-type inflorescences (SI Fig. S1). Development of *P. ovata* fruit on the spike and length of the inflorescence (and consequently yield per plant) are strongly dependent on the plant's health during growth and development. Each fruit or capsule contains two ovules, separated by a maternal disc and joined to the parent plant via a placenta (Fig. 1). *P. ovata* possesses a circumscissile capsule (also called a pyxis) that is firmly attached to the inflorescence at the proximal end of the fruit. When the fruit is mature, the seed dispersal mechanism involves dehiscence at the capsule equator causing the operculum to detach, enabling the seed to dislodge from the capsule. The operculum, the point of attachment to the rachis, and the equator are indicated in Fig. 1A. After pollination, the fruits mature in approximately 1 month. At 2 weeks post-anthesis, the fruit has reached its full length and the seeds continue to develop inside, expanding widthways and filling the fruit.

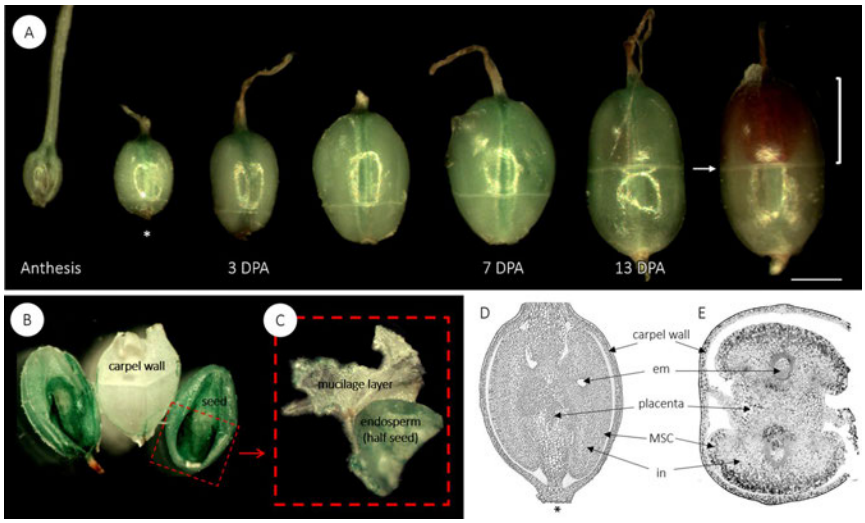


Figure 1. (**A**) The fruit development of *P. ovata*. Each fruit contains two ovules separated by placental tissue. *Plantago* species have a circumscissile capsule, also known as a pyxis. The arrow indicates the equator, where the zone of dehiscence is visible, the square bracket highlights the operculum, which detaches during dehiscence, and *indicates the end that joins the fruit to the rachis. Bar 1 mm. (**B**) A dissected fruit at 13 DPA, showing two immature seeds and in (**C**) one of the seeds has been further dissected to show the mucilage polysaccharide layer, which has been peeled off the seed and is the remnant of the integument tissue. Longitudinal (**D**) and transverse (**E**) cross sections of a developing fruit at 7 DPA, stained with toluidine blue. *Em* embryo sac, *MSC* mucilage secretory cells, *in* integument.

Following successful fertilisation, the parenchyma cells of the integument layers differentiate rapidly (Fig. 2). The MSCs of *P. ovata* seeds are easily observed at 1 DPA. They develop from the outermost single cell layer of the integument and lengthen as they accumulate starch granules (Fig. 2B,C). Substantial growth and elongation of the MSCs is observed from 3 to 5 DPA. Although it is difficult to discern discrete cellular compartments, it is likely that the empty space at the distal end of the cells where the polysaccharides accumulate, is the apoplast (Fig. 2D). By 9 DPA, the accumulated mucilage polysaccharides are hydrophilic enough to rupture the MSCs when they come into contact with aqueous solutions and it is technically challenging to obtain intact sections from this stage onwards. At 15 DPA all MSCs have ruptured and released their mucilage in fixed and sectioned developing seeds but it is possible to observe that the integument has been compressed to just a few cell layers between the MSCs and the seed endosperm. This compressed layer has disappeared almost completely by the time the seed is fully mature, leaving only a thin layer situated between the endosperm and the mucilage polysaccharide layer (Fig. 2H).

Composition of developing ovule cell walls. The cell walls of seed tissues at three key time points across early development were fluorescently immunolabelled with the primary antibodies LM11 (β-1,4-linked xylan backbone[40]) LM19 (un-esterified and partially esterified homogalacturonan[41]) and LM20 (methylesterified homogalacturonan[41]) and the carbohydrate-binding module CBM3a (crystalline cellulose[42]). At anthesis the ovule is minute but there was clear binding by both LM20 (Fig. 3A1–3) and LM19 (SI Fig. S2A), with the latter producing a strong signal in the integument tissue. There was a low level of CBM3a binding to the walls of both MSCs and integument cells (Fig. 3D1–3) and no binding by LM11 (SI Fig. S2B). At 4 DPA, the MSCs are greatly elongated. The strongest signals are generated by LM20 in the MSC layer (Fig. 3B1–3) and CBM3a in both the MSC and integument cells (Fig. 3E1–3) but there was no labelling evident for LM19 (SI Fig. S2C) or LM11 (SI Fig. S2D). The final time point was at 6/7 DPA when the MSCs were becoming fragile due to the accumulation of mucilage polysaccharides. At this point the LM20 labelling was now restricted to the outside edge of the MSC layer and in cell corners bordering the integument tissue (Fig. 3C1–3). The CBM3a signal was still present in both the MSCs and integument though signals in the MSCs had become non-specific and amorphous compared to the integument where labelling of distinct walls was still present (Fig. 3F1–3). By 6/7 DPA there was no labelling by LM19 (SI Fig. S2E) and minimal labelling by LM11 (SI Fig. S2F).

Surface features of mature *P. ovata* seeds. The mature seeds of *P. ovata* have a deep scar on the ventral side resulting in a boat-shaped seed (SI Fig. S3). The patterning of the dry seed surface on the dorsal side is polar. Where the inner surface of the fruit capsule has been pressed against the seed it is smoother (Fig. 4A). The seed

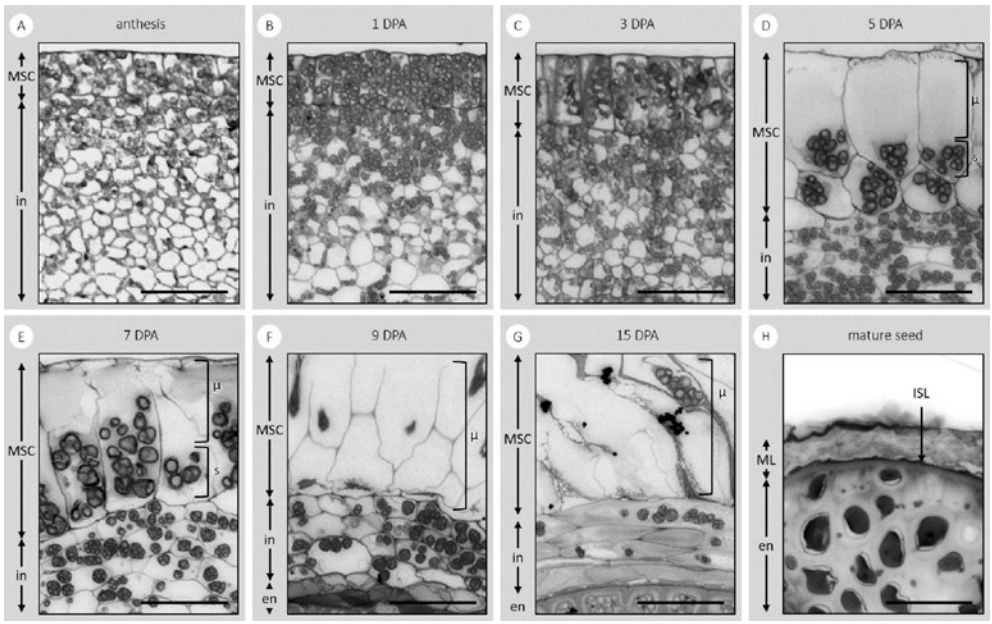


Figure 2. Toluidine blue-stained transverse sections of the developing integument of *P. ovata*. The sections show the tissues and components that are the: endosperm (en); integument (in); mucilage polysaccharides (μ); mucilage secretory cells (MSCs); mucilage polysaccharide layer (ML); intensely stained layer (ISL); and starch granules (s) at days post-anthesis (DPA). Scale bar 50 μm.

surface is covered in hexagonal structures with a distinct wrinkled texture (Fig. 4B). The wrinkled patterning and hexagonal structures on the mature seed surface are lost once seeds have been imbibed in water (Fig. 4D). When the remaining seed mucilage is left to dry back onto the seed after hydration in cold water, the seed surface appears very smooth and high magnification SEM reveals little additional detail (Fig. 4E). The hexagonal structures are no longer visible, and the polarity observed prior to mucilage expansion (Fig. 4A) has also been lost. This is in clear contrast to *Arabidopsis* where, after the same process, the seed surface morphology remains relatively unchanged and the columella is still clearly visible (Fig. 4F). Mature *P. ovata* seeds therefore do not have conventional seed coat cells and there is certainly no columella as found in *Arabidopsis* (Fig. 4C). Rather there is a dehydrated mucilage polysaccharide layer, underlain by a thin dark brown layer (which gives the seed its colour) both of which sit over the outer layer of the endosperm (Fig. 4G). The crushed integument layer is not at all visible in the mature seed.

Mucilage removal from mature *P. ovata* seeds. Different methods were used to remove the expanded mucilage from mature imbibed seeds of *P. ovata*. Although previous studies report no significant compositional differences between the different extraction methods[11], some physical differences were observed on the exposed seed surface (SI Fig. S4). When mature seeds were placed into an aqueous fixative, the sequential washing steps removed most of the mucilage from the seed, leaving a thin resistant layer behind, as did extraction with hot water or 0.2 M KOH (SI Fig. S4). In contrast, when 0.1 M HCl was used for extraction, the mild acid completely removed the mucilage layer from the entire seed (SI Fig. S4D), probably hydrolysing it in situ[10], and in some patches it has also removed the underlying intensely-stained layer (SI Fig. S4F).

Mucilage accumulation may be independent of embryo and endosperm development. From the mutant *P. ovata* population reported in Tucker et al.[43], we selected a line, 69-1, that produces seeds with impaired development across a range of severity: seeds with a thickened translucent outer layer, incomplete endosperm filling, arrested embryo development, or a shrivelled appearance where it was difficult to determine if an embryo was present (Fig. 5A1–5). Ruthenium red staining solution was applied to representative seeds and all types produced mucilage from the seed that was released into the aqueous environment. While different specific architectures were observed, two typical mucilage layers were recognisable in all but the shrivelled phenotype. These seeds may have been aborted early in development rather than representing a developmentally delayed phenotype (Fig. 5B5). While only 5% of seed are WT-like in the 69-1 bulk sample analysed, WT-level total mucilage yields, arabinose and xylan content and ratio were still obtained (Fig. 5C1–3).

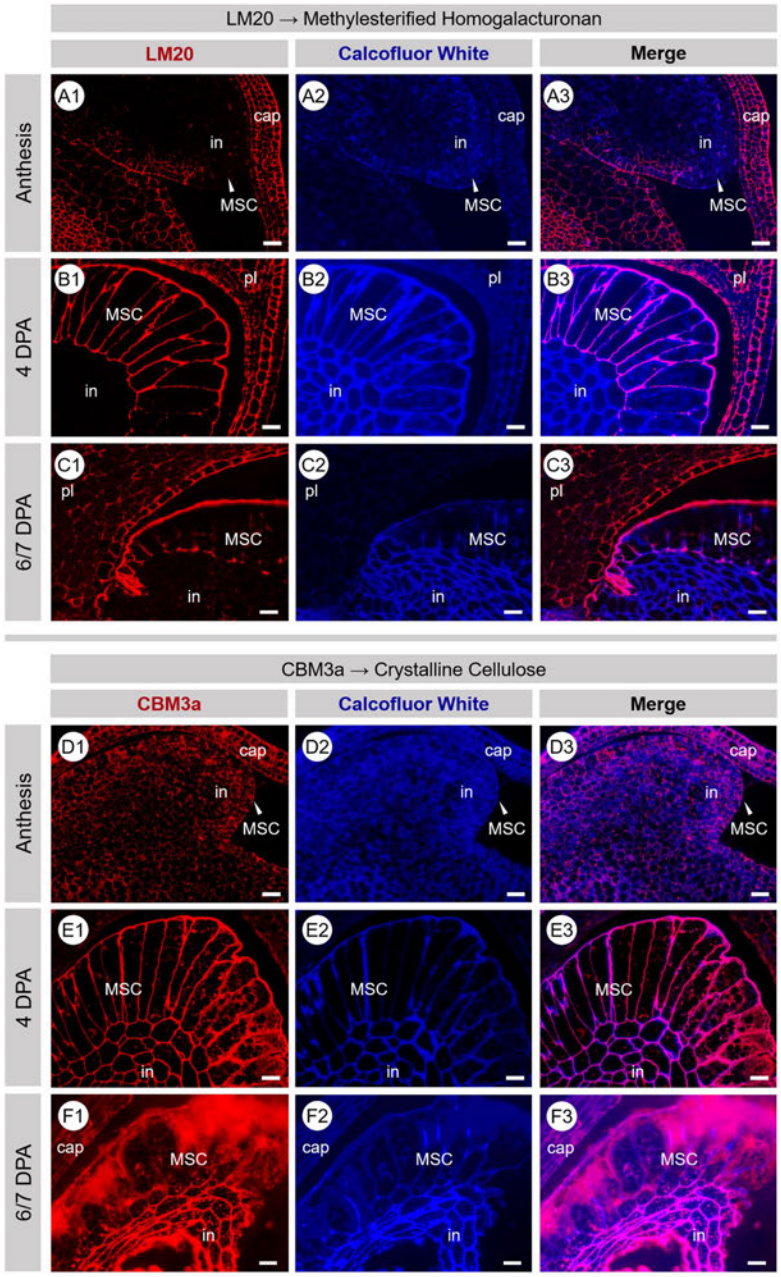


Figure 3. Fluorescence micrographs of transverse sections of developing *P. ovata* seeds labelled with LM20 and CBM3a (red/pink) at anthesis (**A**,**D**), at 4 DPA (**B**,**E**) and at 6/7 DPA (**C**,**F**). MSC cell walls show strong labelling of highly methylesterified HG (LM20) and crystalline cellulose (CBM3a) that diminishes in intensity and organisation as development/mucilage polysaccharide accumulation continues and/or as MSC cell walls disintegrate. Samples are counter-stained with calcofluor white (blue). Scale 20 µm. *DPA* days post-anthesis, *HG* homogalacturonan, *MSC* mucilage secretory cell, *in* integument, *pl* placenta, *cap* capsule.

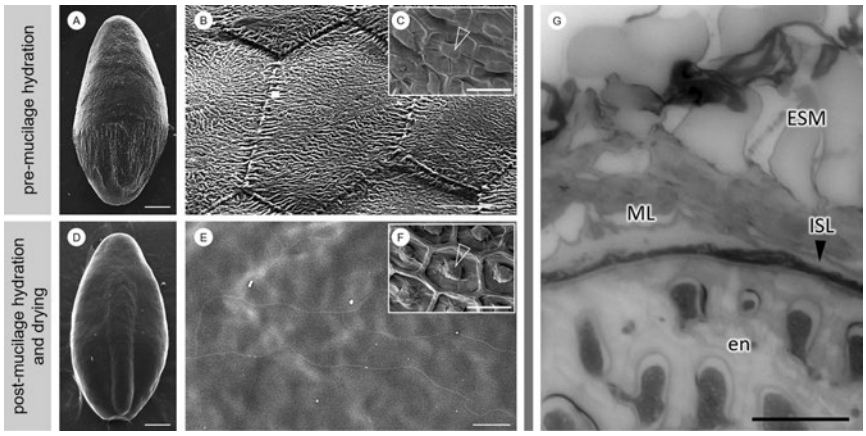


Figure 4. Scanning electron micrographs show that *P. ovata* does not contain a columella (**A**,**B**). Inset (**C**) shows a scanning electron micrograph of the seed surface of *Arabidopsis* with the columella structure indicated with an arrowhead. In *P. ovata*, the wrinkled texture of the dry mucilage polysaccharide layer (ML) and hexagonal shapes of the distal MSC wall remnants disappear after mucilage is hydrated and allowed to dry back onto the seed surface, unfixed (**D**,**E**), leaving it extremely smooth. This contrasts with *Arabidopsis* where the distinct columella structure persists and remains clearly visible after the same process (**F**). Toluidine blue-stained cross sections of the mature seeds fixed in aqueous fixative (**G**) reveal that the seed mucilage (ESM) expands from the ML, which sits on top of an intensely stained layer (ISL) that separates the mucilage polysaccharide layer from the endosperm. *En* endosperm. Scales **A**,**D**,**G** 500 µm; **B**,**E**,**H** 50 µm; **C**,**F** 30 µm.

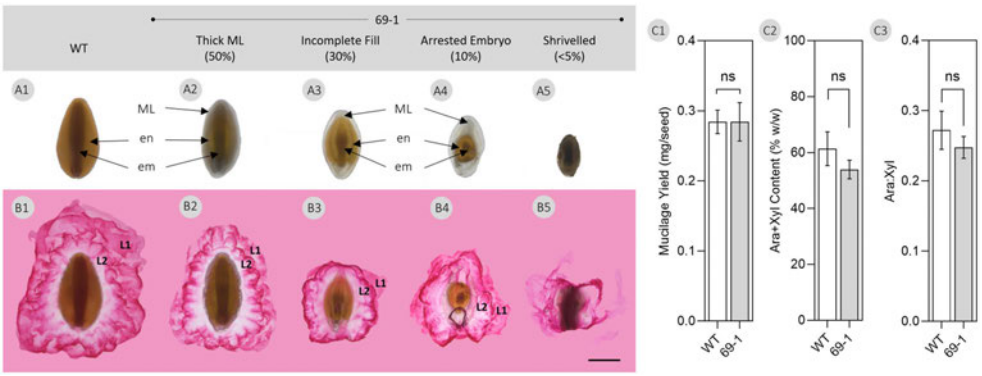


Figure 5. (**A**) Seeds were selected from developmentally-impaired gamma-irradiated *P. ovata* mutant 69-1 generated previously by Tucker et al.[43]. When these seeds were imbibed in a ruthenium red solution (0.01% w/v) for 10 min at room temperature (**B**) mucilage expanded from all seeds with different architectures but two typical mucilage layers (L1 and L2) were present. *em* embryo, *en* endosperm, *ML* mucilage layer, *WT* wild-type. Scale bar 1 mm. (**C**) Analysis of mucilage yield and composition revealed no significant difference (ns) from the wild-type ($p > 0.05$, Student's t test).

Expansion of seed mucilage polysaccharides across time and space. Microscopy and monosaccharide analysis techniques were used to investigate changes in the composition and structure of the mucilage as it expanded from the seed surface. In order to define temporal mucilage expansion we tried to capture and describe the major stages using chemical profiling. By measuring the release of mucilage-related monosaccharides over 14.5 h we found that the major stages of mucilage expansion occurred within 60 min of imbibition (SI Fig. S5). Monosaccharide analysis of serial fractions taken during the first 60 min of mucilage expansion clearly demonstrated a change in the composition of the expanded seed mucilage over time (Fig. 6A). Relative to total extracted sugars, there was a shift from pectin-dominant to heteroxylan-dominant monosaccharide composi-

tion during the initial stages. A sharp increase in the number of heteroxylan-associated monosaccharides (xylose and arabinose) was then observed, peaking at 20 min post imbibition (Fig. 6A) while pectin-derived monosaccharides (rhamnose and galacturonic acid) displayed the inverse, where they were most abundant at the start of seed imbibition and mucilage expansion, before tapering off considerably by 20 min (Fig. 6A).

Mature seeds were stained as described in Yu et al.[9], and real-time mucilage expansion was observed dynamically using confocal microscopy (Fig. 6B). Two distinct layers of expanded mucilage were observed. A third mucilage layer adjacent to the seed coat reported by Yu et al.[9], was not observed. The two mucilage layers, L1 and L2, have distinct structural features. L1 is the first to expand; it lacks clear structure and much of the Calcofluor White staining is associated with this layer. The expansion of L2 from the seed occurs soon after L1 but the sea anemone-like structures (Fig. 6B) are not observed until 2 min post imbibition. By 15 min, L1 has dispersed into the surrounding aqueous environment and by 20 min L2 has expanded in its entirety (Fig. 6B).

Discussion

Mucilage polysaccharide accumulation in the MSCs of *P. ovata* seeds follows a different developmental pattern to that occurring in the MSCs of *Arabidopsis*. The mechanism by which different cell layers are converted into seed tissues and the possible remodelling thereafter appears to be of central importance for MSC development in *P. ovata*. At some point during mid-development and perhaps after polysaccharide accumulation is complete, we propose that the MSC radial cell walls break down and collapse in a concertina-like fashion. The collapse is potentially driven by the outward pressure of the rapidly expanding embryo and endosperm tissues pushing the MSC outer walls against the inner capsule surface, releasing the accumulated mucilage polysaccharides into an amorphous layer that becomes sandwiched between the remnant distal and basal MSC walls (Fig. 2). The process of radial wall remodelling and/or disintegration may already be beginning by 6/7 DPA where discrete labelling present earlier in development is absent or has become non-specific and amorphous (Fig. 3C1–3,F1–3). This is particularly evident for the CBM3a labelling (Fig. 3F1–3). This is unlike the presence of the mucilage polysaccharides contained within intact discrete cells of the *Arabidopsis* (Fig. 4C,F) and flax[36] mature seed coats. From late development onwards it is clear that the MSCs of *P. ovata* do not contain a columella (Fig. 4B) and our hypothesis is that instead, laminated layers of dehydrated mucilage polysaccharides are present, following radial MSC wall disintegration, between the remnant distal MSC walls, inner capsule wall and the expanded endosperm tissue (Figs. 2, 7B). At seed maturity when released from the dehiscent capsule, the dehydrated and highly compressed mucilage polysaccharides, originating from the obliterated MSC cells, (Figs. 2H, 4H, 7C) form a dense layer over the seed surface (Figs. 2C, 6C). Cross sections of the dry mature seed shows that the thickness of this mucilage layer ranges from 10 to 18 µm compared to the 80 to 90 µm thickness of the MSCs at full elongation at 7 DPA (Fig. 2). This supports the compression of the MSCs as the seed matures, after which point the layer is so dense that the constituent polysaccharides do not label with monoclonal antibodies that bind well to the seed mucilage when expanded (Phan et al.[11]; Fig. 2H).

In our SEM analysis of mature seeds, we observe a stark contrast in surface appearance before and after hydration (Fig. 4). After mucilage is hydrated and allowed to dry without fixation, we were no longer able to observe the wrinkled surface or characteristic hexagonal shapes of the underlying distal MSC wall remnants (Fig. 4D–E). We suggest that while the distal walls may have undergone a similar process of remodelling/weakening to the radial walls, they were protected from crushing as they lie perpendicular to the outward force of the expanding endosperm, pushed flat against the inner capsule surface, and thus remain present and visible in the mature seed. However these distal wall fragments are thin, not reinforced with cellulose (Fig. 3F1–3) and appear to be rich in pectin (Fig. 3C1–3) suggesting that they may be highly soluble. We have previously observed 'hexagonal platelets' that stain strongly with Ruthenium Red to be released from *P. ovata* seeds very early in the hydration cascade and rapidly disintegrate or dissolve (Phan et al.[11], Fig. 1H). We suggest that these structures are the soluble remnants of the distal MSC walls. These are different to other cellulose-staining structures like the mucilage discs of the *Arabidopsis* mutant *fly1*[44] or plate cells of other *Plantago* species (Phan et al.[11], Fig. 1I–P; Cowley et al. unpublished data) which persist through mucilage hydration. When hydrated mucilage is left to dry back onto the seed (Fig. 4D,E), the distal MSC wall fragments may have already dissolved/disintegrated and are thus no longer discernible, so the hexagonal shapes are lost, unlike similarly treated *Arabidopsis* seeds which retain the lower portion of the ruptured MSCs (Fig. 4F). The still soft layer of mucilage polysaccharides released by the putative rupture of the *P. ovata* MSC radial walls in later development may also contribute to the polarity of striations on the dorsal side of the mature seed (Fig. 4A). This could occur as the proximal end of the seed is under more compression from the capsule wall, which imprints onto the surface of the polysaccharide layer as it dehydrates. Although the distal end of the seed also comes into contact with the capsule, at this end it does not adhere so tightly and is observed to readily detach when gently touched, thereby enabling seed dispersal. Interestingly, Boesewinkel[45] suggests that the seed coat of *Linum usitassimum* is also polar and that the 'slime-forming matter' is deposited on the outer surface of the epidermal seed coat cells. They also suggest that the cells underneath the outer epidermal seed coat cells, which are on the innermost layer of the inner integument, have thickened cell walls and are pigmented cells. These observations are strikingly similar to what we have observed in *P. ovata*; the mucilage is the outermost layer of the seed and is underlain by an intensely stained layer that may be pigment-rich (Fig. 2H). A similar structure was also described for *P. ovata* by Madgulkar et al.[46]. These observations further support our proposed mechanism of MSC development and disintegration, with the subsequent formation of a cell-free mucilage polysaccharide layer.

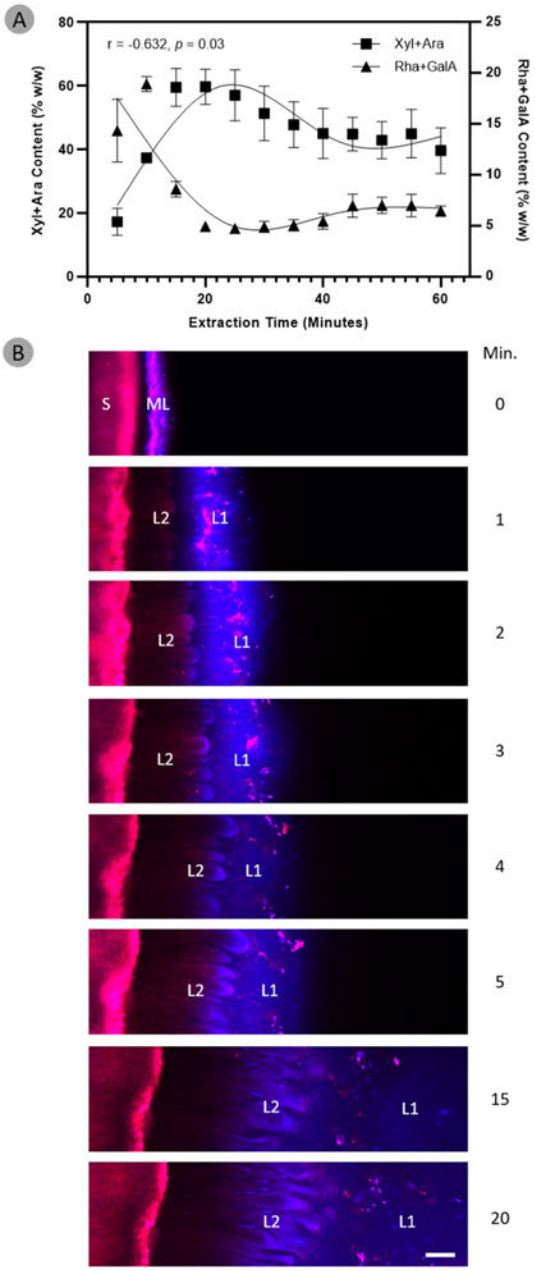
It is interesting to speculate about carbon flow through the *P. ovata* seed during development. There must be a balance between investment in maternal sporophytic tissue and the filial tissues i.e. the carbon supply must be split between mucilage polysaccharide biosynthesis and feeding the rapidly growing embryo and endosperm. It is possible that development of the MSCs from the outermost cell layer of the integument tissue, which is of

Figure 6. The composition and structure of seed mucilage changes over the course of its expansion. (**A**) Compositional analysis reveals pectin-associated monosaccharides (rhamnose and galacturonic acid) are most abundant during the initial expansion of mucilage and rapidly decrease in concentration thereafter. Heteroxylan-associated monosaccharides (xylose and arabinose) are also present in the initial expansion of mucilage, in almost similar amounts to pectin. Contrasting to pectin, the heteroxylan-associated monosaccharides rapidly increase in concentration and go on to make up the bulk of total expanded mucilage. Data have been fitted with cubic spline curves to highlight trends. (**B**) The dynamics of seed mucilage expansion in mature *P. ovata* seeds were observed in real-time over a period of 20 min using confocal microscopy. Seeds were pre-stained with 0.4% Direct Red 23 and 0.1% Calcofluor White. Upon imbibition in water, a sudden "explosion" of an extremely hydrophilic and non-structured mucilage layer emerges (L1). Following L1, a more structured and anemone-like layer of mucilage expands outwards (L2) and by 20 min, L2 has reached its maximal expansion distance and L1 has mostly dissipated into the surrounding aqueous environment. Scale bar 100 µm. *S* mature seed, *ML* mucilage polysaccharide layer on dry seed, *L1* layer 1, *L2* layer 2 N.B. time 0 min is a dry seed that has been pre-stained, water was added after this image was taken.

maternal origin[31], could be favoured over zygotic development. This may explain our observation that MSC development and polysaccharide deposition for mucilage synthesis may occur independently of seed development since even developmentally-stalled and aborted seeds still make mucilage when imbibed (Fig. 5B1–5). While the specific architectures were different, two typical mucilage layers were recognisable and known *P. ovata* quality indicators[10], mucilage yield (Fig. 5C1) and heteroxylan content (Fig. 5C2) and composition (Fig. 5C3) were not significantly different to the wild-type ($p > 0.05$) showing that mucilage synthesis was uninterrupted. Garcia et al.[47], demonstrated that development of the maternally-derived integument and the zygotic embryo and endosperm are coordinated to determine final *Arabidopsis* seed size. Of the various developmentally-impaired *P. ovata* seeds analysed here, none of them reached the same size as the wild-type, suggesting that although integument development and mucilage polysaccharide biosynthesis can occur independently of embryo and endosperm development, some coordination is needed in order to establish the correct size of the mature seed. It is possible that without outward pressure from the growing endosperm not only will the seed not reach mature size, but the MSC contents may not be correctly arranged and/or pressurised causing the diminished mucilage expansion shown here. Unlike *Arabidopsis*, several *Plantago* species are reported to contain specialised nutrient transfer structures called haustoria that develop from the embryo sac. Haustoria have been characterised in the seeds of *P. lanceolata*[48], *P. major*[49], and *P. coronopus*, while those in *P. pumila* (also known as *P. exigua*) and *P. lagopus* are described briefly by Johri et al.[50]. Cooper[48] observed haustoria "penetrating and digesting the outer portion of the ovule adjacent to the developing endosperm", in *P. lanceolata*, and this corresponds to the layer we have designated integument in *P. ovata*. Haustoria may function to directly connect the embryo and endosperm to surrounding integument cells, allowing a networked supply of carbon for growth and development. Eventually, the growing endosperm of *P. lanceolata* absorbs most of the surrounding integument and leaves only a few cell layers that lose most of their cytoplasmic contents and are squashed thin at maturity. Although Cooper[48] did not specifically state what this papery-thin layer could be, it is likely that they were describing the mucilage polysaccharide layer. Haustoria have not yet been reported in *P. ovata*, and we were unable to confirm their presence or absence in the developmental sections presented here. Thus, it remains unclear what mechanisms control the fate of the integument cells and this will be informative to investigate in the future.

The microscopy images of the developing MSCs (Fig. 2) raise questions regarding gene expression and regulation during the different developmental stages. During the early stages of MSC development at 3–5 DPA where rapid cell expansion is observed, the enzymes that are present may be synthesising the backbone of the immature pectin polymer i.e. one that still requires post-synthesis modification to become hydrophilic, as observed in *Arabidopsis*[51–53] and early stages of heteroxylan synthesis may also be occurring. The shift in the esterification status of the pectin in both the MSCs and the integument cell walls is clearly demonstrated in Fig. 3 and SI Fig. S2, where tissues at anthesis label differently when compared to four days later. However, at this magnification it is not possible to unequivocally define the location of the pectin—whether it is in the actual wall of the MSCs or in the apoplast and just pushed tightly against it will require detailed examination at the TEM level. There is a clear increase in the amount of crystalline cellulose in the walls of both the MSCs and the integument cells when tissues at anthesis and later at 4 and 7 DPA are compared. At these early stages there is minimal binding by the LM11 antibody which detects xylan suggesting there is no xylan present yet, even in the cell walls rather than the apoplast (SI Fig. S2B,D,F). This lack of signal is consistent with our previous analyses, where later in development (13 DPA) many of the genes involved in xylan biosynthesis are transcriptionally active, such as the GT61, UXS, and UAM genes[11] which is only two days before the 15 DPA stage when the MSCs appear misshapen and possibly on the verge of extensive disintegration (Fig. 2G). It is possible that at 13 DPA these genes are more involved in polysaccharide post-synthesis modification, modifying the sidechain density and/or length in xylan and pectic polymers to enable correct mucilage expansion and final architecture upon imbibition in aqueous environments, but at this stage details are unknown. Future experiments are aimed at establishing the transcript abundance of gene sub-sets involved in synthesis of the pectic backbone including GT8, GAUT1 and GAUT7[54] members, the addition of minor substituents onto the pectin backbone and the biosynthesis of nascent and mature heteroxylan types. A precise temporal series employing laser capture microdissection of the developing MSCs, followed by RNAseq analysis will prove invaluable in this context.

Our microscopy analyses reveal that, in contrast to *Arabidopsis*, the seed mucilage release mechanism of *P. ovata* may be a physical process not dependent on cellular rupture followed by extrusion (Fig. 6B). Hence, for this species we suggest that it is more appropriate to describe the mechanism as an expansion rather than an


A
r = -0.632, p = 0.03
Xyl+Ara
Rha+GalA
Xyl+Ara Content (% w/w)
Rha+GalA Content (% w/w)
Extraction Time (Minutes)
B
Min.
S
ML
L2
L1
0
1
2
3
4
5
15
20

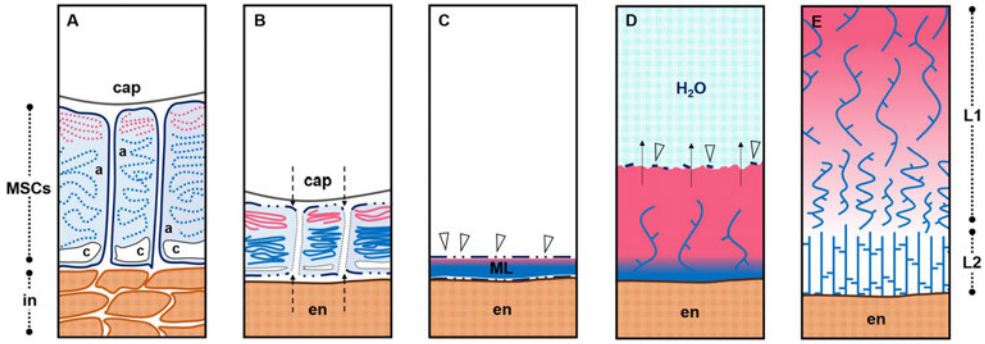


Figure 7. A proposed model of the polysaccharide deposition and mucilage expansion mechanism in *P. ovata*. (**A**) Mucilage polysaccharides pectin (pink) and heteroxylan (blue) are polarly synthesised and deposited into the outer apoplast (a) of mucilage secretory cells (MSCs), (**B**) which become compressed between the endosperm (en) and capsule wall (cap), obliterating the radial walls and releasing their contents into a (**C**) continuous laminated cell-free mucilage polysaccharide layer (ML) on the external seed surface when released from the capsule. (**D**) Upon exposure to an aqueous environment, the hydrophilic mucilage starts to expand outwards from the seed surface. The first layer to expand is rich in extremely hydrophilic and soluble pectin and is topped by fragile remnants of the distal MSC walls (arrowheads) that dissolve as hydration continues. This layer works to provide a hydration cascade to initiate and jumpstart hydration and swelling of the more gel-like, less hydrophilic polymers. (**E**) A hypothesised distribution of mucilaginous polysaccharides. The pink gradient is indicative of the distribution of pectin which is restricted to the periphery (or 'mucilage expansion front') of the expanded seed mucilage (as per Fig. 6) whilst the xylan polysaccharides (blue strands) are evenly distributed (Fischer et al.[12]; Guo et al.[55]; Yu et al.[9]). In L1, the enrichment of pectin may proportionally modulate the solubility/extractability of xylan while in L2, pectin is not present and thus xylan polymers form a more robust, gel-like layer. *In* integument, *c* cytoplasm.

extrusion, similar to the extension of a concertina, where the mucilage starts to expand into the aqueous environment as it hydrates from the wrinkled appressed polysaccharide layer on the seed surface. Data demonstrating the change in mucilage composition and structure over time (Fig. 6) support this type of expansion process. The driving forces behind expansion of *P. ovata* seed mucilage could be derived from the differential hydrophilicity of the constituent mucilaginous polysaccharides. Pectin is most abundant in the first layer of mucilage to expand, L1 (Fig. 6B) and previous studies have described the outermost layer of the expanded seed mucilage to be a highly soluble, pectin-rich fraction[9,12,55]. In *P. ovata*, this seed mucilage fraction was easily extracted using cold water[9,10] and we propose that its function is to act as a primer, initiating mucilage expansion, and providing a hydration cascade triggering the swelling of the more gel-like and structurally-complex heteroxylan polymers located in subsequent fractions/layers (SI Fig. S6). Compositional data supporting such patterns of polymer release have been reported previously[9,10,13], and now a model to illustrate this process, driven by polarised deposition and then expansion, is presented in Fig. 7. To fulfil such a role the pectin-enriched fraction must be synthesised and/or deposited first into the distal end of the MSC, anchoring the mucilage to the seed to form L1, after which the structural polymers, including heteroxylan, are synthesised and/or deposited into the basal end of the cell to make L2 (Fig. 6A). A similar spatio-temporal pattern of polysaccharide synthesis and deposition was recently described by Miart et al.[36] who showed that RG-I (pectin) was synthesised in the two outermost layers of *L. usitatissimum* MSCs prior to synthesis of other polysaccharides in the layers beneath. The authors hypothesised that the arabinoxylan, xyloglucan and cellulose polysaccharides synthesised later in the inner layers provided a structural element that pressurised the outermost contents, enabling efficient mucilage release and anchoring the mucilage to the seed. Similarly, the heterotypic interactions of various polymers including branched xylan, cellulose and arabinogalactan proteins are important for effective mucilage release and adherence in *Arabidopsis*[3,5,6,8,19,56]. In support of this temporal sequence of events, real-time qPCR analysis of cDNA from developing *L. usitatissimum* integument tissue shows that genes involved in pectin biosynthesis are transcriptionally active prior to those associated with xylan biosynthesis (Aubert et al., University of Adelaide, unpublished data). The mucilage component of *P. ovata*, and likely many other species, is a complex network of heterogeneously distributed polysaccharides. Each polymer must be synthesised, deposited, and potentially modified, in a specific sequence and location during seed development to be able to fulfil the required mechanical functions enabling mucilage release, and supporting the structural functions of the material once it extends from the seed surface. The process of mucilage polysaccharide biosynthesis and deposition into the MSCs must therefore be a tightly regulated process, about which we have much to learn.

In future work, it would be valuable to characterise the MSCs of *Plantago* species such as *P. cunninghamii* that we have already confirmed to possess a similar seed surface arrangement to *P. ovata*, but that has a different expanded mucilage architecture[11]. Our preliminary characterisation of the MSCs of *P. ovata* has generated further questions regarding the biosynthesis and deposition of mucilaginous polysaccharides: how, where, and

in what order are these polymers transported and deposited into the MSCs? What genes and regulatory elements are controlling this highly complex process? And what drives the fate of the integument cells? Should we be looking for signs of programmed cell death or a suite of cell wall degrading enzymes in this tissue? Combining further histological analysis of the developing seeds, with a focus on the MSCs and the integument tissue, with characterisation of the temporal regulation of mucilage biosynthetic transcripts may begin to answer some of these questions. Our whole mount immunolabelling data suggests that hydrated heteroxylan is distributed in a specific digit-like pattern whilst immunolabelling of seed sections show that the heteroxylan and pectin are homogenously distributed throughout the expanded seed mucilage (Phan et al.[11]). It remains unclear how these polymers are deposited and distributed in both the developing MSCs, the mature mucilage polysaccharide layer, and the final expanded material. Thorough investigation of the developing MSCs in *P. ovata* may begin to shed some light upon these questions and allow us to fine tune our hypothetical model, whilst eventually providing tools to allow manipulation of the mucilage quantity and quality that could directly impact downstream applications and economics of psyllium use.

Materials and methods

Plant materials and growth. Wild-type *P. ovata* and gamma-irradiated *P. ovata* mutant 69-1 were obtained from a population previously generated by Tucker et al.[43]. Three 69-1 sister lines at M4 were tested to show that > 95% of the seeds in each sister line displayed a developmentally delayed phenotype. The mutant line has not yet been backcrossed to wild type.

Plants were grown as per Phan et al.[11]. To stage wild-type fruit development, fruits with freshly emerged anthers (erect and bright-yellow in colour) were marked and tagged with the date in order to harvest at the relevant day post-anthesis (DPA).

Observing *P. ovata* expanded seed mucilage. *Ruthenium red.* Mature *P. ovata* seeds were individually placed onto microscopy slides in a ruthenium red solution at a concentration of 0.01% (w/v) (ProSciTech, C075, Australia). Seeds were observed under a Zeiss Stemi 2000-C dissecting microscope with an attached AxioCam ERc 5s camera. Seeds with impaired development were selected from the mutant line 69-1[43].

Time-lapse. Mature *P. ovata* seeds were prepared as per Yu et al.[9]. In brief, dry mature *P. ovata* seeds were soaked overnight in stain solution comprised of 0.1% w/v Calcofluor White (Fluorescent Brightener 28, Sigma-Aldrich) and 0.4% w/v Direct Red 23 (Sigma-Aldrich) diluted in 80% ethanol. The seeds were removed from the staining solution and allowed to air dry before being adhered with a cyanoacrylate adhesive to the centre of a Petri dish. The Petri dish was mounted onto the stage of a Nikon A1R Laser Scanning Confocal with DS-Ri1 CCD camera and imaged prior to the addition of deionised water onto the seed at time = 0. Images were captured for 20 min in total at 1 min intervals.

Fixation, embedding, and sectioning of *P. ovata* developing fruit. Samples requiring fixation, embedding, and sectioning were processed as per Burton et al.[57] and embedded tissue was sectioned at 1 µm on an Ultramicrotome (Leica, EM UC6) using a diamond knife (DiATOME, Nidau, Switzerland). For non-aqueous fixation, PBS was replaced with an 80% ethanol solution. Sections were stained with Toluidine Blue (epoxy tissue stain, used undiluted, ProSciTech, C149, Australia). Sections were imaged under transmitted light differential interference contrast (DIC) using a Zeiss Axio Imager M2 (Carl Zeiss, Germany) fitted with an AxioCam MRm3 monochrome camera.

For fluorescence images, samples were fixed, embedded, and sectioned as above for non-aqueous fixation. Survey sections were stained with epoxy tissue stain (used undiluted, ProSciTech, Australia). For immunofluorescence, sections were incubated with monoclonal antibodies raised against pectin (LM19 and LM20) and arabinoxylan (LM11) (PlantProbes Leeds, UK) followed by an appropriate AlexaFluor 555 secondary antibody (Invitrogen, USA). The His-tagged carbohydrate-binding module CBM3a (PlantProbes, Leeds, UK) was used with a triple indirect immunofluorescence labelling procedure as described previously[11,58]. Sections were counterstained using Calcofluor White (Fluorescent Brightener 28; Sigma-Aldrich) and mounted in glycerol. Images were obtained using an AxioCam 105 color camera fitted to a Zeiss fluorescence microscope (Axio Imager M2, Carl Zeiss, Germany) with 254/432 nm excitation/emission wavelength for Calcofluor White and 553/568 nm excitation/emission wavelength for LM11/LM19/LM20/CBM3a.

Scanning electron microscopy. Mature seeds of *P. ovata* and *Arabidopsis thaliana* ecotype *Columbia-0* were air-dried before and after mucilage hydration then sputter-coated with platinum at a thickness of 5 nm. Seeds were imaged at a working distance of 7.5 mm with an accelerating voltage of 3 kV using a Philips XL20 Scanning Electron Microscope (SEM) following Phan et al.[11].

Mucilage extraction and compositional analysis. For temporal mucilage analysis, mucilage was collected by placing 1 g mature seeds into a wide-mouth sieve, placed in a water bath containing 40 mL of deionised water at room temperature with intermittent stirring. Fractions were collected at 5 min intervals by transferring the sieve and seeds into a fresh batch of deionised water. Mucilage extracts were freeze–dried to a constant weight and compositional analysis was conducted as per Hassan et al.[59].

For comparison of mutant 69-1 with the wild-type, mucilage was extracted from 40 seeds for 3 h in 20 mL of deionised water heated to 80 °C and stirred vigorously on a heated magnetic stirrer. While still hot, the mucilage and seeds were transferred into a 50 mL tube and centrifuged for 10 min at 4,000 rpm. The mucilage supernatant

was decanted into a new tube and freeze–dried to a constant weight. Yield per seed was calculated by dividing the freeze–dried mucilage mass by 40. Compositional analysis was conducted as per Hassan et al.[59].

Data availability

The datasets generated and/or analysed in this study are available from the corresponding author on reasonable request.

Received: 12 March 2020; Accepted: 25 June 2020
Published online: 16 July 2020

References

1. Phan, J. L. & Burton, R. A. New insights into the composition and structure of seed mucilage. *Annu. Plant Rev. Online* **1**, 1–41 (2018).
2. Macquet, A., Ralet, M.-C., Kronenberger, J., Marion-Poll, A. & North, H. M. *In situ*, chemical and macromolecular study of the composition of *Arabidopsis thaliana* seed coat mucilage. *Plant Cell Physiol.* **48**, 984–999 (2007).
3. Harpaz-Saad, S. *et al.* Cellulose synthesis via the FEI2 RLK/SOS5 pathway and CELLULOSE SYNTHASE 5 is required for the structure of seed coat mucilage in *Arabidopsis*. *Plant J.* **68**, 941–953 (2011).
4. Yang, B. *et al.* TRM 4 is essential for cellulose deposition in Arabidopsis seed mucilage by maintaining cortical microtubule organization and interacting with CESA 3. *New Phytol.* **221**, 881–895 (2019).
5. Yu, L. *et al.* CELLULOSE SYNTHASE-LIKE A2, a glucomannan synthase, is involved in maintaining adherent mucilage structure in *Arabidopsis* seed. *Plant Physiol.* **164**, 1842–1856 (2014).
6. Voiniciuc, C., Günl, M., Schmidt, M.H.-W. & Usadel, B. MUCI10 produces galactoglucomannan that maintains pectin and cellulose architecture in *Arabidopsis* seed mucilage. *Plant Physiol.* **169**, 403–420 (2015).
7. Hu, R. *et al.* Xylan synthesized by Irregular Xylem 14 (IRX14) maintains the structure of seed coat mucilage in Arabidopsis. *J. Exp. Bot.* **67**, 1243–1257 (2016).
8. Hu, R. *et al.* Irregular xylem 7 (IRX7) is required for anchoring seed coat mucilage in Arabidopsis. *Plant Mol. Biol.* **92**, 25–38 (2016).
9. Yu, L. *et al.* Multi-layer mucilage of *Plantago ovata* seeds: Rheological differences arise from variations in arabinoxylan side chains. *Carbohydr. Polym.* **165**, 132–141 (2017).
10. Cowley, J. M. *et al.* A small-scale fractionation pipeline for rapid analysis of seed mucilage characteristics. *Plant Methods* **16**, 1–12 (2020).
11. Phan, J. L. *et al.* Differences in glycosyltransferase family 61 accompany variation in seed coat mucilage composition in *Plantago* spp.. *J. Exp. Bot.* **67**, 6481–6495 (2016).
12. Fischer, M. H. *et al.* The gel-forming polysaccharide of psyllium husk (*Plantago ovata* Forsk). *Carbohydr. Res.* **339**, 2009–2017 (2004).
13. Ren, Y., Yakubov, G. E., Linter, B. R., Macnaughtan, W. & Foster, T. J. Temperature fractionation, physicochemical and rheological analysis of psyllium seed husk heteroxylan. *Food Hydrocoll.* https://doi.org/10.1016/j.foodhyd.2020.105737 (2020).
14. Jensen, J. K., Johnson, N. R. & Wilkerson, C. G. *Arabidopsis thaliana* IRX10 and two related proteins from psyllium and *Physcomitrella patens* are xylan xylosyltransferases. *Plant J.* **80**, 207–215 (2014).
15. Urbanowicz, B. R., Peña, M. J., Moniz, H. A., Moremen, K. W. & York, W. S. Two Arabidopsis proteins synthesize acetylated xylan in vitro. *Plant J.* **80**, 197–206 (2014).
16. Chiniquy, D. *et al.* XAX1 from glycosyltransferase family 61 mediates xylosyltransfer to rice xylan. *Proc. Natl. Acad. Sci. USA* **109**, 17117–17122 (2012).
17. Anders, N. *et al.* Glycosyl transferases in family 61 mediate arabinofuranosyl transfer onto xylan in grasses. *Proc. Natl. Acad. Sci. USA* **109**, 989–993 (2012).
18. Voiniciuc, C., Gunl, M., Schmidt, M.H.-W. & Usadel, B. Highly branched Xylan made by IRREGULAR XYLEM14 and MUCILAGE-RELATED21 Links mucilage to. *Plant Physiol.* **169**, 2481–2495 (2015).
19. Ralet, M.-C. *et al.* Xylans provide the structural driving force for mucilage adhesion to the Arabidopsis seed coat. *Plant Physiol.* **171**, 165–178 (2016).
20. Jensen, J. K., Johnson, N. & Wilkerson, C. G. Discovery of diversity in xylan biosynthetic genes by transcriptional profiling of a heteroxylan containing mucilaginous tissue. *Front. Plant Sci.* **4**, 20 (2013).
21. Marlett, J. A., Kajs, T. M. & Fischer, M. H. An unfermented gel component of psyllium seed husk promotes laxation as a lubricant in humans. *Am. J. Clin. Nutr.* **72**, 784–789 (2000).
22. McRorie, J. W. *et al.* Psyllium is superior to docusate sodium for treatment of chronic constipation. *Aliment. Pharmacol. Ther.* **12**, 491–497 (1998).
23. Anderson, J. W. *et al.* Cholesterol-lowering effects of psyllium intake adjunctive to diet therapy in men and women with hypercholesterolemia: Meta-analysis of 8 controlled trials. *Am. J. Clin. Nutr.* **71**, 472–479 (2000).
24. Cappa, C., Lucisano, M. & Mariotti, M. Influence of Psyllium, sugar beet fibre and water on gluten-free dough properties and bread quality. *Carbohydr. Polym.* **98**, 1657–1666 (2013).
25. Mancebo, C. M., San Miguel, M. Á, Martínez, M. M. & Gómez, M. Optimisation of rheological properties of gluten-free doughs with HPMC, psyllium and different levels of water. *J. Cereal Sci.* **61**, 8–15 (2015).
26. Fratelli, C., Muniz, D. G., Santos, F. G. & Capriles, V. D. Modelling the effects of psyllium and water in gluten-free bread: An approach to improve the bread quality and glycemic response. *J. Funct. Foods* **42**, 339–345 (2018).
27. Haque, A. & Morris, E. R. Combined use of ispaghula and HPMC to replace or augment gluten in breadmaking. *Food Res. Int.* **27**, 379–393 (1994).
28. Bahrani, A. S. Processes for dehusking psyllium seeds. *US Pat. Number 5020732* (1991).
29. Kumar, J. *Good agricultural practices for isabgol*. (2015).
30. Francoz, E., Ranocha, P., Burlat, V. & Dunand, C. *Arabidopsis* seed mucilage secretory cells: Regulation and dynamics. *Trends Plant Sci.* **20**, 515–524 (2015).
31. Schneitz, K., Huiskamp, M. & Pruitt, R. E. Wild-type ovule development in *Arabidopsis thaliana*: A light microscope study of cleared whole-mount tissue. *Plant J.* **7**, 731–749 (1995).
32. Windsor, J. B., Symonds, V. V., Mendenhall, J. & Lloyd, A. M. Arabidopsis seed coat development: Morphological differentiation of the outer integument. *Plant J.* **22**, 483–493 (2000).
33. Beeckman, T., Rycke, R. D., Viane, R. & Inzé, D. Histological study of seed coat development in *Arabidopsis thaliana*. *J. Plant Res.* **113**, 139–148 (2000).
34. Western, T. L., Skinner, D. J. & Haughn, G. W. Differentiation of mucilage secretory cells of the *Arabidopsis* seed coat. *Plant Physiol.* **122**, 345–356 (2000).

35. Voiniciuc, C., Yang, B., Schmidt, M.H.-W., Gunl, M. & Usadel, B. Starting to gel: How *Arabidopsis* seed coat epidermal cells produce specialized secondary cell walls. *Int. J. Mol. Sci.* **16**, 3452–3473 (2015).
36. Miart, F. *et al.* Cytological approaches combined with chemical analysis reveals the layered nature of flax mucilage. *Front. Plant Sci.* **10**, 1–16 (2019).
37. Witzum, A. Mucilaginous plate pells in the nutlet epidermis of *Coleus blumei* Benth. (Labiatae). *Bot. Gaz.* **139**, 430–435 (1978).
38. Muñoz, L. A., Cobos, A., Diaz, O. & Aguilera, J. M. Chia seeds: Microstructure, mucilage extraction and hydration. *J. Food Eng.* **108**, 216–224 (2012).
39. Hyde, B. B. Mucilage-producing cells in the seed coat of *Plantago ovata*: Developmental fine structure. *Am. J. Bot.* **57**, 1197–1206 (1970).
40. Ruprecht, C. *et al.* A synthetic glycan microarray enables epitope mapping of plant cell wall glycan-directed antibodies. *Plant Physiol.* **175**, 00737 (2017).
41. Verhertbruggen, Y., Marcus, S. E., Haeger, A., Ordaz-Ortiz, J. J. & Knox, J. P. An extended set of monoclonal antibodies to pectic homogalacturonan. *Carbohydr. Res.* **344**, 1858–1862 (2009).
42. Ruel, K., Nishiyama, Y. & Joseleau, J. P. Crystalline and amorphous cellulose in the secondary walls of Arabidopsis. *Plant Sci.* **193–194**, 48–61 (2012).
43. Tucker, M. R. *et al.* Dissecting the genetic basis for seed coat mucilage heteroxylan biosynthesis in plantago ovata using gamma irradiation and infrared spectroscopy. *Front. Plant Sci.* **8**, 326 (2017).
44. Voiniciuc, C. *et al.* Flying saucer1 is a transmembrane RING E3 ubiquitin ligase that regulates the degree of pectin methylesterification in Arabidopsis seed mucilage. *Plant Cell* **25**, 944–959 (2013).
45. Boesewinkel, F. D. Development of ovule and testa of *Linum usitatssimum* L.. *Acta Bot. Neerl.* **29**, 17–32 (1980).
46. Madgulkar, A., Rao, M. & Warrier, D. Characterization of Pysillium (*Plantago ovata*) polysaccharide and its uses. *Polysaccharides* https://doi.org/10.1007/978-3-319-03751-6_49-1 (2014).
47. Garcia, D., FitzGerald, J. N. & Berger, F. Maternal control of integument cell elongation and zygotic control of endosperm growth are coordinated to determine seed size in Arabidopsis. *Plant Cell* **17**, 52–60 (2005).
48. Cooper, G. O. Development of the ovule and the formation of the seed in *Plantago lanceolata*. *Am. J. Bot.* **29**, 577–581 (1942).
49. Mikesell, J. Anatomy of terminal haustoria in the ovule of Plantain (*Plantago major* L.) with taxonomic comparison to other angiosperm taxa. *Bot. Gaz.* **151**, 452–464 (1990).
50. Johri, B., Ambegaokar, K. & Srivastava, P. *Plantaginales. Comparative Embryology of Angiosperms* 777–779 (Springer, Berlin, 1992).
51. Saez-Aguayo, S. *et al.* PECTIN METHYLESTERASE INHIBITOR6 promotes *Arabidopsis* mucilage release by limiting methylesterification of homogalacturonan in seed coat epidermal cells. *Plant Cell* **25**, 308–323 (2013).
52. Macquet, A. *et al.* A naturally occurring mutation in an *Arabidopsis* accession affects a β-D-galactosidase that increases the hydrophilic potential of rhamnogalacturonan I in seed mucilage. *Plant Cell* **19**, 3990–4006 (2007).
53. Walker, M. *et al.* The transcriptional regulator LEUNIG_HOMOLOG regulates mucilage release from the *Arabidopsis* testa. *Plant Physiol.* **156**, 46–60 (2011).
54. Atmodjo, M. A., Hao, Z. & Mohnen, D. Evolving views of pectin biosynthesis. *Annu. Rev. Plant Biol.* **64**, 747–779 (2013).
55. Guo, Q., Cui, S. W., Wang, Q. & Christopher Young, J. Fractionation and physicochemical characterization of psyllium gum. *Carbohydr. Polym.* **73**, 35–43 (2008).
56. Sullivan, S. *et al.* CESA5 is required for the synthesis of cellulose with a role in structuring the adherent mucilage of Arabidopsis seeds. *Plant Physiol.* **156**, 1725–1739 (2011).
57. Burton, R. A. *et al.* Over-expression of specific HvCslF cellulose synthase-like genes in transgenic barley increases the levels of cell wall (1,3;1,4)-β-D-glucans and alters their fine structure. *Plant Biotechnol. J.* **9**, 117–135 (2011).
58. Guillon, F. *et al.* Brachypodium distachyon grain: Characterization of endosperm cell walls. *J. Exp. Bot.* **62**, 1001–1015 (2011).
59. Hassan, A. S. *et al.* A Genome Wide Association Study of arabinoxylan content in 2-row spring barley grain. *PLoS One* **12**, 1–19 (2017).

Acknowledgements
The authors would like to thank Associate Professor Matthew R Tucker for ongoing scientific discussions. The authors acknowledge the facilities, and the scientific and technical assistance, of the Australian Microscopy and Microanalysis Research Facility at Adelaide Microscopy, The University of Adelaide, Waite Campus, for their assistance in using the SEM and confocal microscope and Dr Long Yu for assistance in developing a method to stain the mucilage of mature seeds. This work was supported by the Australian Research Council (ARC) Centre of Excellence in Plant Cell Walls (CE110001007) and Plant Energy Biology (CE140100008). JP was supported by a University of Adelaide's CJ Everard PhD Scholarship, a Grains Research and Development Corporation (GRDC) scholarship and a SARDI Bursary, JMC was supported by the Australian Government's Research Training Program and LH was supported by a University of Adelaide's Adelaide Graduate Research Scholarship.

Author contributions
J.P. and R.A.B. conceived the study. J.P., J.M.C., K.A.N., L.H. and L.A.O. prepared materials and performed experiments. All authors contributed to the data analysis and interpretation. J.P., J.M.C. and R.A.B. wrote the manuscript. All authors approved the final manuscript.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-68685-w.

Correspondence and requests for materials should be addressed to R.A.B.

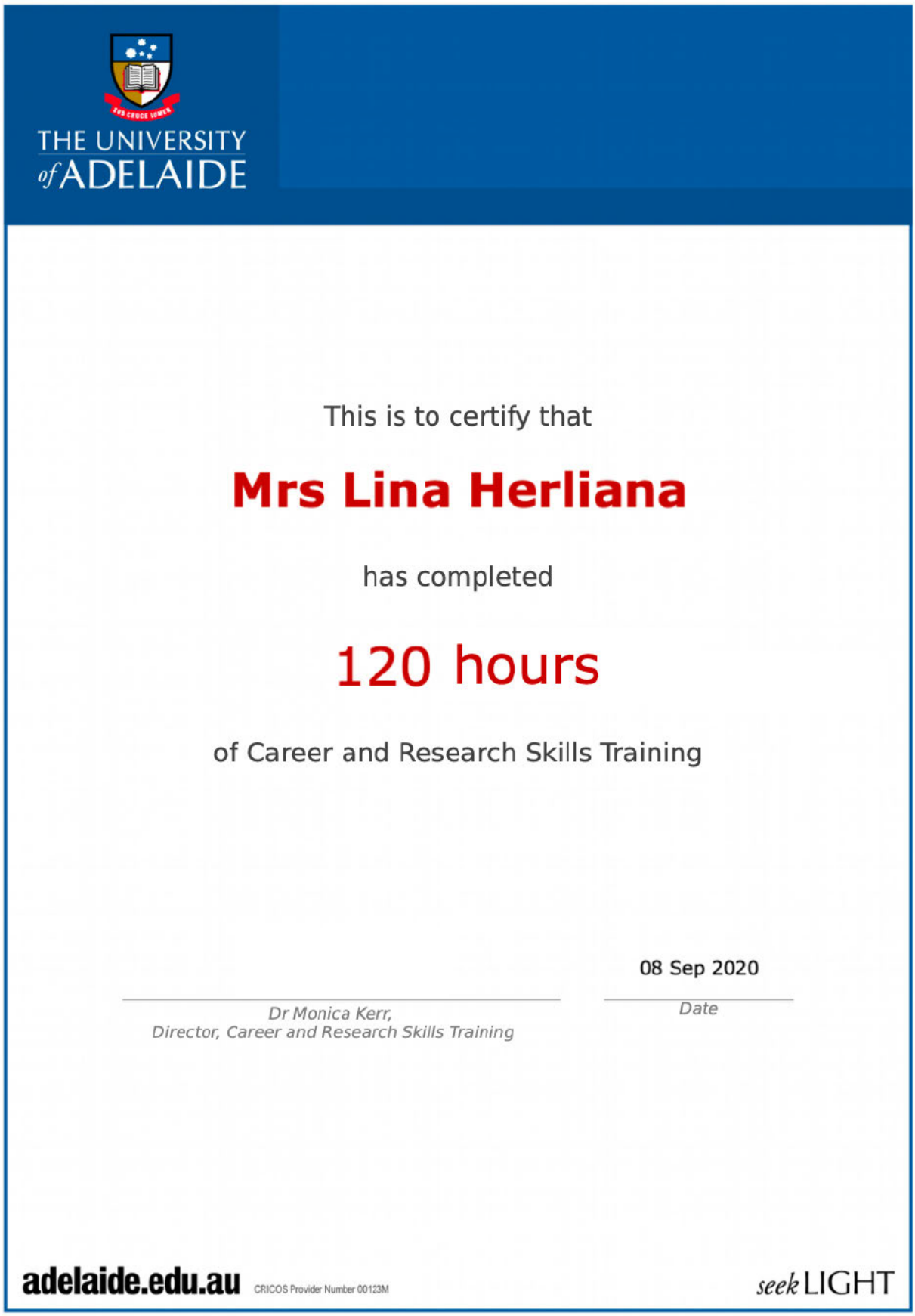
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix III

Career and Research Skills Traiing (CaRST)

ꕥ•ꕥ



THE UNIVERSITY of ADELAIDE

This is to certify that

Mrs Lina Herliana

has completed

120 hours

of Career and Research Skills Training

08 Sep 2020

Dr Monica Kerr,
Director, Career and Research Skills Training

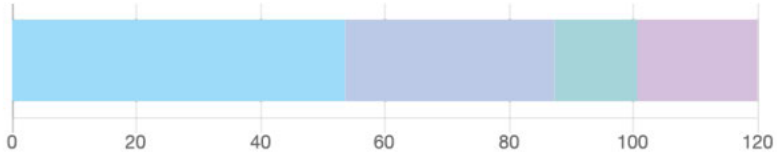
Date

adelaide.edu.au CRICOS Provider Number 00123M

seek LIGHT

Appendix III – CaRST Record

CaRST Credits Overview



120

Recognised Credits

A: 104 credits in Knowledge and intellectual abilities
B: 37.5 credits in Personal effectiveness
C: 15 credits in Research governance and organisation
D: 21.5 credits in Engagement, influence and impact

Date Completed	Activity Description	Domain	Hours
20-Jul-18	Postgraduate Research Induction	Domain B	2
21-Jul-18	Research Methods in Literature Review - Epigeum	Domain A	3
28-Jul-18	How to Plan Your PhD - Online	Domain C	2
3-Aug-18	EndNote: Basics - Lecture	Domain A	1
8-Aug-18	Research Integrity - Epigeum	Domain C	5
16-Aug-18	Mendeley and Zotero: Which is the right free reference management tool for you?	Domain A	0
18-Aug-18	Presenting your Research with Confidence - Online	Domain D	2
18-Aug-18	IP 101 - Online	Domain C	1
18-Aug-18	Thriving in your Life as an HDR - Online	Domain B	2
20-Aug-18	Biological Safety Management	Domain C	1
20-Aug-18	Chemical Safety Management	Domain C	1
21-Aug-18	CaRST Information Session	Domain C	1
23-Aug-18	Communicating the Impact of Your Research - Online	Domain D	2
23-Aug-18	Pronouncing scientific English: Identifying and addressing your personal challenges	Domain A	4
24-Aug-18	The Imposter Syndrome - Online	Domain B	2
25-Aug-18	Communication and 'win-win' skills as an HDR - Online	Domain D	2
25-Aug-18	Avoiding Plagiarism - Epigeum	Domain C	1
28-Aug-18	Negotiating for Positive Outcomes - Online	Domain D	3
29-Aug-18	The Balanced Researcher - Online	Domain C	2
30-Aug-18	Leadership and the Art of Influence - Online	Domain D	3
31-Aug-18	Managing your Research Data	Domain C	3

1-Sep-18	Introduction to Entrepreneurship - Online	Domain D	2
2-Sep-18	Statistical Methods in Natural Sciences – Epigeum	Domain A	12
6-Sep-18	Commercialisation: Are You Ready? - Online	Domain D	1
8-Sep-18	Reviving your Life while Doing an HDR - Online	Domain B	2
9-Sep-18	The Self-Reflective HDR - Online	Domain B	2
13-Sep-18	The science behind GMs	Domain A	2
14-Sep-18	Demystifying Research Metrics Lecture - Online	Domain A	1
17-Sep-18	Career Control for Researchers	Domain B	13
17-Sep-18	Strategic Networking	Domain B	3
18-Sep-18	Introduction to SPSS for Statistics	Domain A	6
19-Sep-18	Postgraduate Symposium 2018	Domain A	7
24-Sep-18	Managing Large Data Lists with Excel - Lecture (Demonstration Only)	Domain A	1
25-Sep-18	Introduction to R for Statistics	Domain A	6
28-Sep-18	Analysing Data with Pivot Tables in Excel - Lecture (Demonstration Only)	Domain A	1
3-Oct-18	EpiCSA 3rd Annual General Meeting	Domain B	3
18-Oct-18	Seven Secrets of Highly Successful Research Students	Domain B	2.5
2-Nov-18	Integrated Bridging Program Research (IBP-R)	Domain A	30
16-Nov-18	Turbocharge your Writing - Online	Domain D	2.5
16-Nov-18	Improve your Confidence with Improv	Domain B	2
6-Jun-19	NVivo Basics	Domain A	3.5
13-Jun-19	NVivo Intermediate	Domain A	3.5
25-Sep-19	Imaris 3D/4D Image Analysis Workshop	Domain A	10
3-Feb-20	2020 Symposium Down Under: Mechanisms Controlling Plant Reproduction	Domain A	13
15-Mar-20	Scientific Writing Workshops	Domain D	4
19-Aug-20	Improve Your English Pronunciation and Fluency - A Workshop for International Students	Domain B	2