Soybean nodule development and senescence:

The role of cysteine proteases and their inhibitors

By

STEFAN GEORGE VAN WYK

submitted in partial fulfilment of the requirements for the degree

PHILOSOPHIAE DOCTOR

in the

Faculty of Natural and Agricultural Sciences

Department of Plant Production and Soil Sciences

UNIVERSITY OF PRETORIA

Pretoria

SUPERVISORS:

DR. B.J. VORSTER

PROF. C.A. CULLIS

May 2015

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF.

I, Stefan George van Wyk, declare that this thesis, which I hereby submit for the degree Philosophiae Doctor (Biotechnology) at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature: 

4 May 2015

ABSTRACT

Background: Root nodules in soybean play an important role in fixation of atmospheric nitrogen used for plant growth. Premature senescence of nodules can negatively impact on nitrogen availability for plant growth and, as such, a better understanding of nodule development and senescence is required. Cysteine proteases are known to play a role in nodule senescence, but knowledge is still fragmented regarding the function of their inhibitors (cystatins) during the development and senescence of soybean nodules.

Results: RNA-Seq expression analysis showed that transcription of cystatins Glyma05g28250, Glyma15g12211, Glyma15g36180 increased during onset of senescence, possibly regulating proteolysis when nodules senesce and undergo programmed cell death. Biochemical inhibitory assays with recombinantly expressed and purified cystatins showed that most cystatins had preferential affinity to cathepsin L-like cysteine proteases. Both actively- and non-actively transcribed nodule cystatins inhibited cathepsin-L- and B-like activities in different age nodules and they also inhibited papain and cathepsin-L activity. The localisation of these proteins could not be determined with the approach that was followed.

Conclusions: This PhD study provided the first evidence with regard to cystatin expression during nodule development combined with biochemical characterization of their inhibition strength. Knowledge about the expression, localization and possible function of the cysteine protease-cystatin system during soybean nodule development was generated. Overlap in activities and specificities of actively and non-actively transcribed cystatins, raises the question if non-transcribed cystatins provide a reservoir for response to particular environments. This data might be applicable to the development of strategies to extend the active life span of nodules or prevent environmentally induced senescence.

THESIS COMPOSITION

Chapter 1 of this PhD thesis provides a background of the importance of soybean and other legumes in society and agriculture. Furthermore, the study of nodule development and senescence and the agricultural importance of symbiotic nitrogen fixation in these plants are illustrated. The aim, hypothesis and objectives set for this study is supplied at the end of the chapter. **Chapter 2** outlines all the resources and methods utilised during this study to achieve the aim and objectives set for this study to investigate the working hypotheses under investigation. **Chapter 3** of thesis provides all the results obtained from the experimental work conducted. The transcriptomic analysis, phylogenetic work, and *in vitro* and *ex vivo* experiments, provide a wealth of information regarding soybean cystatins and cysteine proteases at work during nodule development and senescence. The localisation of these proteins is investigated at the end of this chapter. **Chapter 4** discusses the results obtained and interprets the overall importance of findings to advance the understanding of the expression and possible function of cysteine proteases and their inhibitors and further explains the relevance to the original working hypothesis being addressed. At the end of the chapter the main findings of the study are summarised with the relevance to the field of knowledge and further highlights prospective research to further advance the main findings of this study. **Chapter 5** consists of all the literature sources used in the literature survey of this thesis, as well as other information sources related to tools and techniques utilised during this study. **Chapter 6** is composed of the appendices and provides additional information important and relevant to the study.

ACKNOWLEDGEMENTS

This work was funded by the International Foundation of Science (IFS grant C/5151-1), the NRF National Bioinformatics functional Genomics program (86947) (BJV). The funding received from the Genomic Research Institute, University of Pretoria, is hereby also acknowledged. SGVW thanks the NRF/DST for bursaries. The assistance of Kyle Logue and David Serre for developing the RNA-Seq data is hereby acknowledged.

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS AND SYMBOLS

BLAST	Basic Local Alignment Search Tool
bp	Base pairs
cDNA	complementary DNA
DNA	Deoxyribonucleic acid
FPKM	Fragments Per Kilobase of exon model per Million mapped
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
KOG	EuKaryotic Orthologous Groups
mRNA	messenger Ribonucleic acid
NCBI	National Centre for Biotechnology Information
PCD	Programmed Cell Death
PCR	Polymerase Chain Reaction
QPCR	Quantitative Real-Time PCR
RT-qPCR	Reverse Transcriptase qPCR
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
TAIR	The *Arabidopsis* Information Resource
VPE	Vacuolar-Processing Enzyme(s)

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 Significance of soybean

Soybean (*Glycine max* [L.] Merr.) is currently one of the world's most important crops used as sources for vegetable oil, in food and biodiesel, for vegetable protein and used world-wide in both food and animal feed. Although soybean originates from East Asia and was domesticated in China since 3500 B.C., its use in modern agriculture on a global scale has been only in the last 50 years (Keyser and Li 1992, Siddiqi 2000, James 2010, Oldroyd et al. 2011). The global soybean cultivation totalling 271.55 million metric tons (Figure 1.1) is mainly concentrated in the USA (31.58 %), Brazil (27.16 %), Argentina (16.46 %), China (9.41 %), India (4.7 %), Paraguay (2.47 %), Canada (1.89 %), Uruguay (1.14 %), Ukraine (0.95 %), and Bolivia (0.81 %). Currently, production in USA, Brazil, Argentina, China and India contributes to about 90 % of global demand of soybean-based products (Food and Agriculture Organization of The United Nations Statistics Division 2014). Local South African production of soybean has increased almost 21 % between 2012 to 2013, while imports have risen by over 1500 % during the same period (Department of Agriculture, Forestry and Fisheries 2013). The interest in soybean, and soybean-derived products, will continue to grow as global demand for food, feed and biofuels continue to increase along with population growth (USDA Agricultural Projections 2014). Sustainable production of high quality protein is, therefore, necessary for the economic and nutritional welfare of not only South Africa, but also the rest of the world (Dakora and Keya 1997, Food and Agriculture Organization of The United Nations Statistics Division 2014).

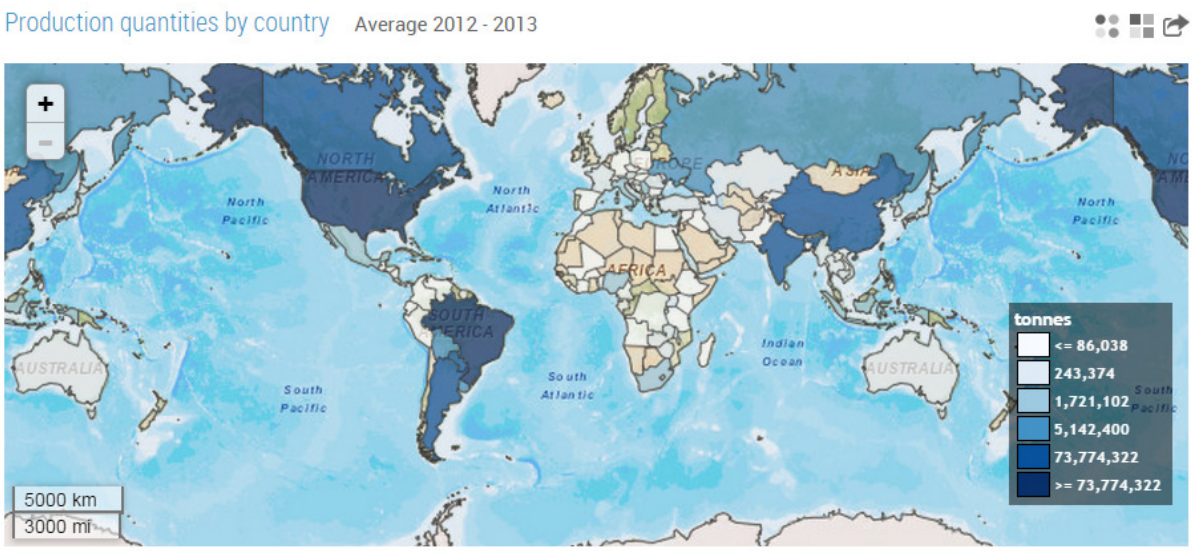


Figure 1.1 Global production of soybean during the 2012 – 2013 period. Figure taken from FAO website Global crop production statistics (Food and Agriculture Organization of The United Nations Statistics Division 2014).

Soybean seed is composed of up to 40 % protein, 21 % oil, 34 % carbohydrate and with 5 % ash (Snyder and Kwon 1987, Keyser and Li 1992, Liu 1997). Soybean protein is considered to be a complete protein containing significant amounts of all the essential amino acids required by the human body, however the majority of soybean meal is used as feed (James 2010). Soybean and its by-products provide an improved balance of amino acids and increased content of digestible sugars compared to other legumes, which enables highest possible milk and protein production required in modern agriculture (Bateman and Clark 2000, Ishler and Varga 2000, Grummer et al. 1994). Although the soybean protein quality is considered comparable to animal protein, soybean is more efficient and can produce more protein per area of land used than other commercial crops, in particular when compared to livestock (Dovring 1974, National Soybean Research Laboratory 2014). But to drive this accumulation of protein rich products, a nitrogen rich environment has to be supplied. This is typically achieved with nitrogen fertilizers, which is often too costly for small subsistence

farmers (Devienne-Barret et al. 2000, King and Purcell 2001). Fortunately, soybean, as well as several other legume species (e.g. alfalfa, clover, peas, beans, lentils, peanuts, etc.), can form a symbiotic association, via their root systems, with soil bacteria called Rhizobia (Schultze and Kondorosi 1998), which have evolved the special ability of fixing atmospheric nitrogen (N_2) into ammonia (NH_3) by the following chemical reaction:

$$N_2 + 8\ H^+ + 16\ ATP + 8\ e^- \rightarrow 2\ NH_3 + H_2 + 16\ ADP + 16\ P_i$$

The produced ammonia is subsequently converted to ammonium:

$$NH_3 + H^+ \rightarrow NH_4^+$$

This renders root nodules as a source for plant nitrogen, making legumes relatively rich in plant proteins (Howard and Rees 1996, Rees et al. 2005, Terpolilli et al. 2012). The plant uses the products from this bacterial nitrogen assimilation to synthesise essential macromolecules such as proteins. The plant in-turn supplies the bacteria with an energy source as well as produces the carbon source (e.g. sucrose) for the bacteria's own metabolic processes. This symbiosis drives the plant's development and ultimately determines the plant's yield (Chrispeels and Sadava 2003, Lee et al. 2004, Puppo et al. 2005). Symbiotic nitrogen fixation (SNF) therefore offers an important advantage in soybean when compared with most grain crops in that soybean fixes the nitrogen required for its growth and for the production of high-protein and oilseeds. It is estimated that under optimal conditions 300 kg N/ha can be fixed during this symbiotic interaction, providing a considerable contribution to the plant's requirements and also reduces drastically the production cost (Sanginga 2003, Hungria et al. 2006). This is achieved by minimizing the need for constant nitrogen supplementation with commercial fertilizers and furthermore enriching the surrounding soil with fixed nitrogen, which also makes soybean an ideal rotation crop (Keyser and Li 1992, Bergersen 1997, Duranti and Gius 1997, Espinosa-Victoria et al. 2000, Sanginga 2003, Puppo et al. 2005). However, this symbiotic interaction in soybean is short, lasting only about

11-12 weeks, after which nitrogen fixation declines rapidly as the nodules age and undergoes senescence, and has almost ceased by the time pod-filling starts (Espinosa-Victoria et al. 2000, Alesandrini et al. 2003a, Puppo et al. 2005). Nodule longevity and extent of nodule formation is regulated by the plant's hormone, signalling and metabolic pathways. The number and size of nodules are further positively correlated with the potential amount of nitrogen that can be fixed (Swaraj et al. 1995, Al-Karaki 2000, Ferguson and Mathesius 2003). SNF is, however, sensitive to abiotic stresses, such as drought, cold, salt or heavy metal stresses, which leads to premature senescence. Once nodule senescence is initiated, SNF capacity decreases and such early losses in nitrogen fixation capacity lead to nitrogen limitation within the plant. This has a major impact on seed production, crop quality and final yield (Pfeiffer et al. 1983, Alesandrini et al. 2003a).

1.2 Nodule physiology and activity

The development of the nodule, overall morphology, anatomy and nitrogen fixating capacity is determined by the plant's genome and also associated species of *Rhizobium* (Pereira et al. 1993, Swaraj et al. 1995). The nodules can be either determinate (in species such as common bean and soybean) or indeterminate (in species such as alfalfa and clover). Determinate nodules have vascular tissue arranged on the periphery of the nodule with cortical cell divisions and cell expansion giving the globular structure (Crespi and Galvez 2000, Van de Velde et al. 2006). Indeterminate nodules have a lateral arrangement of the vascular tissue and an apical meristem producing new cells on the terminal ends of the nodules leading to the elongated structure. Both nodule types have a central infection zone, the inner cortex with the distribution zone of small cells with large intercellular spaces, followed by the middle cortex with tightly packed boundary layer with few intercellular spaces, and finally the outer cortex,

which acts as physical barrier to maintain a low O_2 level within the nodule. An anaerobic environment is crucial for activity of bacterial nitrogenase, which is inhibited by high O_2 in the nodule (Appleby 1984). Figure 1.2 represents the morphology and physiology of a determinate nodule, such as that of soybean, and also shows the senescence initiation zone.

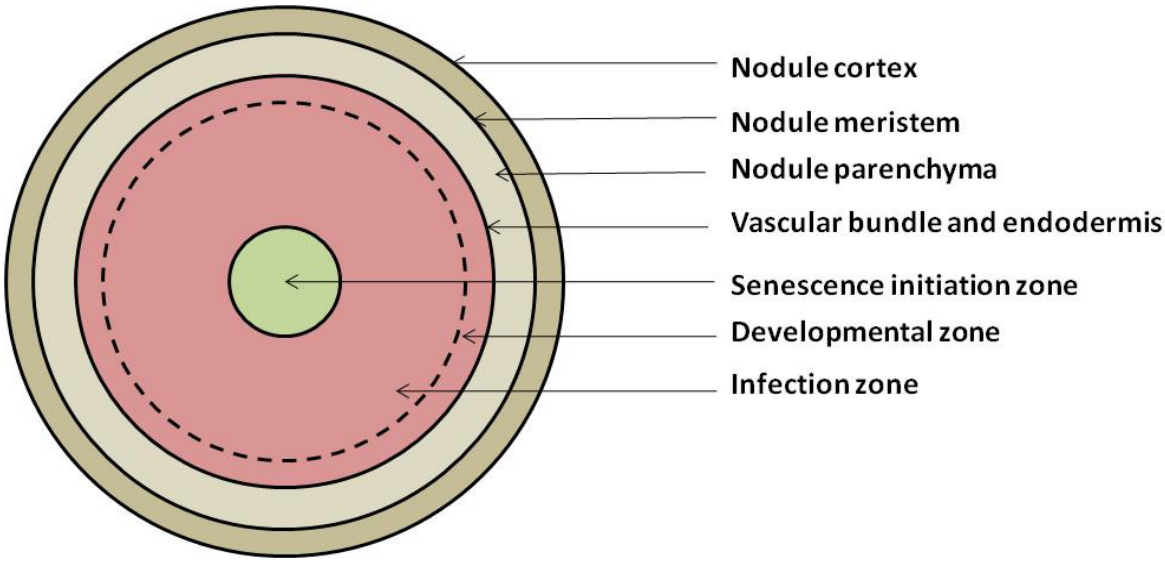


Figure 1.2 Diagrammatic representation of determinate root nodule physiology and morphology. The central infection zone and the inner cortex are represented by the central pink region, which is the primary site of nitrogen fixation. The middle cortex is represented by the light grey area, which includes the vascular bundle, nodule endodermis, nodule parenchyma and nodule meristem, and finally the outer cortex is the outermost layer. Figure adapted from (Luyten and Vanderleyden 2000) and (Hirsch 1992).

Nodulation in soybean occurs after infection of the root-hair cells by *Bradyrhizobium japonicum*. The bacterium is attracted by plant-derived flavonoids, or isoflavonoids (secondary phenolic compounds), at which stage the bacteria releases lipochito-oligosaccharides, called Nod factors, which stimulate a signal transduction cascade leading to differentiation into the root nodules (Long 1996, Long 2001, Oldroyd et al. 2011). This interaction is highly specific between host plant species and compatible Rhizobium strains, which is determined by compatible signal detection. During the infection process (Figure 1.3), bacteria are entrapped by the curling of the root hair due to an influx of Ca^{2+}. The plant cell wall then degrades and the cell membrane folds inward to form the infection thread. Upon entering the infection thread, the bacteria release the nodulin proteins which stimulate mitosis and starts differentiation into bacteroids.

The bacteria enter individual plant cells by endocytosis leading to the formation of the peribacteroid. This plant membrane isolates the bacteria from the plant cell's cytoplasm and creating an artificial environment favourable for nitrogen fixation. While supplying nutrients for bacterial growth and development, the plant cell also supplies a low O_2 environment within the symbiosome. This is essential to the activity of nitrogenase (EC1.18.6.1) expressed by the bacteria (Long 1996, Long 2001, Ferguson and Mathesius 2003, Colebatch et al. 2004, Puppo et al. 2005, Oldroyd et al. 2011).

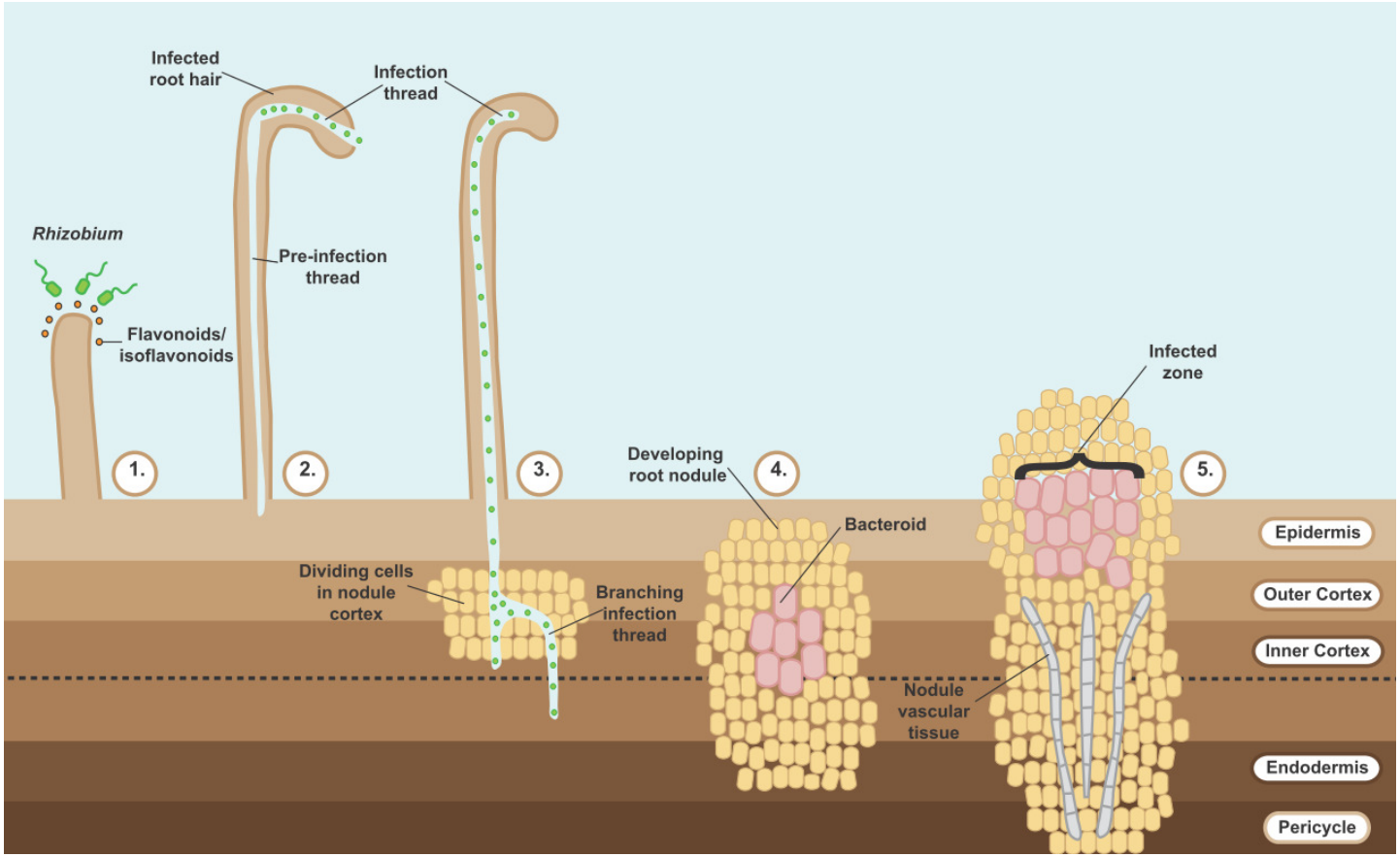


Figure 1.3 Diagrammatic representation of interaction between the root and *Rhizobium* leading to the formation of the nodules. *Rhizobium* bacteria are attracted to the plant's root hair by flavonoids/ isoflavonoids released from the plant (1), the bacteria releases Nod factors which stimulate the formation of the infection thread (2), causing differentiation of the nodule cells (3) and subsequent formation of the root nodules and bacteroids (4). The nodule vascular tissues supply nutrients and resources to the bacteroids, now capable of fixing nitrogen in the newly created artificial environment (5). Figure adapted from (Puppo et al. 2005), (Oldroyd et al. 2011) and A Companion to Plant Physiology, 5th Ed.

Symbiotic nitrogen fixation requires an abundant ATP supply, produced by oxidative phosphorylation and therefore high respiration rates in the infected cells, which requires ample O_2, but since nitrogenase activity is sensitive to the presence of O_2, the supply and presence of O_2 has to be regulated and balanced. The oxygen diffusion barrier in the nodule ensures a low O_2 concentration in the infection zone of the nodule to maintain nitrogenase activity. The carrier molecule, leghemoglobin, is capable of scavenging O_2 allowing a high O_2 flux for rapid gas exchange between a low concentration of O_2 in the infection zone and the interconnecting air spaces of the inner cortex (Appleby 1984, Downie 2005, Ott et al. 2005).

The high respiration rates required for nitrogenase activity is coupled with the carbon metabolism from photosynthesis as main source of energy (ATP) and reducing power ($NADP^+$) for N_2 fixation (Driscoll and Finan 1993, Ladrera et al. 2007, Dupont et al. 2012, Kaschuk et al. 2012). Sucrose is the main carbon source supplied to the nodule and is converted to di-carboxylates, malate and succinate in the uninfected nodule cortex. This hydrolysis reaction is initiated by sucrose synthase (EC 2.4.1.13) or alkaline invertase (EC 3.2.1.26). Uridine di-phosphate glucose (UDP-Glc) and hexoses are also produced from this reaction. These compounds enter the oxidative pentose phosphate pathway after phosphorylation by hexokinases to produce phosphoenol pyruvate (PEP). This is subsequently converted to oxaloacetate, which is subsequently utilised for L-malate synthesis by PEP carboxylase (EC 4.1.1.31) and malate dehydrogenase (EC 1.1.1.37), or potentially to synthesise amino acids as well as amides or ureides (Walsh 1990, Koch 2004, Stacey 2007, Qureshi et al. 2010).

Nitrogenase is an iron (Fe) protein and is a 60kDa homo-dimer protein binding ATP during nitrogen fixation. The iron-molybdenum (MoFe) cofactor protein binds the substrate during the reducing reaction. Nitrogenase uses electron-transfer via nucleotide binding and hydrolysis to catalyse reduction of N_2 to NH_3 (Shah and Brill 1977, Howard and Rees 1996). Ammonia (NH_3) is subsequently converted to ammonium NH_4^+ because NH_3 cannot diffuse across membranes, but can diffuse out of the peribacteroid into the infected cell cytosol. Glutamate synthetase (GS) and glutamate-oxoglutarate amino transferase (GOGAT) assimilates the ammonium into glutamine, which is subsequently converted into asparagines and exported to the xylem from infected cells. Glutamine can also be exported to the plastids and converted by glutamate synthase to amides (asparagines and glutamine) or ureides (allantoin and allantoic acid) in uninfected cells. These ureides are the main form of fixed nitrogen exported from the nodules to above ground parts, e.g. shoot and leaves (Day et al. 2001, Cabello et al. 2009). Remaining glutamate can also be recycled by GS for NH_3 assimilation and main products of SNL are further used to produce nitrogen containing compounds such as nucleic acid and amino acids.

1.3 Nodule senescence

The process of senescence involves degradation of cellular proteins which are re-mobilised with other nutrients to be re-utilised elsewhere in the plant (Callis 1995, Martinez et al. 2007a). Soybean nodule senescence is generally the sequence of biochemical and physiological changes from the fully mature developed nodule state to nodule death (Lim et al. 2003, Puppo et al. 2005). According to the decay model, a point of no return is suggested to exist and passing the threshold, either by age or degree of stress, initiates the senescence

process leading to the breakdown of homeostatic processes (Huffaker 1990, Noodén et al. 1997).

Reactive O_2 species (ROS) are produced by various cellular processes. An unbalanced production, or removal, of ROS results in oxidative stress triggering nodule senescence. Although ROS function in certain signalling pathways, ROS cellular concentration is controlled by antioxidant defences and redox pathways. According to Foyer and Noctor (2005a, 2005b), ROS does not trigger senescence or cell death, but rather functions as a signal, changing gene expression pathways leading to programmed cell death (PCD). Under adverse environmental conditions, a wide range of responses can be found depending on the stress type and degree of stress. This leads to changes in gene expression, metabolic rates, growth and development processes, ultimately changing growth rates and final crop yields. Stress-induced nodule senescence is initiated by a physiological/ biochemical trigger and is faster than natural senescence, which is age dependent, but follows similar progression than natural senescence (Puppo et al. 2005). Nodule senescence causes several changes, from loss of leghemoglobin and nitrogen fixation capacity to enhanced proteolytic activity. Active nodules are further pink in colour due to the heme-group of leghemoglobin. With the onset of senescence, loss of leghemoglobin causes a visual change in nodule tissue colour from pink to green (Roponen 1970). SNF requires high reducing conditions for electron transfer from ferredoxin, uricase and hydrogenase that are susceptible to auto-oxidation resulting in superoxide production (Dalton et al. 1991). Auto-oxidation of oxygenated leghemoglobin to ferric leghemoglobin also produces superoxide radicals, and causes oxidation of bacteroid proteins. Under normal functioning, the array of antioxidant metabolites and ROS scavenging enzymes protect cells from oxidative damage. However, with the onset of senescence, the ascorbate-glutathione antioxidant pool is decreased associated with a decrease in the nodule

carbon to nitrogen ratio in the nodule mediated by the abscisic acid signalling pathway which ultimately activates proteases (Puppo et al. 2005). Martinez et al. (2007b) found that in leaf tissue during senescence 2500 genes were activated with 7% of these genes coding for different types of hydrolases and proteases. Large-scale protease activation, however, causes extensive protein remobilisation, and ultimately degradation of the symbiosome (Vance et al. 1979, Pladys and Vance 1993).

1.4 Proteases and protease inhibitors and senescence

Cellular proteins have a limited life-span and are degraded, and/ or subsequently replaced as part of the homeostatic process, or as part of normal cell death. The degradation of these proteins is performed by the activity of proteases which catalytically hydrolyse specific peptide bonds, cleaving either internally (e.g. endo-proteases) of the target proteins or cleaving from the amino- or carboxy- terminal ends (e.g. exo-proteases) of the target proteins (Lohman et al. 1994, Salas et al. 2008). Proteolytic enzymes are not only involved in cellular death and senescence processes, but also normal cellular development and differentiation of proteins. Proteases can be found in plants, micro-organisms, animals, almost all organisms, and perform a variety of different processes and are integral in various physiological functions (Beers et al. 2004). The classifications of specific functional classes of proteases are determined by the essential amino acid residue in the active site of the enzyme and optimal pH range of activity. Main protease groups are the aspartic-, cysteine-, metallo- and serine proteases, with the most abundant plant proteases being the cysteine proteases (Grudkowska and Zagdanska 2004).

Cysteine proteases are further sub-grouped into families, with the papain-like C1 group and legumain-like C13 group being the most predominant of the cysteine proteases (Oliveira et al. 2003, Salas et al. 2008). The cysteine proteases and their inhibitors, cystatins, play an important role during plant growth and development, and other physiological processes such as hormone signalling, embryogenesis and morphogenesis (Salas et al. 2008). During the germination of cereal seeds (such as wheat) cysteine proteases were the major contributor to the degradation of the storage proteins prolamins (Bottari et al. 1996). In germinating barley seeds 42 proteinases were involved of which 27 were cysteine proteases (Zhang and Jones 1995). Cysteine proteases are also involved in different processes of PCD, e.g. process of tissue-differentiation or different stages of senescence and even in response to abiotic and biotic stress conditions (Grudkowska and Zagdanska 2004). The expression of cystatins was found to coincide with the activation of these signal transduction cascades, perhaps involved in regulating the non-specific activity of the protease enzymes involved in PCD during the hypersensitive response (Solomon et al. 1999, Belenghi et al. 2003). Cysteine proteases also contribute to nutrient remobilization from senescing tissues to actively growing parts of the plant, by dismantling cellular molecules and the organelles (Beers et al. 2000). Beyene et al. (2006) found transcripts encoding the tobacco cysteine protease (*Nt*CP2) expressed in mature leaves had increased under drought conditions, indicating possible involvement in PCD. Cysteine proteases were found to be abundant during senescence in late nodules and was classified as either nodule specific cysteine proteases or nodule enhanced proteases (Lee et al. 2004).

Cysteine protease expression during nodule senescence has been previously reported by several research groups (Pfeiffer et al. 1983, Kardailsky and Brewin 1996, Espinosa-Victoria et al. 2000, Lee et al. 2004, Oh et al. 2004). Proteolytic activity in infected nodules limits the

bacterial symbiosis and nitrogen fixation, with cytosolic leghemoglobin and the bacteroid as targets (Pladys and Vance 1993). In *Medicago trunctula* anti-sense inhibition of the cysteine protease CYP15A caused a delay in nodule senescence (Sheokand et al. 2005) and nodule lifespan was prolonged, when a nodule-specific papain-like cysteine protease (AsNODF32) was silenced (Li et al. 2008). However, despite strong evidence for cysteine protease involvement in nodule development and senescence, only limited detailed information is currently available on any specific cysteine protease inhibitor (cystatin) function and activity in these development and senescence processes (Pfeiffer et al. 1983, Alesandrini et al. 2003a, Oh et al. 2004, Vorster et al. 2013). Benchabane et al. (2010) suggested that regulation of cystatin production at the transcriptional level has a significant function in protein recycling during the process of senescence as well as under abiotic stress conditions. The most detailed analysis of participation of an endogenous cystatin in interaction with an endogenous plant cysteine protease during senescence has been the coordinated expression of the mRNAs of a cysteine protease and a cystatin, in senescent spinach leaves where a senescence-related cysteine protease–cystatin complex was identified (Tajima et al. 2011). Further evidence of the *in vivo* regulation of cysteine protease has been provided by Pillay et al. (2014) demonstrating that co-expression of the rice cystatin OCI in tobacco plants protected recombinant proteins from degradation by lowering overall cysteine protease activity.

For each class of proteases a corresponding class of protease inhibitors exists. Each class has a specific pH optima and specific amino acid residues in its structure required to inhibit targeted proteases (Fan and Guo-Jiang 2005). In plants, cystatins are natural and specific inhibitors of cysteine proteases of the papain C1A family that generally block C1A proteases by a tight and reversible interaction (Chu et al. 2011). Cystatins used for housekeeping purposes and physiological regulation, are thought to have a broad range of expression

patterns, whereas cystatins involved in developmental cues and in stress responses, are thought to have a restricted and preferential expression pattern (Massonneau et al. 2005). The involvement of cystatins in various defence and developmental processes has been inferred based on the expression profiles of these cystatins during these different processes. During seed development, expressed cystatin transcripts accumulate, eliminating protease activity and enabling storage protein deposition. The cystatins continue to protect the deposited proteins until seed germination, at which time cysteine proteases are expressed and storage protein remobilisation is allowed to occur (Benchabane et al. 2010). Cystatins with a carboxy-terminal extension containing a SNSL amino acid motif (SNSL) are able to inhibit legumain-like cysteine proteases of the C13 family (Martinez et al. 2007b). Several cystatin functions have been proposed, but all involve a balanced interplay with a cysteine protease to regulate proteolytic activity (Grudkowska and Zagdanska 2004, Benchabane et al. 2010).

Research has so far provided strong evidence that plant cystatins regulate endogenous protein turnover during growth and developmental processes, including senescence and PCD, and are further involved in accumulation and mobilization of storage proteins (Belenghi et al. 2003, Severin et al. 2010). A further key function is protection against plant pests, such as Coleopteran insects and nematodes, where cystatins prevent cysteine protease activity required for protein digestion in pests (Benchabane et al. 2010, Diaz-Mendoza et al. 2014). The potential of using cystatins to regulate cysteine protease activity and thereby agronomical important traits (such as stress tolerance, delayed leaf senescence, etc.) has been shown. The ectopic expression of OCI, had not only affected the physiology, growth and development (such as retarded stem elongation and leaf expansion) of tobacco (Van der Vyver et al. 2003, Prins et al. 2008), soybean and *Arabidopsis thaliana* plants (Quain et al. 2014), but had also enhanced the tolerance of these plants to abiotic stresses, such as chilling stress, heat stress,

drought stress, high and low light stress (Prins et al. 2008, Demirevska et al. 2010, Quain et al. 2014).

Plant cysteine proteases, and by implication their inhibitors (cystatins), play an important role during plant development processes (Pfeiffer et al. 1983, Espinosa-Victoria et al. 2000). Lee et al. (2004) found cysteine proteases to be abundant during senescence in late nodules and had categorized these cysteine proteases as either nodule specific or nodule-enhanced proteases. Although there are published data available for the expression of cysteine proteases during nodule senescence, there is very limited or fragmented information available about the expression of cysteine proteases during early nodule development (Pfeiffer et al. 1983, Espinosa-Victoria et al. 2000, Lee et al. 2004, Oh et al. 2004) with even less information about expression of cystatins in soybean nodules during these processes. The release of the database for the soybean genome (Schmutz et al. 2010) has now made it possible to identify all the different cysteine proteases and cystatins present in soybean. However, it is not known whether all of these protease and inhibitor sequences are expressed and functional in the soybean proteome and where and when they are expressed. My interest in studying the cysteine protease-cystatin system in greater detail is based on a recent EU-IRSES “LEGIM” project together with University of Leeds and Ghent University which was aimed to extend our knowledge and concepts of nodule development and sustainability.

1.5 Problem statement

Several studies have detected the activities of cysteine protease enzymes in both developing and senescent nodules of legume species (Malik et al. 1981, Kardailsky and Brewin 1996, Alesandrini et al. 2003b, Lee et al. 2004, Oh et al. 2004, Li et al. 2008, Vorster et al. 2013). Only some information is currently available about the expression of some members of the cysteine protease gene family in soybean nodules. The plant cysteine protease-cystatin system plays an important role during plant development processes and is also involved in nodule senescence. The cysteine proteases are very likely to function in protein remobilization and break down of the symbiosome during nodules senescence. However, all the members of the entire gene family has yet to be fully characterized, as some members might have different functions, as well as different expression profiles. Furthermore, the proteases might physically not be able to interact with the inhibitors due to localisation differences or weak interactions. Very limited expression data about the members of the cystatin gene family in soybean nodules is available (Vorster et al. 2013). With the release of the complete genome of soybean, all putative cysteine proteases and cystatin sequences can now be identified. However, it is not known whether all of these putative sequences are expressed, where and when they are expressed, which individual cysteine proteases and cystatins are at work during nodule development and senescence. In previous research a protease activity profile during nodule development has been established and it was found that protease activity, including cysteine protease activity, increases dramatically during early developmental stages as well as during senescence. However, detectable cysteine protease activity, which was completely inhibited by the commonly used cysteine protease inhibitor E64, but could not be blocked by a natural purified plant cystatins from either rice or papaya, raises the question if this interaction with a natural inhibitor is prevented due to the structure

of the inhibitor. Through characterization of the individual components of the protease–protease inhibitor system one might be able, as a possible outcome, either silence a particular cysteine protease or express a specific natural or engineered recombinant cysteine protease inhibitor(s) (Kiggundu et al. 2006) in soybean nodules possibly delaying either natural or stress-induced premature nodule senescence. Unravelling the involvement of key genes and processes, which lead to natural senescence, might give an indication of targets at work at the onset of stress-induced senescence and will be crucial targets for future genetic improvements by breeding programmes or genetic manipulation.

1.6 Hypothesis

The working hypothesis for this PhD study was that members of the soybean cysteine protease and cystatin gene families are differently expressed during soybean nodule development allowing a balanced interplay between individual cysteine proteases and cystatins at work during nodule development with only particular cystatins interacting with senescence-induced cysteine proteases.

1.7 Research aim and objectives

The aim of this investigation was to advance our knowledge of expression, localization and possible function of the cysteine protease-cystatin system during natural soybean nodule development to ultimately apply this gained knowledge to improve soybean nodules to also better withstand premature senescence induced by environmental stresses. To achieve this aim the PhD study had the objectives to (i) characterise the process of natural soybean nodule development including senescence, (ii) to identify all members of the soybean cysteine protease and cystatin gene families in soybean through homology searches, (iii) to analyse the expression of these identified gene sequences during nodule development and senescence using RNA-Seq, (iv) to analyse the interaction affinities between these individual proteins identified expressed concurrently and finally (v) to determine their sub-cellular localisation of these respective proteins.

CHAPTER 2

MATERIALS AND METHODS

2.1 Plant material and RNA preparation

Soybean (*Glycine max* L. Merr.) seeds of the commercial cultivar Prima 2000 were obtained from Pannar Seed in South Africa. Each pot was inoculated with 0.5 g of SoyGro inoculum (SoyGro Bio-Fertilizer Limited) containing *Bradyrhizobium japonicum* of the strain WB74-1 prior to planting in fine vermiculite (Mandoval PC). Plants were grown under a controlled 13 hours photoperiod at a light intensity of 600 $mmolm^{-2}s^{-1}$, with 3 hours of supplementary light from metal-halide lamps and using a day/night temperature of 25°C/17°C and 60% relative humidity. Distilled water was used for plant watering and twice a week plants were watered with a nitrogen-poor nutrient solution (Fenta et al. 2012). This watering regime promotes the symbiotic relationship between the plant and the *Rhizobium* stimulating nodules with high symbiotic nitrogen fixation (Van Heerden et al. 2007). Four crown nodules were harvested from a minimum of three plants at time points 4, 8 and 14 weeks of development and were flash frozen in liquid nitrogen and stored at -80°C until RNA extraction. Three biological replicates were pooled for RNA extraction with the Qiagen RNeasy® kit (Qiagen, Germany). RNA quantity was measured with a Thermo Scientific NanoDrop 2000 with RNA quality analysed on a 2% agarose gel prior to sequencing at Case Western Medical Institute. The Illumina® mRNA-SEQ kit was applied for sample preparations and RNA-Seq libraries were generated with an Illumina® Genome Analyzer II$\varkappa$.

2.2 Transcriptome sequencing, data processing, normalization and data mining

Sequenced RNA was analysed with the Galaxy server [http://galaxy.bi.up.ac.za/] (Bioinformatics Unit, Forestry and Agricultural Biotechnology Institute, University of Pretoria). Glyma1.89 genomic assembly and transcriptome models, available on Phytozome (Schmutz et al. 2010), were used as reference for annotation and mapping of reads. RNA-Seq reads were first converted to a Sanger FASTQ format with FASTQ Groomer (version 1.0.4), FastQC:ReadQC (version 0.52) was applied to assess the read quality scores and FASTQ Quality Trimmer (version 1.0.0) was used to improve the read quality scores (Blankenberg et al. 2010, Cock et al. 2010). Trimmed paired reads were mapped to reference genome with Tophat2 (version 0.6) tool (Kim et al. 2013) and Cufflinks (version 0.0.5) tool was applied to assemble aligned reads into transcript/ exon-isofoms (Trapnell et al. 2010). Figure 2.1 illustrates the various bioinformatic tools used in the PostQC pipeline to analyse the generated RNA-Seq data. The Cuffcompare (version 0.0.5) tool subsequently applied to track transcripts across the time-points (4, 8 and 14 weeks of nodule age) and comparison of assembled transcripts to reference annotation. Finally, the Cuffdiff (version 0.0.5) tool was applied to find significant changes in transcription between the time points (Trapnell et al. 2010). Specific parameters used for each of these tools are listed in Appendix A, Tables A1-3. All other programmes used were with default set of parameters. RNA-Seq data have been made publically available on SoyBase at the following web address http://soybase.org/projects/SoyBase.A2014.01.php.

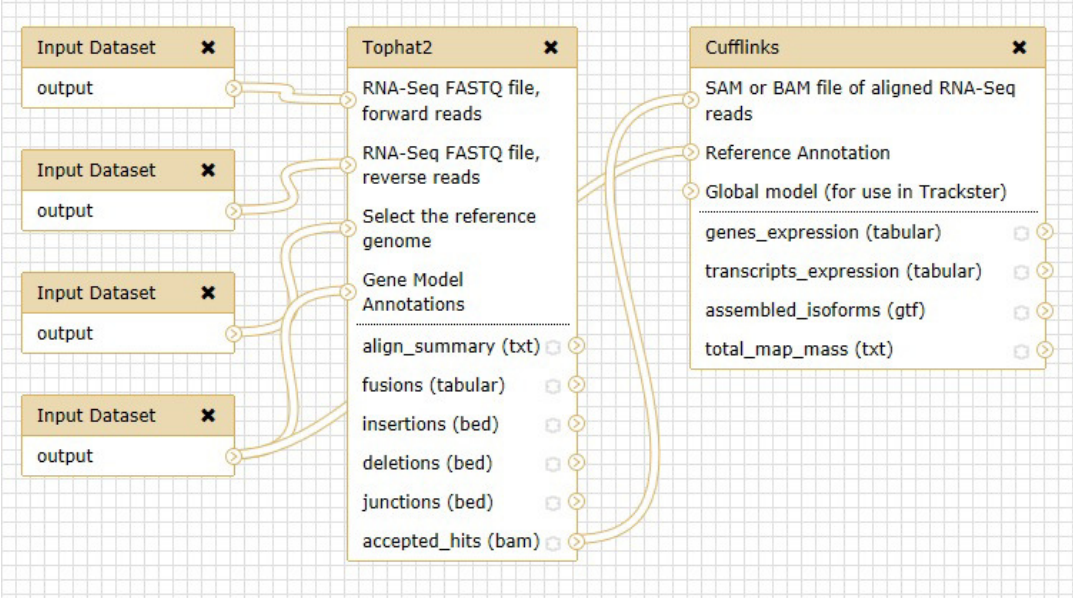


Figure 2.1 The PostQC pipeline applied to map reads to the genome, using the exon model as guide. This pipeline was employed for each of the respective time-points, prior to applying the outputs generated in Cuffdiff to compare expression across time-points.

The software package MapMan (version 3.5.1R2) from the Max Planck Institute for Molecular Plant Physiology (MPIMPP) was utilised to gain maximum information from the RNA-Seq outputs and the available GO resources. Furthermore, the online resource AgriGO (http://bioinfo.cau.edu.cn/agriGO/index.php) was applied for functional classification and annotation of genes. The Plant GO Complete gene ontology was generated using the hypergeometric statistical test method and the Bonferroni correction method for multiple test adjustment, with a significance level of 0.05 and a minimum number of mapping entries of 1. These settings were used throughout all AgriGO analyses.

2.3 Transcript quantification and RNA-Seq validation

Confirmation of transcription data obtained from RNA-Seq experiments was carried out by RT-qPCR on a minimum of 3 biological replicate sample preparations. The RNA samples were prepared as mentioned before, followed by DNase I (1 U/µl) treatment of RNA and cDNA synthesis with the Thermo Scientific RevertAid® First Strand cDNA Synthesis Kit according to manufacturer's instructions. Reverse transcription was carried out in a 20 µl reaction volume with 1 µg RNA, 0.5 µg Oligo(dT)$_{18}$ primer (100 µM) and 1 µl of RevertAid™ M-MuLV Reverse Transcriptase (200 U/µl). The reaction was carried out at 42°C for 60 min prior to inactivation at 70°C for 5 min. Primers for RT-qPCR were designed with the IDT's PrimerQuest Design Tool [http://eu.idtdna.com/PrimerQuest/Home/] and primer sets were applied at 300 nM (Table A.4, Appendix B). The Bio-Rad CFX96-C1000 Touch Lightcycler was applied for thermal cycling with an initial 95°C for 10 min followed by cycling with 95°C for 15 sec, 60°C for 30 sec and 72°C for 30 sec over 40 cycles. Specificity of PCR amplification was confirmed by melting curve analysis (75°C to 95°C) and sequencing of PCR amplicons. Amplicon specificity was screened by BLAST searches to detect any off-targets. Reverse transcriptase negative controls (NTC) were used once for each RNA sample to detect any genomic DNA contamination. All reactions were setup in triplicates. The Bio-Rad CFX Manager v2.1 software was applied for data analysis and calculating C_qs. Any outliers were determined by the Grubbs's test and were removed from subsequent analysis (Burns et al. 2005, Bustin et al. 2009). Housekeeping genes used as reference genes for normalization were ribosomal protein 40S subunit S8 (40S) or elongation factor 1 beta (ELF1) (Mortier et al. 2011) and a SYBR Green I NTCs threshold of C_qs 40 was used. Relative quantification and normalisation was done with the ΔΔCq method and transcript quantification was carried out twice to determine reproducibility. Each standard

curve for each primer set was measured in triplicate and was checked for validity and primer pairs were only accepted if their standard curves had a slope between -3.3 and -3.8. Only R^2 and PCR efficiencies between 90% and 110% ($.90 \leq Cq \leq 1.1$) were accepted (Bustin et al. 2009).

2.4 Identification of cysteine proteases and cystatins in soybean

The Soybean Genome Database [http://soybase.org/], Phytozome Database [http://www.phytozome.net/soybean] and NCBI-BLAST [http://blast.ncbi.nlm.nih.gov/] online resources were searched to identify the entire family of cystatins and cysteine proteases in *G. max*. For identification of cystatin homologues, oryzacystatin-I [PDB: 1EQK_A] from rice was applied as model representative of the I25 family of cysteine protease inhibitors. For identification of cysteine protease homologues in soybean, the cysteine protease papain [E.C.3.4.22.2; GenBank: P00784] from *Carica papaya* was used as model representative for the C1A cysteine protease family. BLASTn, tBLASTx and BLASTp programmes were applied to identify all I25 cystatins and all C1 cysteine proteases with an E-value cut-off of 1E-1.0 to identify homologous gene sequences. Since the database was first accessed during July and November of 2011, the gene nomenclature was maintained to correspond to the Glyma 1.89 reference assembly (Schmutz et al. 2010) which was applied for RNA-Seq read mapping. Gene sequences identified for investigation are listed in Appendix C, Table A.5 and A.6. The FPKM data (Fragments Per Kilobase of exon model per Million mapped fragments) generated were compared to the data of Severin et al. (2010) available online at SoySeq database [http://soybase.org/soyseq/]. Gene sequences were searched for any signal peptides with the online resource TargetP

[http://www.cbs.dtu.dk/services/TargetP/] to determine any cellular localisation, results are summarised in Appendix D, Table A.7.

2.5 Phylogenetic analysis of cysteine proteases and cystatins

Full-length protein sequences for each of the cystatins and cysteine proteases were aligned and phylogenetic trees were generated with the CLC Main Workbench v6.7.1. The Neighbour Joining algorithm was applied with 100 Bootstrapping replicates. Model representative sequences for the different cystatin subfamilies identified by Martinez et al. (2009) were included for phylogenetic analysis: Hv-CPI1 (CAA72790), Hv-CPI2 (CAG38123), Hv-CPI3 (CAG38124), Hv-CPI4 (CAG38130), Hv-CPI5 (CAG38126), Hv-CPI6 (CAG38127), Hv-CPI7 (CAG38131), Hv-CPI8 (CAG38129), Hv-CPI9 (CAG38125), Hv-CPI10 (CAG38128), Hv-CPI11 (CAG38132), Hv-CPI12 (CAG38133), Hv-CPI13 (CAG38134), as well as Monellin cystatin (At5g47550), Cystatin A (At2g40880), Cystatin B (At3g12490), Cystatin 2 (At2g31980) and a representative of the I25B cystatin from *Vigna unguiculata*. Out-group for the cystatin phylogenetic analysis consisted of papain.

Model representative sequences for the eight different cysteine proteases subfamilies described by Richau et al. (2012) were RD21A (At1g47128), RD21B (At5g43060), RD21C (At3g19390), RDL2 (At3g19400), XBCP3 (At1g09850), XCP1 (At4g35350), XCP1 (At1g20850), THI1 (At1g06260), SAG12 (At5g45890), RD19A (At4g39090), RD19B, (At2g21430), RD19C (At4g16190), AALP (At5g60360). ALP2 (At3g45310) and CTB3 (At4g1610) were also included in the phylogenetic trees to infer possible functional activity of the proteases. Out-group used for the C1 cysteine protease phylogenetic analysis was OCI (Os01g58890) and a further I25B cystatin from *V. unguiculata* (Q06445).

2.6 Recombinant cystatin expression

Gene sequences for selected cystatins (Glyma04g10360, Glyma07g39590, Glyma08g11210 and Glyma13g27980 as well as each of the domains from Glyma14g04260, Glyma15g36180 and Glyma18g12240) were synthesized by GenScript (USA). Sequences were synthesized with a 5'-*Bam*HI and 3'-*Eco*RI restriction enzyme cut site for subsequent sub-cloning. Gene sequences of remaining cystatins (Glyma05g28250, Glyma13g04250, Glyma14g04250, and Glyma20g08800) were isolated from cDNA preparations with gene specific primers listed in Appendix E, Table A.8. Forward primers had a 5'-*Bam*HI restriction enzyme site and reverse primers had a 3'-*Eco*RI restriction enzyme recognition site for sub-cloning. Identified putative gene sequences were cloned into the plasmid p*GEX*-3X (Amersham Pharmacia Biotech, UK) as *Bam*HI-*Eco*RI fragments and the *E. coli* strain BL21 (DE3) (Invitrogen, USA) was used for recombinant cystatin expression. All chemicals for bacteria culturing and the GenElute™ plasmid extraction kit for plasmid preparations were sourced from Sigma Aldrich (UK). All molecular biology enzymes, e.g. polymerases used for PCR isolation of gene sequences and enzymes used for cloning were sourced from Thermo Scientific (USA). Thermo Scientific GSH-agarose was applied during the protein purification procedure and Factor Xa (NEB, UK) was applied during the recombinant protein purification process. These products were used according to manufacturers' instruction. Analysis of protein preparations throughout the recombinant protein expression process was done by SDS-PAGE (Laemmli 1970) and protein quantification was carried out with a commercial protein determination assay (Bradford 1976).

2.7 Determination of *Ki* values

The fluorogenic substrate Z-Phe-Arg-MCA (cathepsin L-like substrate from Sigma-Aldrich) was used at 10 µM final concentrations from a 400 µM stock dissolved in DMSO (Sigma-Aldrich, UK). Papain (Sigma; EC 3.4.22.2, UK), cathepsin-L (Sigma; EC 3.4.22.15, UK) and cathepsin-B (Sigma; EC 3.4.22.1, UK) were applied as protease standards. Z-Phe-Arg-7-amino-4-methylcoumarin (Z-FR-AMC) and Z-Arg-Arg-7-amido-4-methylcoumarin (Z-RR-AMC) were used as cysteine protease substrates to assay for cathepsin-L and cathepsin-B like activity (Z-FR-AMC / Z-RR-AMC), cathepsin-F (Z-FR-AMC), cathepsin-H (Z-RR-AMC) and cathepsin-L (Z-FR-AMC) cysteine protease activity. Cysteine protease activity was determined and the K_i values for each of the different recombinant cystatins were determined. Dissociation (inhibition) constants (K_i) for the interaction between the different recombinant cystatins, with model cysteine proteases were determined according to Goulet et al. (2008). Substrate hydrolysis progress curves were monitored as described by Salvesen and Nagase (1989), the linear equation was determined as described by Henderson (1972). Papain (pH 7.0), cathepsin L (pH 5.5) and cathepsin B (pH 6.0) activity was measured in 50 mM sodium phosphate buffer, 4 mM EDTA and 8 mM L-cysteine at their respective enzyme pH optima and hydrolysis was at 25°C. Cysteine protease activities were determined with a Fluostar Galaxy fluorimeter (BMG, Germany) with a 360 nm excitation filter and a 450 nm emission filter. K_i values were calculated, by firstly empirically estimating the K_i (app) and K_m values, and then using the equation $K_i = K_{i(app)}/(1 + [S]/K_m)$. K_m values were 13.6 µM for papain, 2.0 µM for cathepsin B and 1.0 µM for cathepsin L (Goulet et al. 2008). The slope per sec (FU/sec) was calculated with the MARS Data Analysis Software v2.10 (BMG, Germany). E-64 (Sigma-Aldrich, UK) was applied as a broad spectrum inhibitor (positive control) for cysteine proteases at a concentration of 10 nM (Michaud et al. 1993). Different concentration

of the model cystatin OCI was first tested to obtain a reduction in proteolytic activity of 40-60% under assay conditions and an identical concentration was used to assay inhibitory potency of the different recombinant soybean cystatins. The blank is represented by the slope/sec of buffer and substrate without enzymes, whereas the negative control is represented by the slope/sec of the uninhibited protease standards. All reactions were carried out in triplicate and averaged.

2.8 Measurement of cystatin potency

Total plant protein extracts were applied as sources for cysteine protease activity in assays to measure cystatin potency. Extracts were prepared from soybean crown nodules corresponding to the different time points (4, 8 and 14 weeks). Nodules were homogenised by crushing in liquid nitrogen and 50 mM sodium phosphate buffer, pH 6.0 was added in a 1:3 ratio (50 mg : 150 µl; sample - buffer). The solution was incubated for 30 min on ice before centrifugation at 15000 g for 15 min at 4°C to remove any debris. The supernatant was removed, the total protein concentration determined, and a total of 100 ng protein was used per enzyme reaction. The IC_{50} for OCI was calculated as the concentration of cystatin required to achieve 50 % inhibition of proteolytic activity. The other cystatins were then used at this same concentration to compare cystatin potency. All reactions were carried out in technical triplicates and averaged. The protease activity measured was expressed as percentage relative to absence of inhibitor.

2.9 Nodule histology and immunohistochemistry

Plants were grown to the time points as outlined above, crown nodules were harvested and processed according to methods outlined by Johansen (1940), Jensen (1962), O'Brien and McCully (1981). Firstly, tissues were fixed with a solution of FAA (5 % formaldehyde, 5 % acetic acid and 90 % ethanol) for a minimum of 24 h. The tissue was dehydrated by increasing concentrations of ethanol as follows, 30 % ethanol for 8 – 12 h, 50 % ethanol for 8 – 12 h, 70 % ethanol for 8 – 12 h and 100 % ethanol for 8 – 12 h which was done twice. The ethanol was subsequently extracted with incremental concentrations of xylene as follows, 30 % xylene in 100 % ethanol for 8 – 12 h, 50 % xylene for 8 – 12 h, 70 % xylene for 8 – 12 h and 100 % xylene for 8 – 12 h. Wax pellets were then dissolved in the solution at 60°C and repeatedly removed until all xylene was displaced from the tissues. The samples were kept at 60°C until mounted on a microtome slide. The mounted nodules were sliced in 10 µm cross sections with a Reichert-Jung 2040 rotary microtome. Haupt's adhesive solution was used to mount the cross sections to the microscope slides and fixed in formaldehyde vapours. The slides were stained as follows: submerged in a 2.5 % Safranin-O solution, prepared with ethanol, for 30 min; rinsed with water and then submerged in incremental concentrations of ethanol for 5-10 sec, firstly 30% followed by 70 % and finally twice in 100 % ethanol. The slides were then dipped into a 1 % fast green solution, prepared with ethanol, for 30 sec and de-stained in xylene solutions of increasing concentration, prepared with ethanol. Each step was about 5-10 sec, firstly in 30 % xylene, followed by 50 %, 70 % and finally 100 %. For the crystal violet staining, 2 g of crystal violet was dissolved in 20 ml 95 % ethanol, prior to dipping the slides for 1 min followed by de-staining in running distilled water. Eukitt mounting medium, from Agar Scientific, was applied to fasten the coverslip to the slide. An Olympus SZX2-TR30 bright field microscope was used for preliminary investigation of the

stained nodule cross sections, followed by investigation with a Carl Zeiss Axioskop EL 451485 bright field microscope on a 10X magnification and photos were taken with the Carl Zeiss AxioCam ICc5.

For the detection of targeted proteins with antibodies, slides with mounted cross sections were first de-paraffinated and rehydrated, followed by a HIER (heat induced epitope retrieval) step and finally followed by a typical Western blotting approach. The approach followed was done according to the technical guideline of ABCAM IHC-Paraffin Protocol and the protocol of NordiQC Epitope retrieval: Demasking antigens (NordiQC 2005). Mounted cross sections were de-paraffinated and rehydrated as follows: washed twice in xylene for 3 min, then 3 min in a 1:1 xylene to 100 % ethanol mixture, then twice in 100 % ethanol for 3 min each, followed by 95 % ethanol for 3 min, 70 % ethanol for 3 min, and 50 % ethanol for 3 min and finally rinsed with tap water. The de-paraffinated and rehydrated slides were transferred to the antigen retrieval buffer (10 mM Tris, 1 mM EDTA, 0.05 % Tween 20, pH 8.0) and left at 60°C overnight. The following day slides were submerged in TBS (20 mM Tris, 150 mM NaCl, pH 7.6) containing 1 % (w/v) BSA for 2 h at room temperature with gentle agitation. The slides were rinsed 3 times in TBST (20 mM Tris, 150 mM NaCl, 0.025 % Tween 20, pH 7.6), each for 5 min, by gentle agitation. The primary antibody solutions were prepared with the respective antibodies (Rabbit anti-OCI (Van der Vyver et al. 2003)/ Rabbit anti-R79 (Vincent and Brewin 2000)) in a 1:100 dilution in TBS with 1% BSA and antibody binding was done at 4°C for 16 – 24 h with slight agitation. The slides were again rinsed 3 times in TBST, each for 5 min, by gentle agitation. The secondary antibody, goat anti-rabbit alkaline phosphatase conjugate, was prepared with in a 1:1000 dilution in TBS with 1 % BSA and antibody binding was done at room temperature for 1 h with slight agitation. The slides were again rinsed 3 times in TBST, each for 5 min, by gentle

agitation. The AP conjugates were detected using the Bio-Rad AP conjugate kit according to manufacturer's instructions for 5 min. An Olympus SZX2-TR30 bright field microscope was applied for preliminary investigation of the stained nodule cross sections, followed by investigation with a Carl Zeiss Axioskop EL 451485 bright field microscope on a 10X magnification and photos were taken with the Carl Zeiss AxioCam ICc5.

2.10 Statistical methods and analysis

To determine significant transcription changes in the RNA-Seq data, a False Discovery Rate (FDR) of 0.05 was applied and significance in change was determined after correction with the Benjamini-Hochberg correction for multiple-testing. For generation of heat maps with the MeV software package (Saeed et al. 2003), the Pearson's correlation coefficient was used. A one-way ANOVA with Bonferroni post-tests was performed with GraphPad Prism Software version 5.00 for Windows (www.graphpad.com).

CHAPTER 3

RESULTS

3.1 Transcriptome analysis and pathway enrichment

The transcriptome of soybean nodule development at time points 4, 8 and 14 weeks was investigated. Three biological replicates for each time point (4w, 8w and 14w) were produced and pooled. The RNA sequencing produced a total of ~40 million paired-end reads for each time point (Table 3.1). The average read length from the Illumina Genome AnalyzerIIx, was 58 bp before read QC filtering and trimming based on PHRED scores. The per base sequence quality, the per sequence quality score, the per base sequence content, the per base GC content, the per sequence GC content, the per base N content, the sequence length distribution, the sequence duplication, the overrepresented sequences and the overrepresented Kmer, quality parameters were assessed and the appropriate action was taken until acceptable ranges for each parameter was obtained (Figures 3.1-3.4). The soybean transcriptome was expected to have high sequence duplication due to the genome's paleopolyploidy nature, with 75 % of the genome present in multiple copies (Schmutz et al. 2010, Severin et al. 2010). After QC steps were taken, the usable read length was 44 bp and the total reads generated was within acceptable ranges (>85 % of bases had Q score > 30) for the Illumina® Genome Analyser IIx sequencing platform. From these reads that passed the QC process, > 96 % of reads from each time point, were uniquely mapped to the reference assembly.

Table 3.1 The paired-end read quality statistics obtained for each time point after read quality was assessed and processed.

Time point	PostQC read length	Total reads PostQC	Total bases (% aligned)
4w	44 bp X 44 bp	40 823 386 (x2)	1 823 107 818 (96.6 %)
8w	44 bp X 44 bp	43 520 791 (x2)	1 939 048 070 (97.1 %)
14w	44 bp X 44 bp	42 777 586 (x2)	1 908 272 774 (97.1 %)

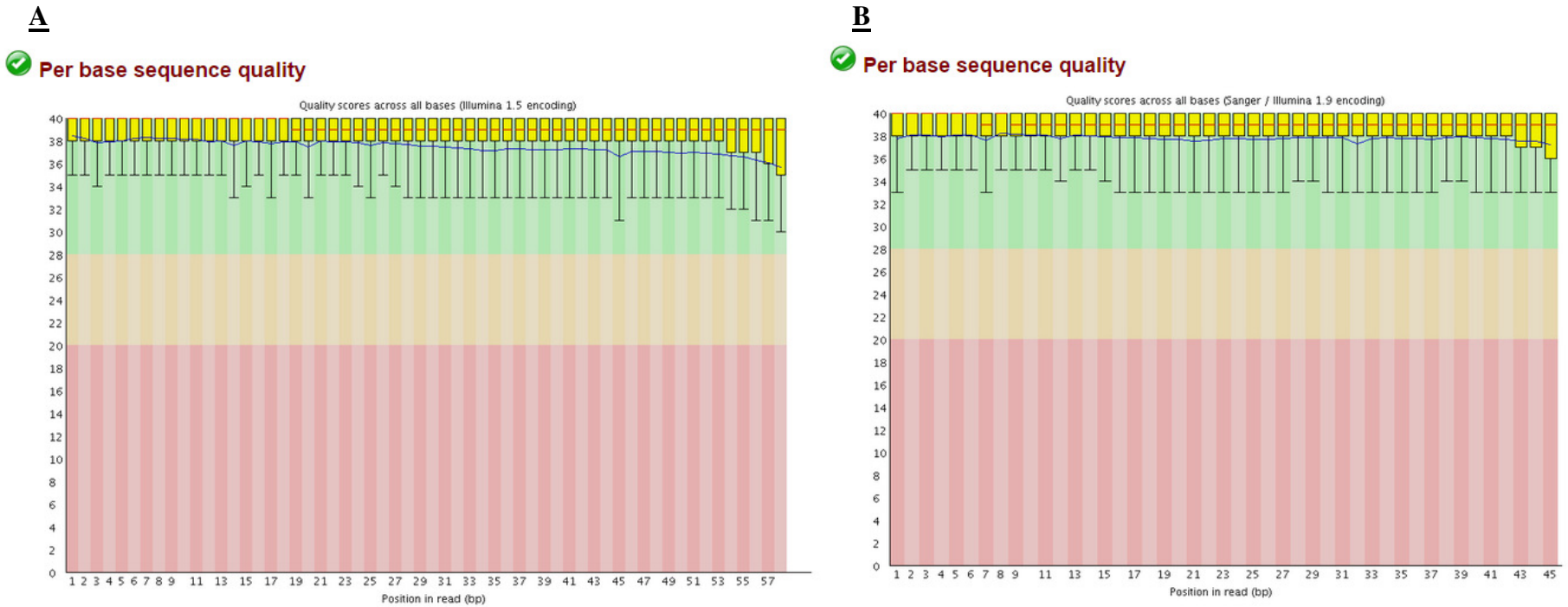


Figure 3.1 Typical per base sequence quality scores obtained from FASTQC, prior to read quality improvement (A) and after read quality improvement (B). The position of each base is indicated on the x-axis and the quality score at each position is plotted on the y-axis. The median score per base is indicated by the red line and the standard error bars are indicated in black whiskers in both directions.

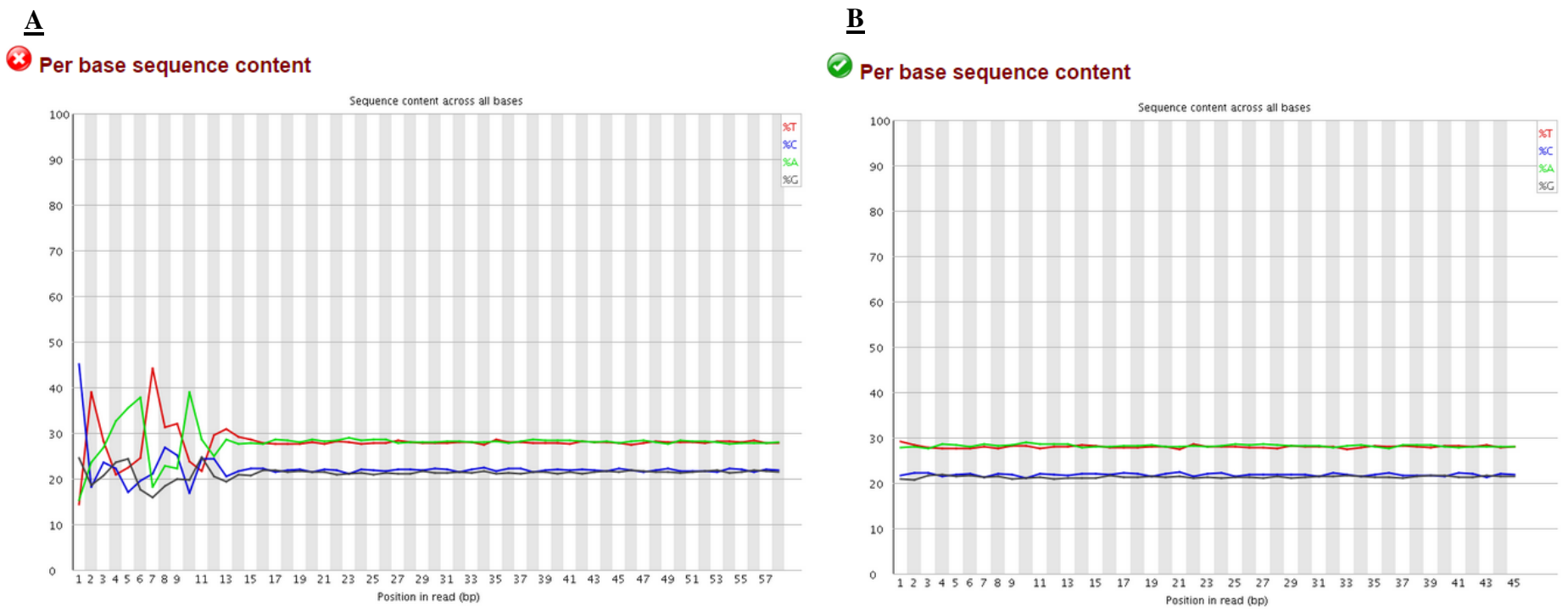


Figure 3.2 Typical per base sequence content quality scores obtained from FASTQC, prior to read quality improvement (A) and after read quality improvement (B). The position of each base is indicated on the x-axis and the percentage per base is plotted on the y-axis.

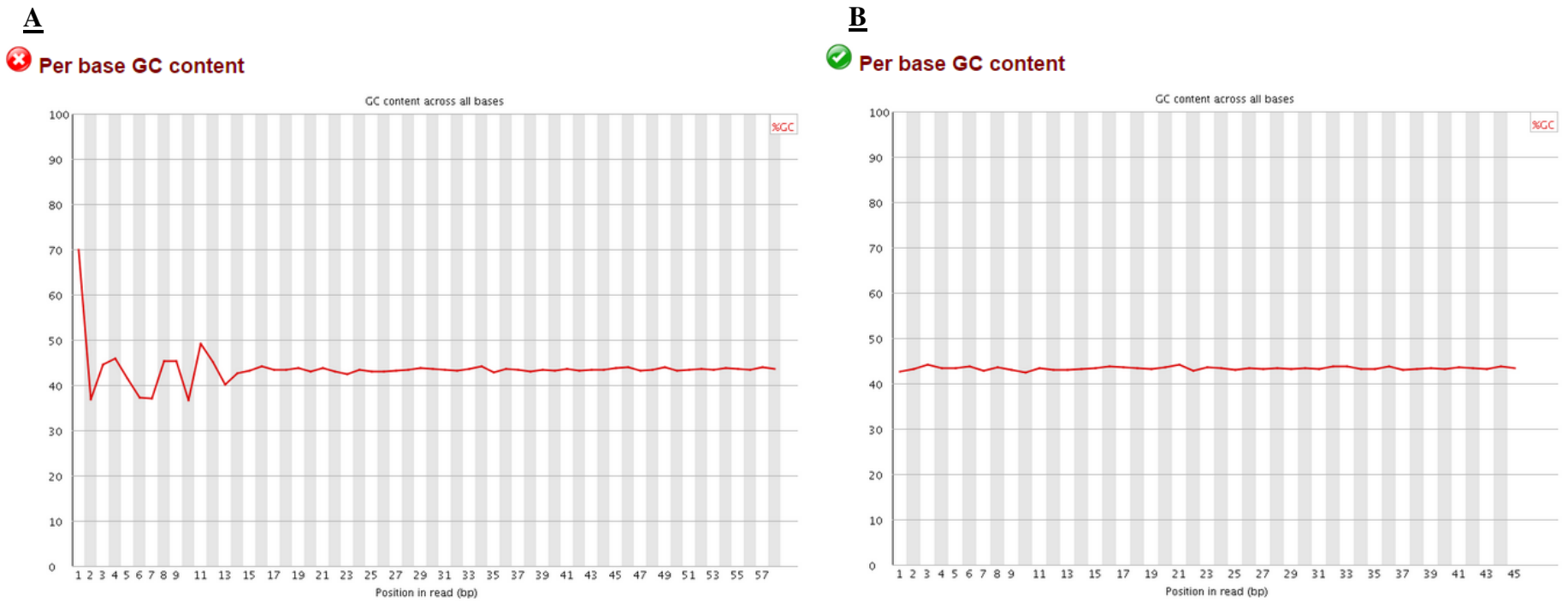


Figure 3.3 Typical per base GC content quality scores obtained from FASTQC, prior to read quality improvement (A) and after read quality improvement (B). The position of each base is indicated on the x-axis and the percentage GC content at each position is plotted on the y-axis.

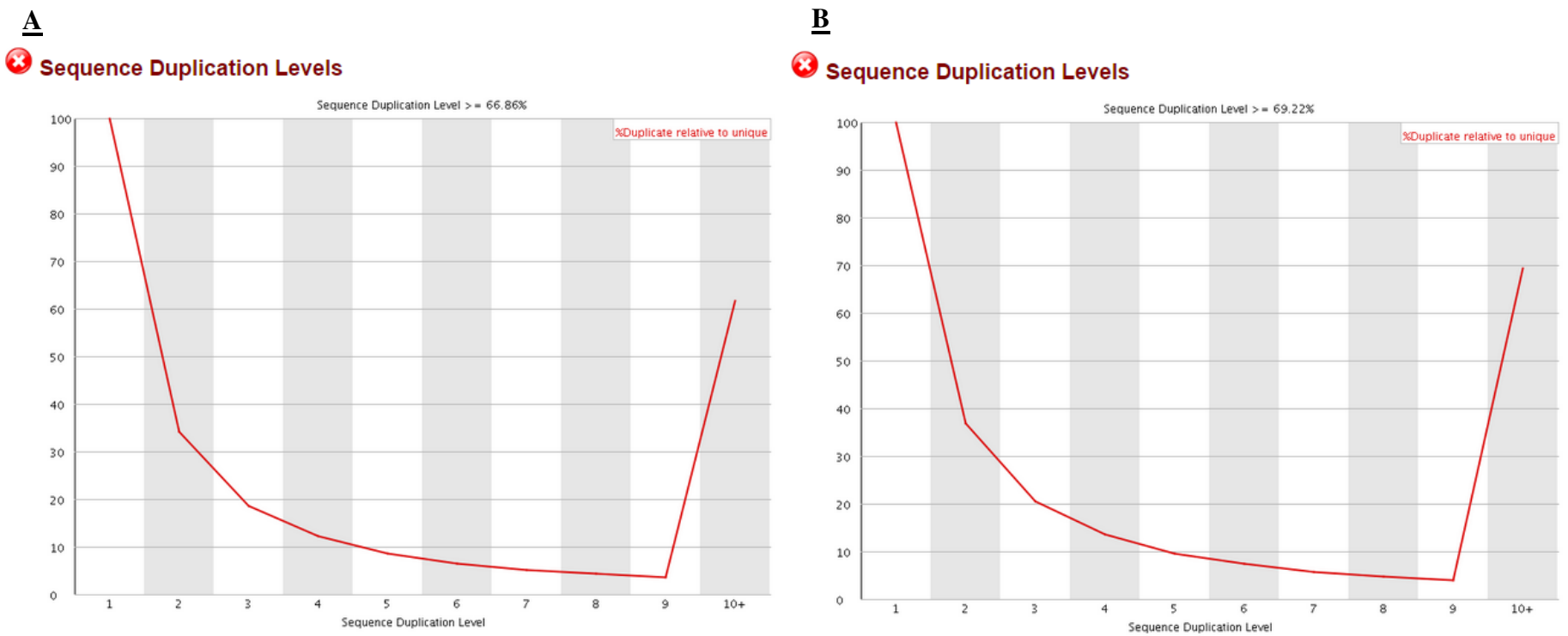


Figure 3.4 Typical sequence duplication levels obtained from FASTQC, prior to read quality improvement (A) and after read quality improvement (B). The level of sequence duplication is indicated on the x-axis and the number of sequences with a given duplication is plotted on the y-axis.

Gene expression unique to each respective time-point as well as overlapping gene activities between time-points with a 0.5 FPKM cut-off applied (Warden et al. 2013) are shown in Figure 3.5. Of the 54176 genes in the soybean reference assembly, 2.44 % (1321) were specific to 4 weeks, 2.23 % (1209) to 8 weeks and 3.71 % (2009) were specific to 14 weeks of nodule development. Approximately 4.39 % (2377) of genes were present in both 4 and 8 weeks, 2.6 % (1410) of genes were present in both 8 and 14 weeks, 1.66 % (899) was present in both 4 and 14 weeks and finally 77.19 % (41819) of genes were present at all 3 time points of nodule development with expression > 0.5 FPKM.

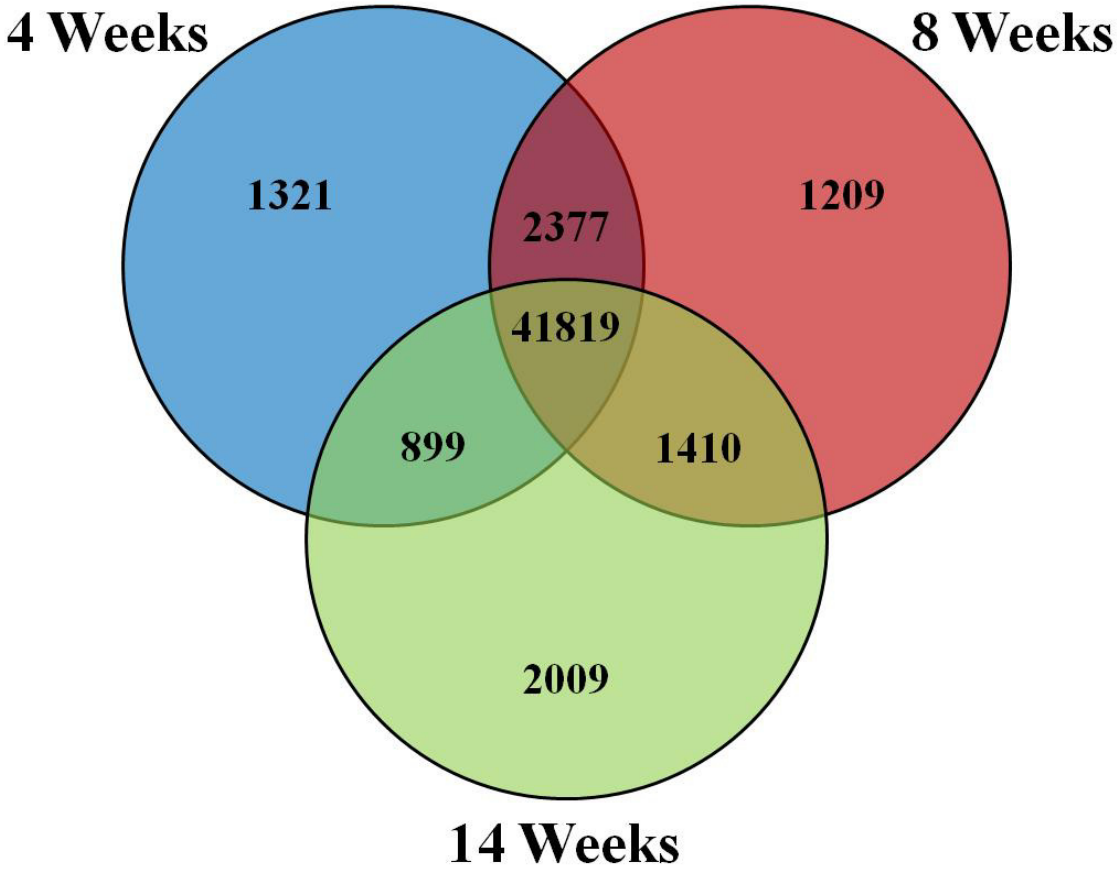


Figure 3.5 Venn diagram of unique and overlapping genes with an FPKM < 5. Values generated with MapMan (v3.5.1R2) and figure generated with Microsoft® PowerPoint.

Genes with low coverage were detected with FPKM values close to 0, but with high-fold changes between time-points. This results in a false impression of gene expression and introduces a bias into the results, as the true expression of these genes might not differ as greatly, or even be, physiologically relevant (Warden et al. 2013). Furthermore, a single fragment mapping to a gene region would not necessarily equate to a single transcript, therefore a more modest estimate of full length transcripts was applied. A gene was considered transcriptionally active if a FPKM ≥ 5.0 was obtained (Trapnell et al. 2010) at any of the time points. If a gene was not transcriptionally (FPKM < 5.0) active in at least one of the 3 time points, the gene was considered to be inactive in the nodules. Figure 3.6 shows gene expression unique to each respective time-point as well as overlapping activities between time-points. From the 54176 genes in the soybean reference assembly, 1.68 % (908) was specific to 4 weeks, 0.96 % (522) specific to 8 weeks and 6.60 % (3573) was specific to 14 weeks of nodule development. About 2.55 % (1383) of genes were present in both 4 and 8 weeks, 2.85 % (1543) present in both 8 and 14 weeks, 1.56 % (843) in both 4 and 14 weeks and again 77.19 % (18972) of genes were present at all 3 time points of nodule development with expression ≥ 5 FPKM. The total number of "active" genes in developing nodules (4 weeks of development) was 40.8 % (22106), in mature nodules (8 weeks of development) 41.38 % (22420) and in senescing nodules (14 weeks of development) was 46.01 % (24931).

For the analysis of the differential gene expression a cut-off of > 1.5-fold-change and a FDR of < 0.05 was used as suggested by Warden et al. (2013). Although several genes, not meeting these criteria are still physiologically relevant, focussing on the genes meeting these criteria ensured statistically sound results and identification of genes strongly involved in the processes of nodule development and senescence.

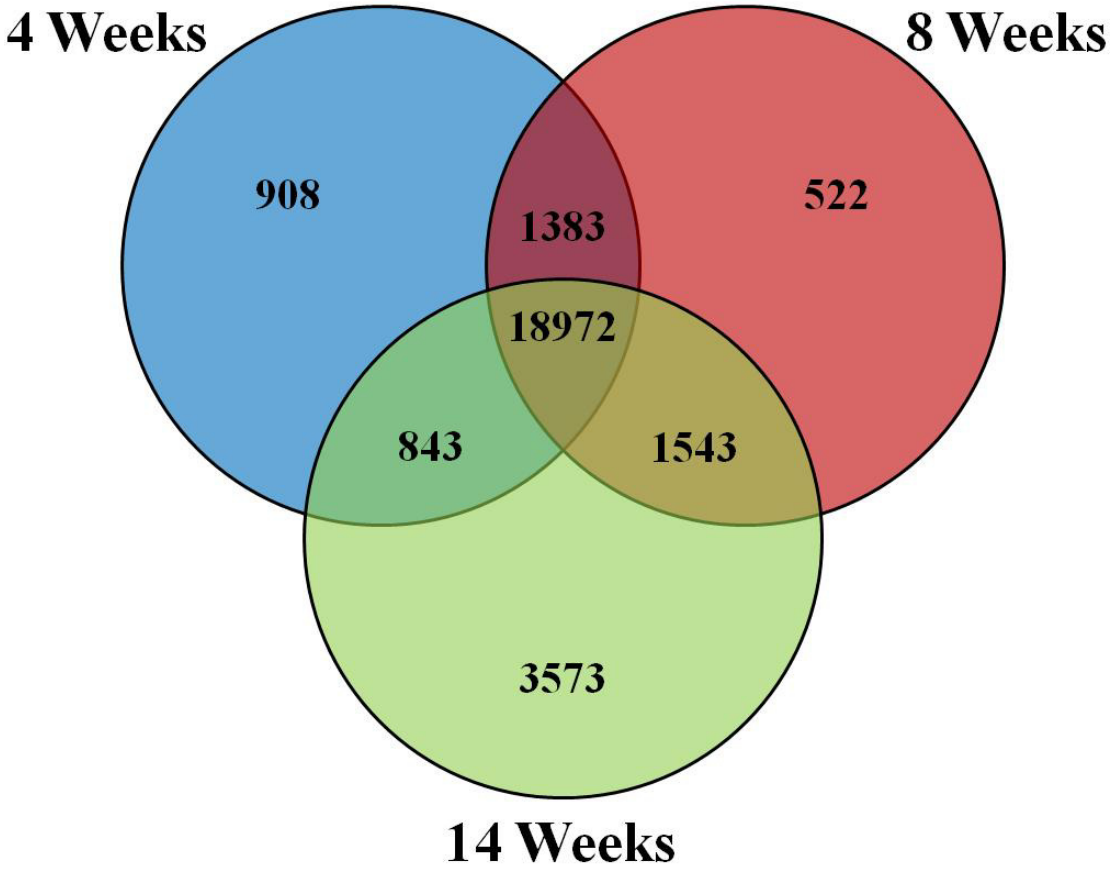


Figure 3.6 Venn diagram of unique to each time-point as well as genes overlapping activity with FPKM ≥ 5. Values generated with MapMan (v3.5.1R2) and figure generated with Microsoft® PowerPoint.

Figure 3.7 shows the summary of genes that had significant changes differential expression across the respective time-points. From the 54176 genes in the soybean reference assembly, 0.04 % (22) of these genes had significant changes in gene expression from 4 to 8 weeks of nodule development, with only 8 (0.01 %) genes unique to the transition from 4 to 8 weeks and 14 (0.02 %) are in common between the transition from 4 to 8 weeks and 4 to 14 weeks. A hundred and seventy (170) genes (0.3 %) showed significant changes in gene expression from 8 to 14 weeks of nodule development, with 52 genes (0.09 %) unique to the transition from 8 to 14 weeks and 118 genes (0.2 %) in common between the transition from 8 to 14

weeks and 4 to 14 weeks of nodule development. Finally, 328 genes (0.6 %) of genes showed significant changes in expression from 4 to 14 weeks of nodule development, of which 196 (0.35 %) of genes were unique to the transition from 4 and 14 weeks of nodule development. These Venn-diagrams (Figures 3.5, 3.6 and 3.7) graphically represent the relative changes in gene expression between the respective stages of nodule development, mature nodules and nodule senescence.

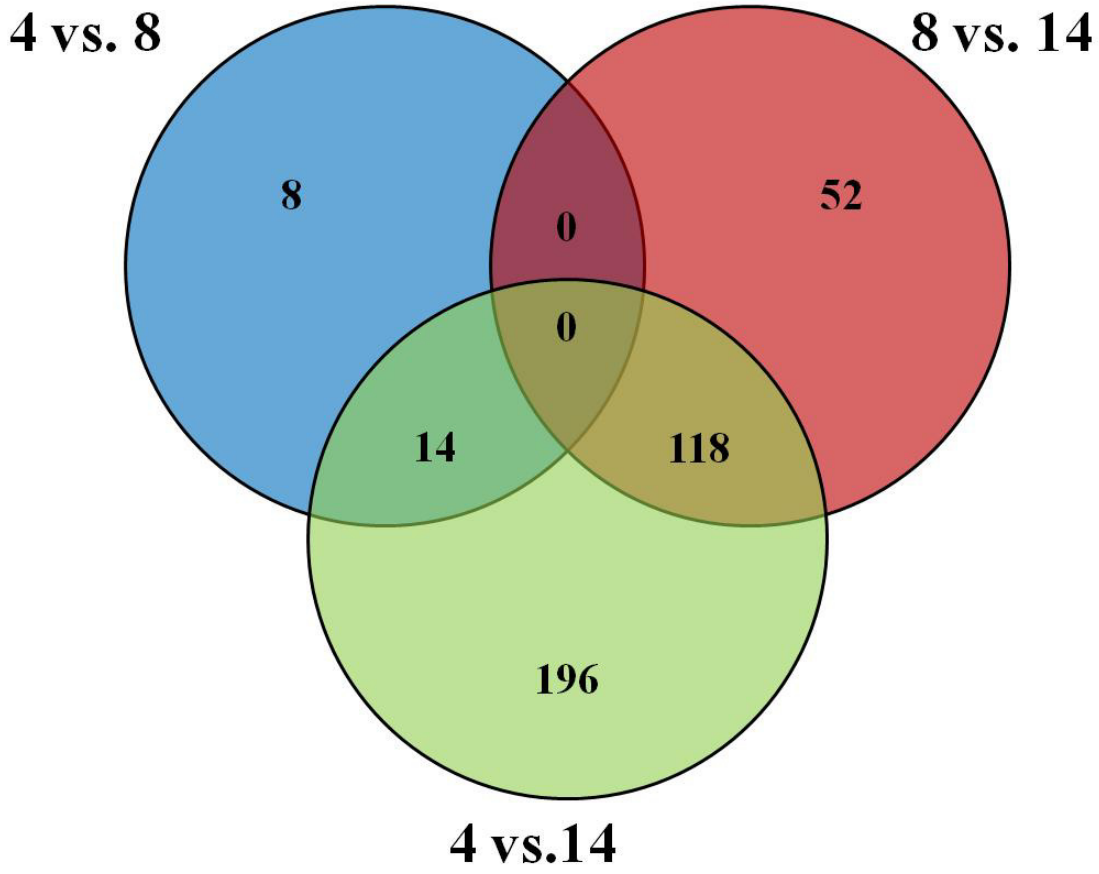


Figure 3.7 Venn diagram of genes showing significant changes in gene expression between the respective time-points. Values generated with MapMan (v3.5.1R2) and Figure generated with Microsoft® PowerPoint.

To determine which GO terms are significantly over-represented in each subset of genes, AgriGO (http://bioinfo.cau.edu.cn/agriGO/) was used to identify the most important biological processes, cellular compartments and molecular functions involved at each time point of nodule development and senescence (Du et al. 2010). This was done for the subsets of genes that were unique to each respective time-point, as well as for each subset of genes which overlap between each time-point and overlap with all of the time-points. This allows for the construction of complex network maps of these various biological processes, cellular components or molecular functions involved. It also shows how the different processes crosslink and influences one another. In case no gene ontology data was available, the best Arabidopsis TAIR10 hits were used. But if none were found, the PFAM > Panther > KOG > KEGG ontology data were applied to possibly elucidate a gene function.

Bar charts (Figures 3.8, 3.10 and 3.12) show over-represented terms from the three categories, biological process, cellular compartment and molecular function. The X-axis indicates the specific GO term and grouped with a bracket according to each category. The Y-axis indicates the percentages of genes mapped to each GO term (shown in green), calculated from the number of genes mapped to the specific GO term divided by the total number of genes mapped in the input list. The reference list is generated with the same calculation from a Background/Reference list compiled by AgriGO from data collected from the Phytozome database *G. max* reference assembly v1.1, which was generated from leaf tissue from 2-week old plants and after etiolation for 5 days prior to harvest. The reference list GO term percentages will not be considered in comparisons as the different tissue types (leaf vs. nodules) and ages of tissues (2 weeks vs. 4 or 8 or 14 weeks) are unfortunately not directly comparable.

The hierarchical tree graphs (Figures 3.9, 3.11 and 3.13) generated by SEA (singular enrichment analysis) indicate over-represented GO terms in each of the three categories, biological process, cellular compartment and molecular function. The ranking of GO terms is set from top to bottom, with each box representing a GO term labelled with GO ID, term definition and statistical information. Non-significant terms are indicated in white boxes with significant terms (adjusted $p \leq 0.05$) indicated according to the level of enrichment and are positively correlated to the degree of colour saturation. Furthermore, solid, dashed, and dotted lines indicates two, one and zero enriched GO terms at both ends connected by the line, respectively. Red arrows indicate positive regulation of dependent GO terms and green arrows indicate negative regulation of dependent GO terms.

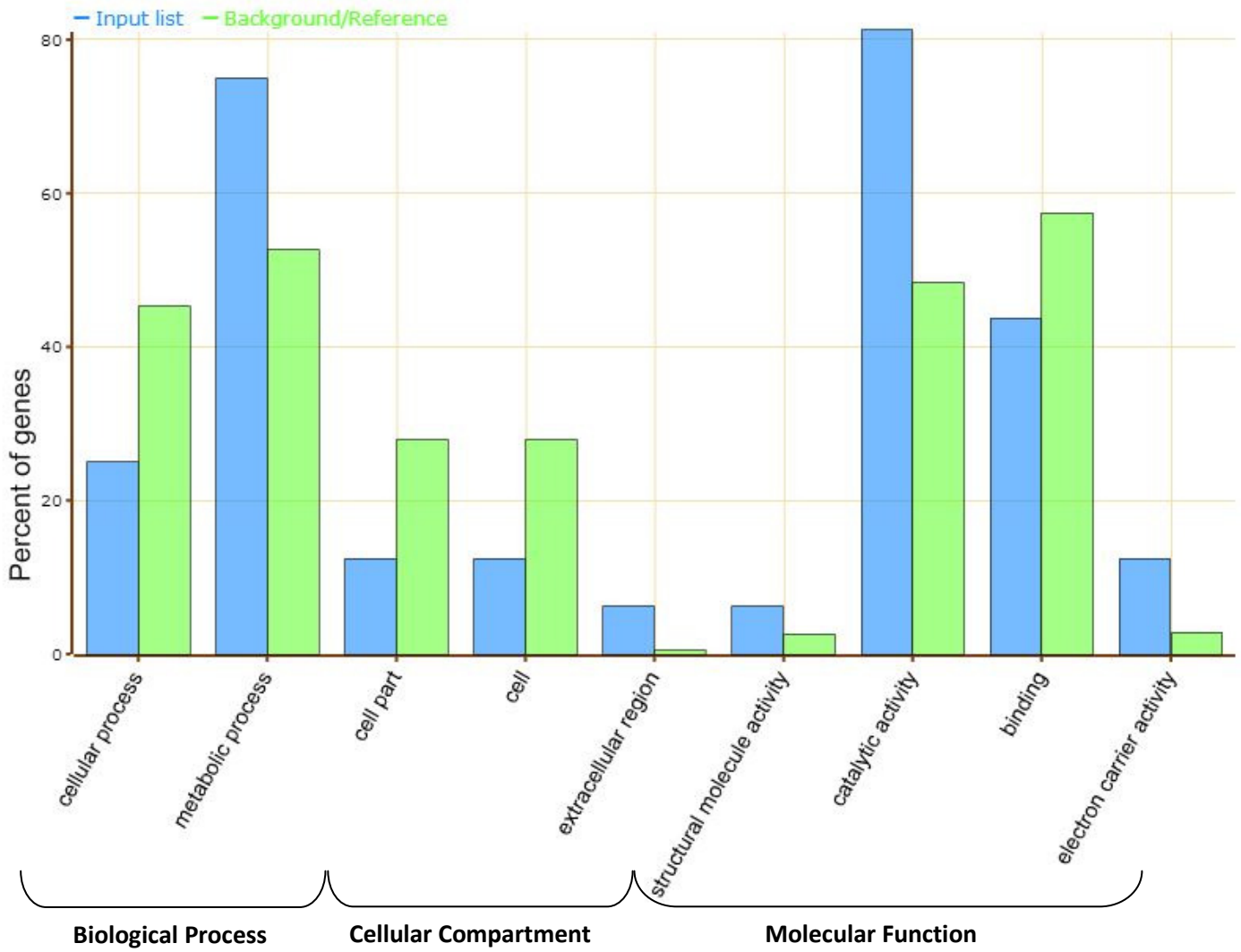


Figure 3.8 Overrepresented GO terms from the dataset of genes with significantly changed expression from 4 to 8 weeks of nodule development from the three categories, biological process, cellular compartment and molecular function.

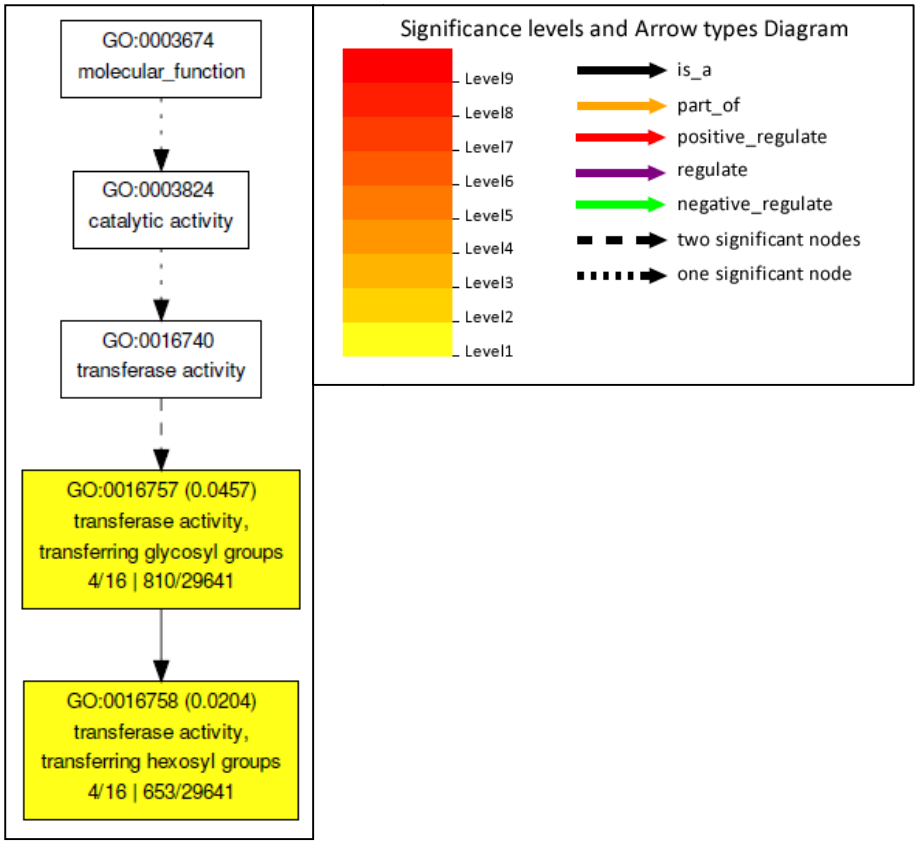


Figure 3.9 Overrepresented GO terms from the dataset of genes with significantly changed expression from 4 to 8 weeks of nodule development in hierarchical tree graphs for molecular function. No GO terms from the biological process and cellular compartment categories were overrepresented.

The only GO terms overrepresented from the gene list with significant changes between the 4 and 8 weeks of nodule development (Figures 3.8 and 3.9) are genes involved in molecular function of transferase activity, GO:0016758 involved in transferring of hexosyl groups and GO:0016757 involved in transferring of glycosyl groups. From the dataset four genes (Glyma17g14750, Glyma11g05680, Glyma02g11640 and Glyma15g34720) were from these GO terms and these genes were significantly down regulated from 4 to 8 weeks of nodule development.

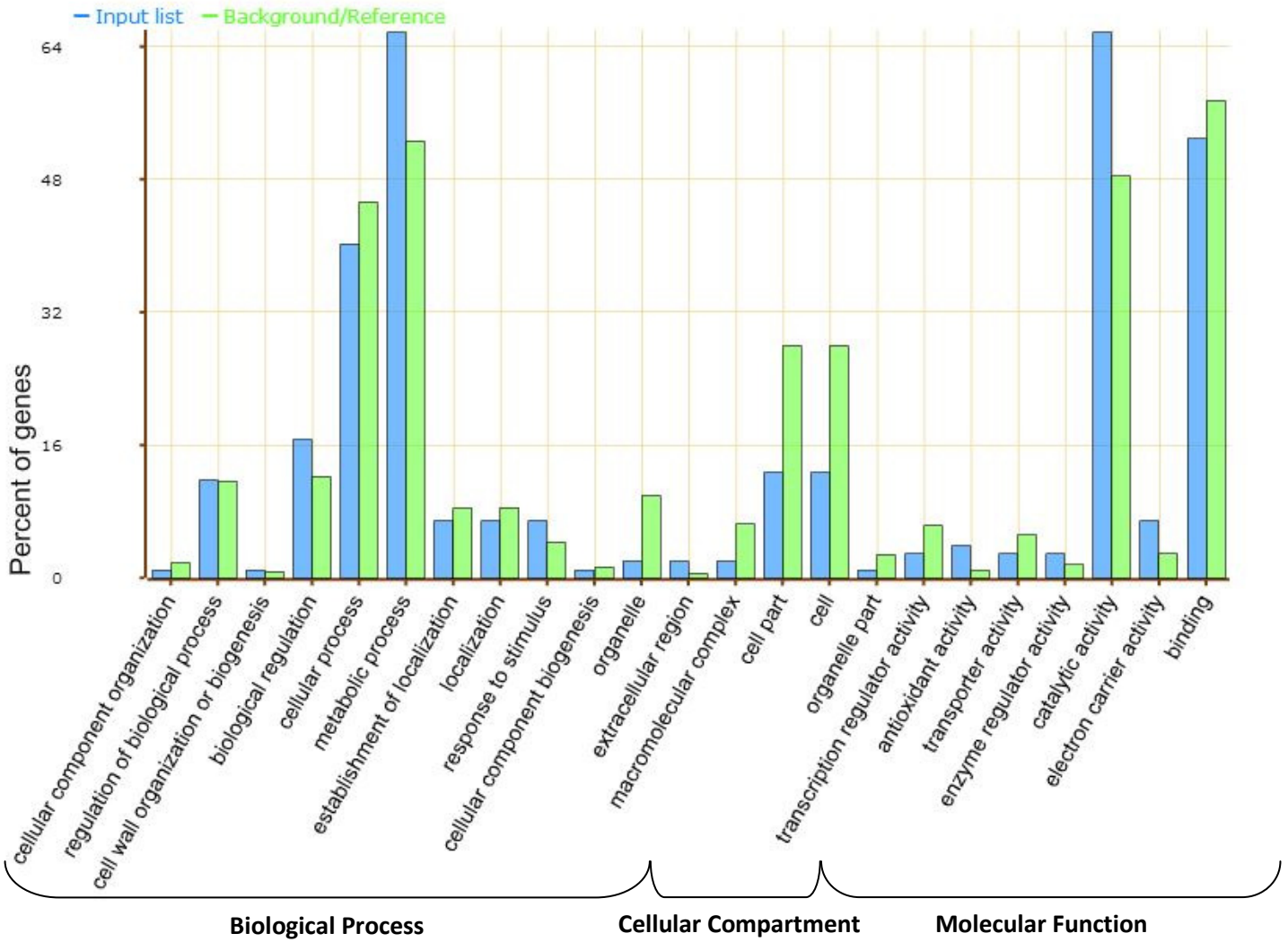


Figure 3.10 Overrepresented GO terms from the dataset of genes with significantly changed expression from 8 to 14 weeks of nodule development from the three categories, biological process, cellular compartment and molecular function.

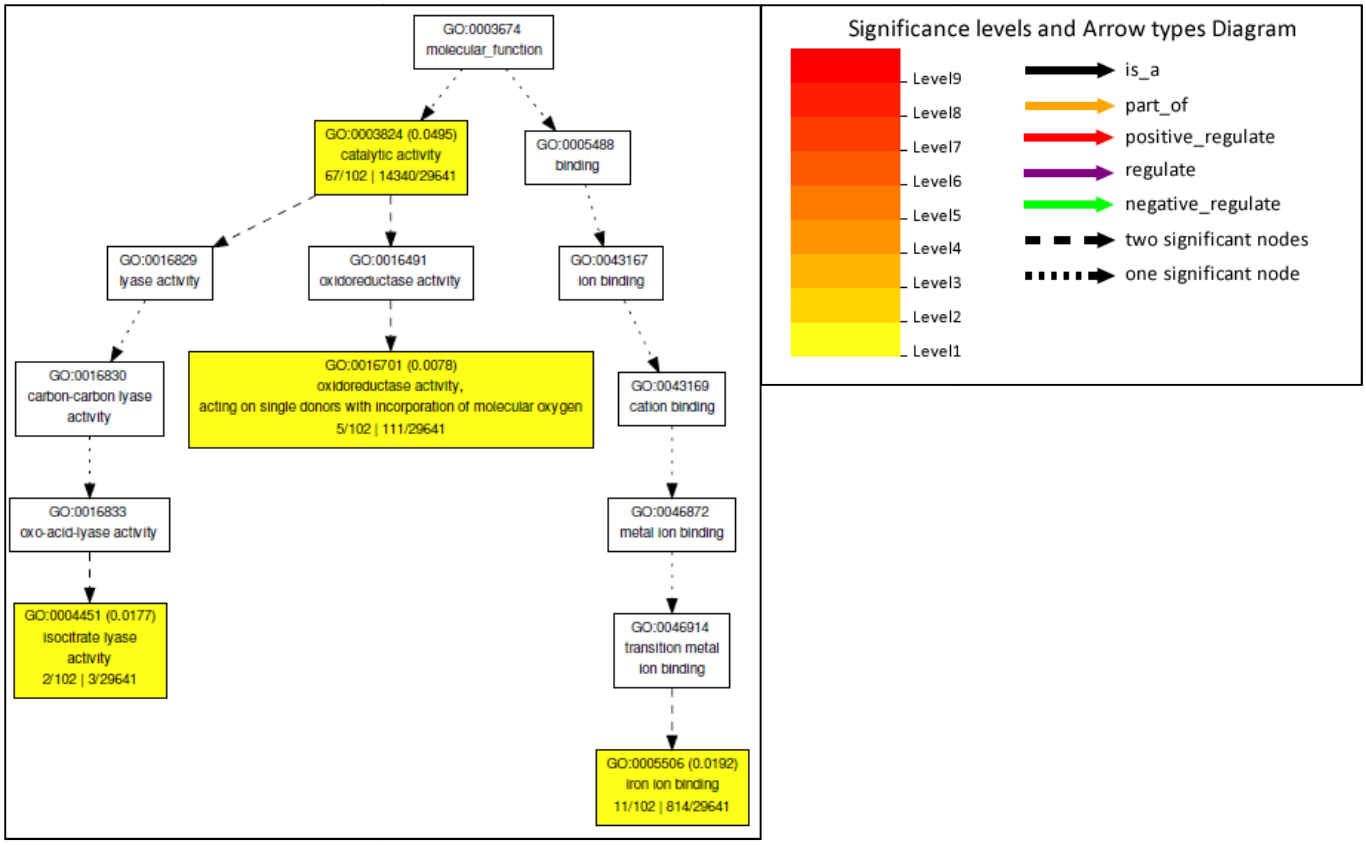


Figure 3.11 Overrepresented GO terms from the dataset of genes with significantly changed expression from 8 to 14 weeks of nodule development in hierarchical tree graphs for molecular function. No GO terms from the biological process and cellular compartment categories were overrepresented.

The GO terms overrepresented from the gene list with significant changes between the 8 and 14 weeks of nodule development (Figure 3.10 and 3.11) are genes involved in the molecular function of catalytic activity (GO:0003824), furthermore dependent subcategories GO:0016701 and GO:0004451 are involved in oxidoreductase activity (acting on single donors with incorporation of molecular oxygen) and isocitrate lyase activity, respectively. The isocitrate lysase genes (Glyma06g45950 and Glyma12g10780) had significant increase in expression from 8 to 14 weeks of nodule development. The only dependent GO term from the binding category was GO:0005506 involved with binding iron ions.

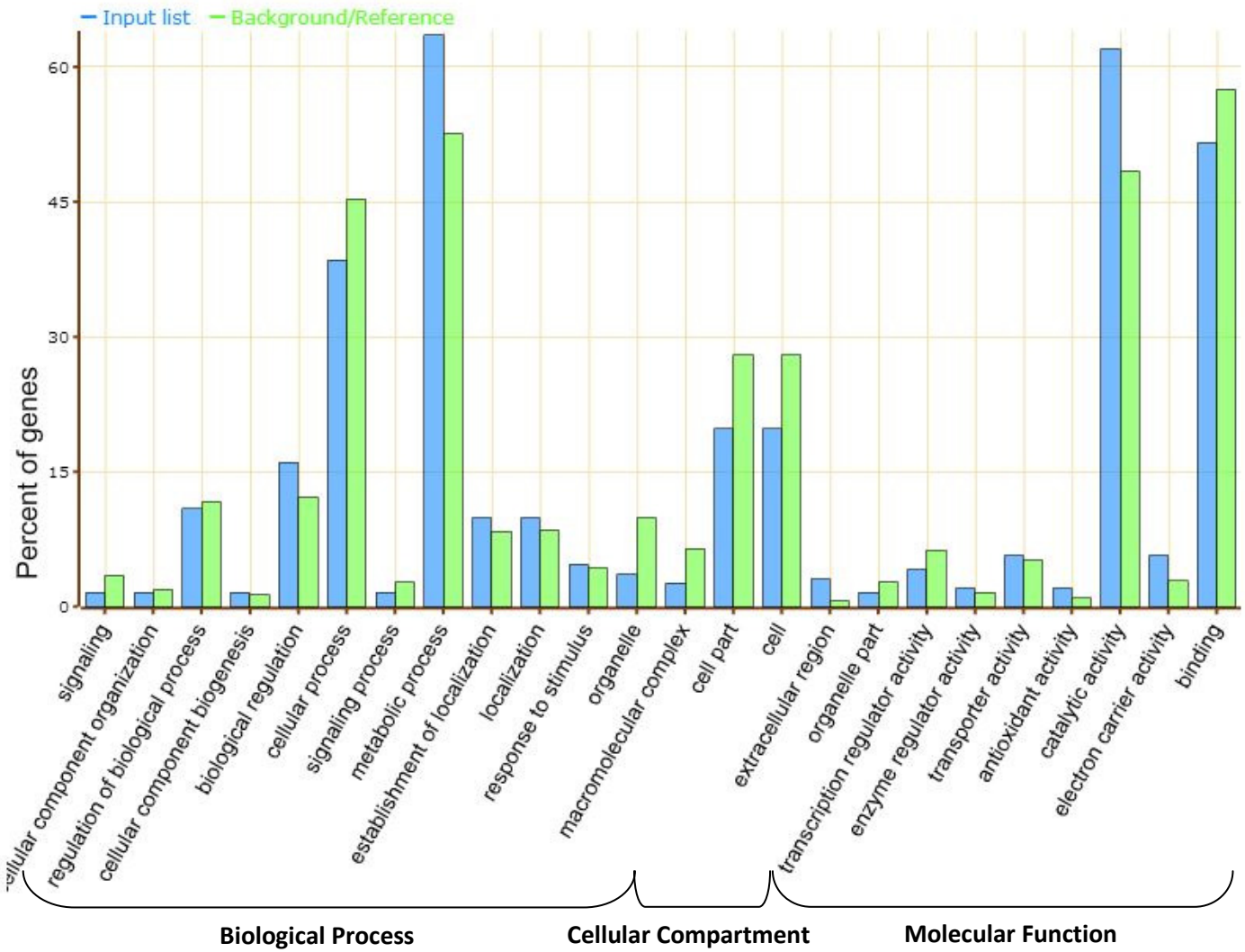
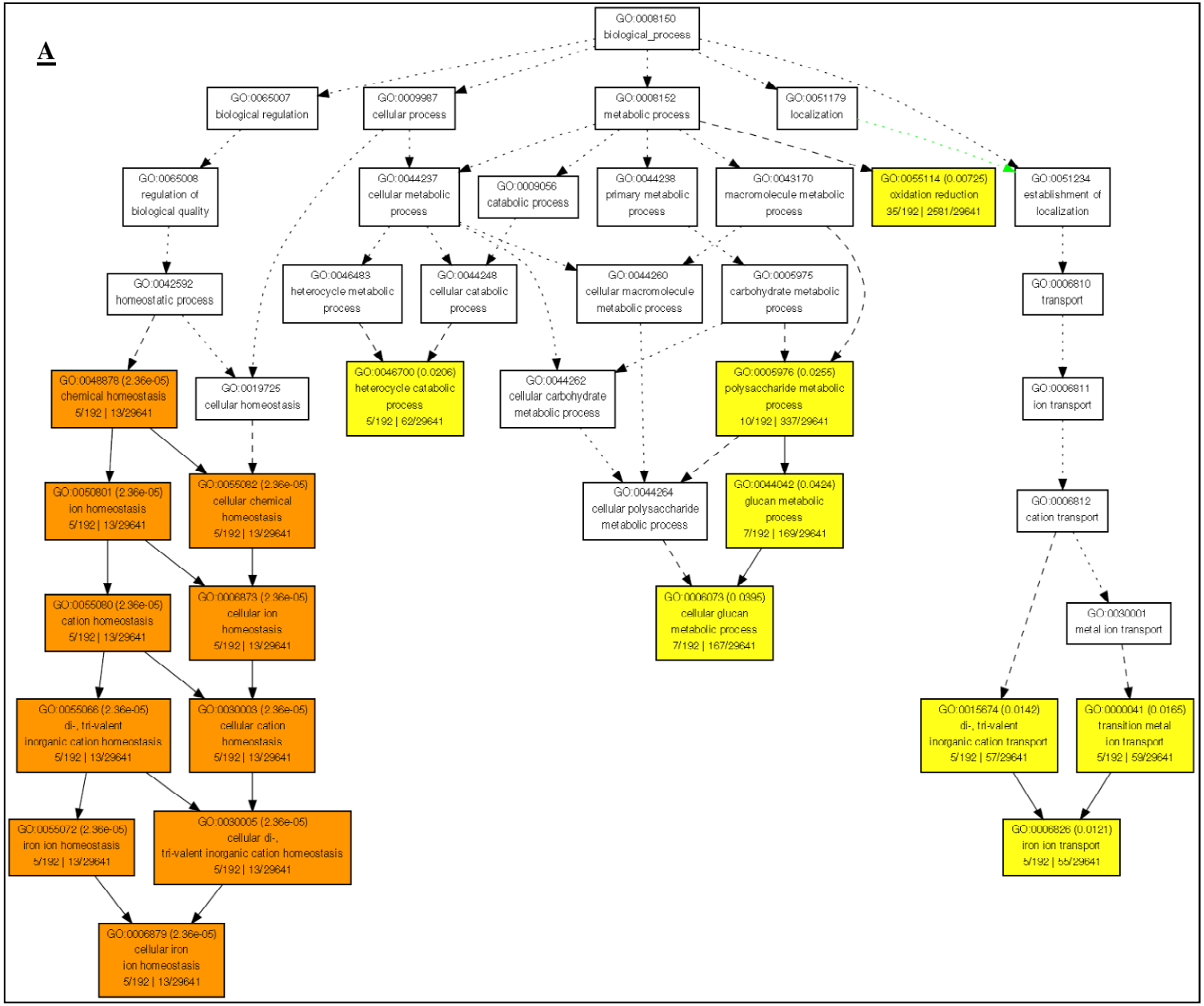


Figure 3.12 Overrepresented GO terms from the dataset of genes with significantly changed expression from 4 to 14 weeks of nodule development from the three categories, biological process, cellular compartment and molecular function.



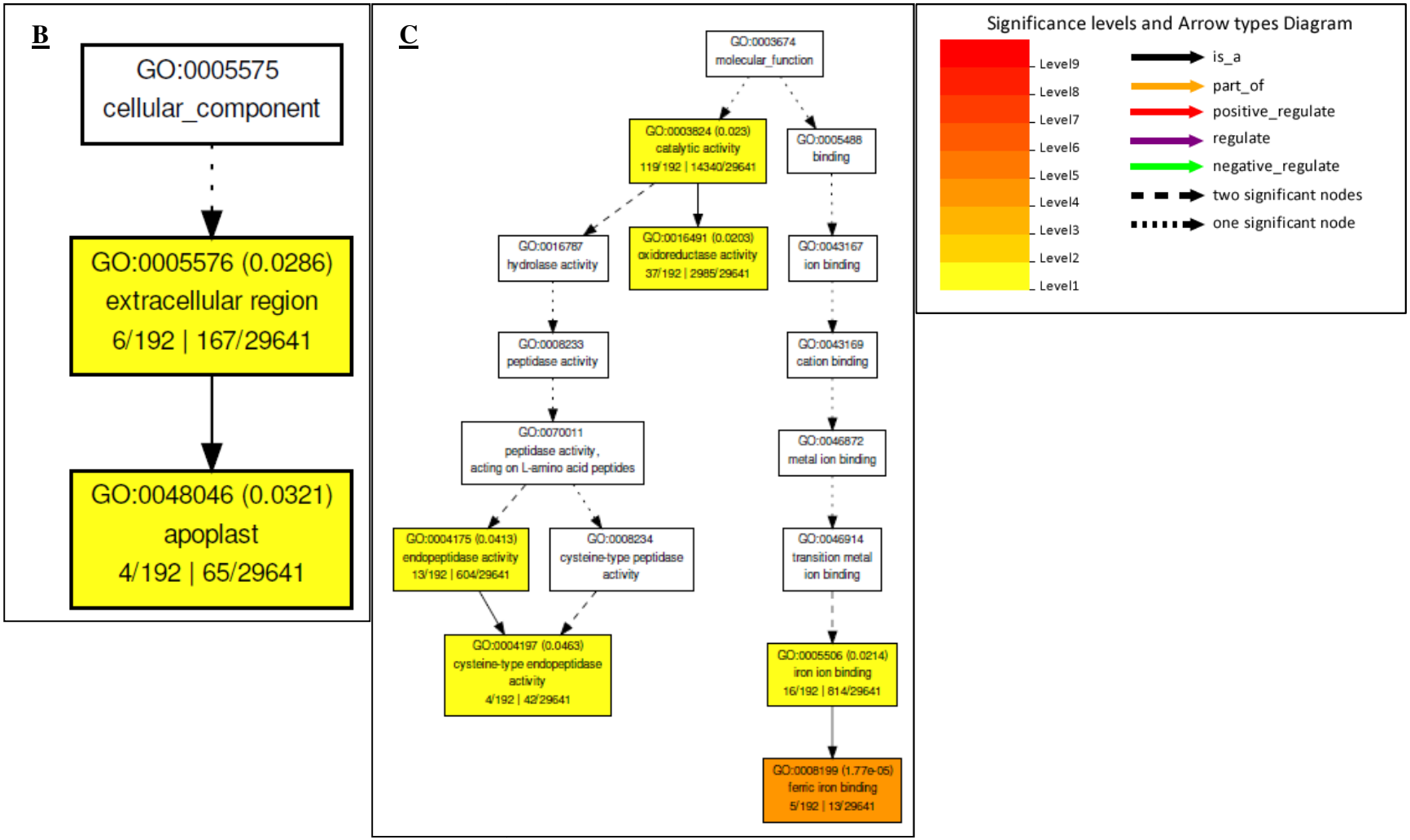


Figure 3.13 Overrepresented GO terms from the dataset of genes with significantly changed expression from 4 to 14 weeks of nodule development in hierarchical tree graphs for each of the three categories, biological process (A), cellular compartment (B) and molecular function (C).

From the Biological Process category several GO terms from the homeostatic processes were overrepresented from the gene list showing significant changes in gene expression (Figure 3.12 and 3.13A). The terms: GO:0055082 (cellular chemical homeostasis), GO:0055072 (iron ion homeostasis), GO:0030005 (cellular di-, tri-valent inorganic cation homeostasis), GO:0050801 (ion homeostasis), GO:0006879 (cellular iron ion homeostasis), GO:0006873 (cellular ion homeostasis), GO:0055080 (cation homeostasis), GO:0030003 (cellular cation homeostasis), GO:0048878 (chemical homeostasis) and GO:0055066 (di-, tri-valent inorganic cation homeostasis). The genes Glyma07g19060, Glyma03g06420, Glyma14g06160, Glyma18g43650 and Glyma02g43040 contributing to these GO terms showed significant increase in expression from 4 weeks to 14 weeks of nodule development. The gene products from these genes are Ferritin proteins which are involved in the storage and transport of iron ions, possibly acting as antioxidants in response to the oxidative stress associated with the aging of the nodule (Lucas et al. 1998, Matamoros et al. 1999). The GO terms GO:0006826 (iron ion transport), GO:0015674 (di-, tri-valent inorganic cation transport), GO:0000041 (transition metal ion transport) are also assigned to these genes, but are a dependent of the transport (GO:0006810) biological process. Further GO terms overrepresented from the Biological Process categories include: GO:0046700 (heterocycle catabolic process), GO:0005976 (polysaccharide metabolic process), GO:0006073 (cellular glucan metabolic process) and GO:0044042 (glucan metabolic process), several of these genes (Glyma13g01140, Glyma13g00280, Glyma09g34140 and Glyma13g01150) are xyloglucan endo-transglycosylase genes involved with the metabolism of xyloglucan of cell walls. They showed significant increase in gene expression from 4 to 14 weeks of nodule development. Genes contributing to the overrepresented GO term oxidation reduction (GO:0055114), showed both increase and decrease in expression, which confounded the contribution of these individual genes. The only GO terms of the cellular compartment

(Figure 3.12 and 3.13B) that were overrepresented are GO:0005576 (extracellular region) and GO:0048046 (apoplast), which are the same genes (Glyma13g01140, Glyma13g00280, Glyma09g34140 and Glyma13g01150) for the xyloglucan endo-transglycosylase ontology. This indicates that these genes function extracellular in the apoplast. Furthermore, the serine-type endopeptidase inhibitor (Glyma18g46560) functions ontologically in the extracellular region and had a significant increase in expression from 4 to 14 weeks of nodule development. The overrepresented GO terms for molecular function (Figure 3.12 and 3.13C) include: GO:0008199 (ferric iron binding) and GO:0005506 (iron ion binding). Under the catalytic activity (GO:0003824) GO term, the dependent GO:0016491 (oxidoreductase activity) term had genes contributing to the overrepresented GO term. They showed both increase and decrease in expression, which confounded the contribution of the individual genes. The GO terms GO:0004175 (endopeptidase activity) and GO:0004197 (cysteine-type endopeptidase activity) are part of the gene family under investigation in the present PhD study. These genes (Glyma14g10620, Glyma17g34900, Glyma17g14680 and Glyma05g04230) had a significant expression increase from 4 to 14 weeks of nodule development.

To further investigate importance and physiological relevance of observed changes in gene expression, ontological classification of these genes were done in conjunction with the MapMan software package. The same stringent threshold, mentioned above, of a $\log_2$-fold-change of ≥ 1.5 and a false discovery rate (FDR) < 0.05, determined by the Benjamini and Hochberg method (Trapnell et al. 2010), was applied to provide statistically sound and physiologically reliable data (Warden et al. 2013). The $\log_2$-fold-change of ≥ 1.5 is roughly equivalent to a 3 fold change in gene expression ($FPKM_{T0}/FPKM_{T1}$).

From Table 3.2, a total of 22 genes showed significant changes in expression between 4 weeks and 8 weeks of nodule development, 7 genes were significantly ($p \leq 0.05$) up-regulated. Glyma11g33922 of unknown function or annotation but the closest BLAST hit with a Metallo-carboxypeptidase inhibitor from *Medicago truncatula*, had the greatest change with almost 7 $\log_2$-fold up-regulation. These types of proteins are produced in response to mechanical damage and activation of the systemin signalling pathway (Ryan 2000), as well as during some developmental processes such as anthesis (Martineau et al. 1991). Strong up-regulation clearly indicates association with mature nodules (8 weeks), with a FPKM of 0 in 4 weeks of nodule development and a significant increase to a FPKM of 593 after 8 weeks.

In total 15 genes were significantly ($p \leq 0.05$) down-regulated, of which Glyma15g03050, with a homology to lipoxygenase 1 (AtLOX1, AT1G55020.1), showed the greatest change with almost 4 $\log_2$-fold down-regulation. This protein has been found to be downstream of the systemin signalling pathway, and acting in plastids during wounding in leaves, but also has activity during growth and development (Peng, et al. 1994, Schilmiller and Howe 2005). Although Glyma08g12340, a granulin-repeat containing cysteine protease family protein, had a comparable 4 $\log_2$-fold down-regulation, this gene was no longer considered after the stringent FDR < 0.05 cut-off was applied. This granulin-containing cysteine protease is closely related to the RD21 sub-family of cysteine proteases involved in cell death preparation during leaf senescence and responds to wounding (van der Hoorn, et al. 2004). The bin map in Figure 3.14 shows the distribution of these genes with differential gene regulation grouped based on their physiological functions.

Table 3.2 Differentially regulated genes comparing 4 and 8 weeks of nodule development after significance and FDR cut-offs have been applied. Proteases or protease inhibitors are indicated by bold letters in the Gene IDs column.

Gene ID	FPKM		Log_2-fold	Significant	TAIR	TAIR Define	Closest Blast Hit Annotation
	4w	8w					
Glyma11g33922	0.00	593.29	**6.90**	Yes			Metallo-carboxypeptidase Inhibitor [*MedicagoTruncatula*]
Glyma13g01140	4.81	142.01	**4.88**	Yes	AT5G57560.1	Xyloglucan Endo-transglucosylase/Hydrolase Family Protein	
Glyma17g01170	3.98	37.76	**3.25**	Yes	AT4G17280.1	Auxin-Responsive Family Protein	
Glyma13g11965	42.03	253.45	**2.59**	Yes			RRNA Intron-Encoded Homing Endonuclease [*MedicagoTruncatula*]
Glyma09g33650	5.29	31.48	**2.57**	Yes	AT4G37870.1	Phosphoenolpyruvate Carboxykinase 1	
Glyma13g25273	9.89	51.20	**2.37**	Yes	AT1G61120.1	Terpene Synthase 04	
Glyma17g23870	9.40	47.41	**2.33**	Yes			UNKNOWN
Glyma17g14750	31.90	6.84	**-2.22**	Yes	AT1G62660.1	Glycosyl Hydrolases Family 32 Protein	
Glyma11g05680	38.62	8.20	**-2.24**	Yes	AT4G34138.1	Udp-GlucosylTransferase 73b1	
Glyma02g11640	41.73	8.47	**-2.30**	Yes	AT4G34131.1	Udp-GlucosylTransferase 73b3	
Glyma05g28260	26.54	5.36	**-2.31**	Yes	AT5G48570.1	Fkbp-Type Peptidyl-ProlylCis-Trans Isomerase Family Protein	
Glyma08g46520	95.75	17.92	**-2.42**	Yes	AT5G06900.1	Cytochrome P450, Family 93, Subfamily D, Polypeptide 1	
Glyma17g13420	60.63	10.12	**-2.58**	Yes	AT3G26330.1	Cytochrome P450, Family 71, Subfamily B, Polypeptide 37	
Glyma15g34720	54.95	8.80	**-2.64**	Yes	AT4G34131.1	Udp-GlucosylTransferase 73b3	
Glyma13g24490	223.21	29.08	**-2.94**	Yes	AT1G07400.1	Hsp20-Like Chaperones Superfamily Protein	
Glyma07g32050	171.56	20.84	**-3.04**	Yes	AT1G53540.1	Hsp20-Like Chaperones Superfamily Protein	
Glyma15g23830	227.84	26.20	**-3.12**	Yes			PF02095 - Extensin-Like Protein Repeat
Glyma06g17860	28.70	3.27	**-3.13**	Yes	AT5G50400.1	Purple Acid Phosphatase 27	
Glyma01g16140	61.57	6.89	**-3.16**	Yes	AT1G65680.1	Expansin B2	
Glyma18g12660	38.32	3.80	**-3.33**	Yes	AT1G78570.1	Rhamnose Biosynthesis 1	
Glyma15g03050	55.16	3.97	**-3.80**	Yes	AT1G55020.1	Lipoxygenase 1	
Glyma08g12340	38.46	1.91	**-4.33**	No	AT5G43060.1	Granulin Repeat Cysteine Protease Family Protein	

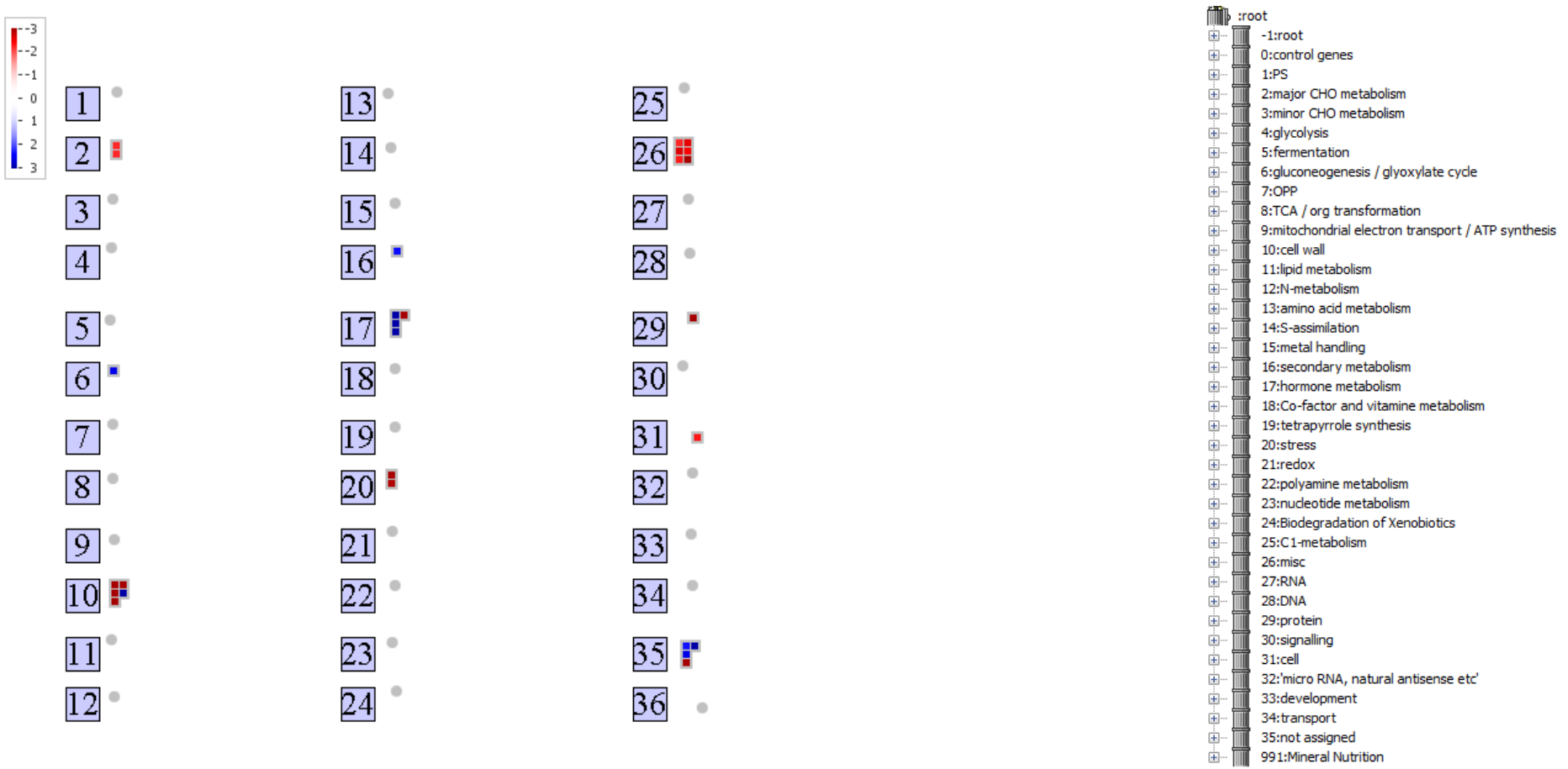


Figure 3.14 Bin maps generated by MapMan groups of differentially regulated genes between 4 and 8 weeks of nodule development. All genes shown had significant ($p \leq 0.05$) change in expression between the two respective time points. The degree of change is depicted based on the colour scale with red indicating down-regulation and blue indicating up-regulation.

A total of 170 genes were differentially regulated and had significant ($p \leq 0.05$) changes in expression between 8 and 14 weeks of nodule development (Table 3.3). From these, 110 genes were significantly ($p \leq 0.05$) up-regulated with the greatest change of nearly 5.5 log_2-fold up-regulation for Glyma09g29410, a oligo-peptide transmembrane transporter protein, which is suspected to be involved with the transport of regulatory molecules such as hormone-peptide conjugates (Koh et al. 2002). The Kunitz-family trypsin protease inhibitor protein, Glyma08g45520 had approximately a 4 log_2-fold change in expression from a FPKM of 15 to 266. These proteins have been implicated as seed storage proteins, in pest resistance, as well as in senescing nodule cells, regulating a plant or bacteroid-derived protease (Manen et al. 1991, Major and Constabel 2008). The alpha-vacuolar-processing enzyme, Glyma14g10620, increased in expression from FPKM 9 to 110, which had approximately 3.5 log_2-fold increase in expression. This group of proteins are thought to be involved in PCD with caspases-like functions and during nodule senescence activating other inactive proteases by post-translational maturation (Van de Velde et al. 2006).

The greatest down-regulation change of 5 log_2-fold was found for Glyma08g21420, a member of the HAD superfamily, subfamily IIIB, an acid phosphatase protein and also for Glyma20g33450, a cycloartenol synthase 1 protein homologue, involved in plant cell viability and plastid biosynthesis (Babiychuk et al. 2008). The subtilase-family protein, Glyma17g14270, had a 4.18 log_2-fold down-regulation in expression. Although these proteases are thought to have similar activity and function to caspases in other plant tissues (Vartapetian et al. 2011), their function in nodules certainly differ as their expression is reduced from a FPKM of 38 to 2. The aspartyl protease family protein, Glyma04g09740, had a 2.5 log_2-fold down-regulation in expression. The bin maps in Figure 3.15 show the distribution of these genes with differential gene regulation grouped into broad physiological functions.

Table 3.3 Differentially regulated genes after 8 and 14 weeks of nodule development after significance and FDR cut-offs have been applied. Proteases or protease inhibitors are indicated by bold letters in the Gene IDs column.

Gene ID	FPKM		Log_2-fold	Significant	TAIR	TAIR Define	Closest BLAST Hit Annotation
	8w	14w					
Glyma09g29410	1.15	50.12	**5.44**	yes	AT1G65730.1	Yellow Stripe Like 7	
Glyma06g45950	5.15	135.71	**4.72**	yes	AT3G21720.1	IsocitrateLyase	
Glyma06g38410	2.61	47.91	**4.20**	yes	AT4G27410.2	Nac (No Apical Meristem) Domain Transcriptional Regulator Superfamily Protein	
Glyma08g45520	14.99	266.59	**4.15**	yes	AT1G17860.1	Kunitz Family Trypsin And Protease Inhibitor Protein	
Glyma12g10780	1.26	22.31	**4.14**	yes	AT3G21720.1	IsocitrateLyase	
Glyma13g34520	4.68	75.51	**4.01**	yes	AT1G78860.1	D-Mannose Binding Lectin Protein With Apple-Like Carbohydrate-Binding Domain	
Glyma18g42920	3.30	51.03	**3.95**	yes	AT5G01520.1	Ring/U-Box Superfamily Protein	
Glyma19g02580	5.83	82.63	**3.82**	yes	AT5G39820.1	Nac Domain Containing Protein 94	
Glyma17g13730	5.64	79.16	**3.81**	yes	AT5G03860.1	Malate Synthase	
Glyma15g07710	1.49	20.90	**3.81**	yes			Pf12143 - Protein Of Unknown Function (Duf_B2219), Pf00264 - Common Central Domain Of Tyrosinase, Pf12142 - Polyphenol Oxidase Middle Domain
Glyma06g46450	3.30	46.03	**3.80**	yes	AT2G32540.1	Cellulose Synthase-Like B4	
Glyma07g00873	2.53	33.79	**3.74**	yes	AT1G55020.1	Lipoxygenase 1	
Glyma07g35630	2.96	38.97	**3.72**	yes	AT1G69490.1	Nac-Like, Activated By Ap3/Pi	
Glyma13g35550	4.71	61.63	**3.71**	yes	AT3G15500.1	Nac Domain Containing Protein 3	
Glyma06g43970	30.23	395.41	**3.71**	yes	AT4G35160.1	O-Methyltransferase Family Protein	
Glyma11g15180	7.28	93.51	**3.68**	yes	AT3G46130.1	MYB Domain Protein 48	
Glyma16g31290	18.86	235.84	**3.64**	yes	AT3G61260.1	Remorin Family Protein	
Glyma19g28476	6.28	78.27	**3.64**	yes	AT3G04070.2	Nac Domain Containing Protein 47	
Glyma14g10620	9.16	109.64	**3.58**	yes	AT2G25940.1	Alpha-Vacuolar Processing Enzyme	
Glyma07g04810	9.60	114.78	**3.58**	yes	AT1G54070.1	Dormancy/Auxin Associated Family Protein	
Glyma13g01420	2.48	29.27	**3.56**	yes	AT2G18700.1	Trehalose Phosphatase/Synthase 11	
Glyma16g01390	5.89	68.24	**3.53**	yes	AT1G54070.1	Dormancy/Auxin Associated Family Protein	

Glyma03g39050	3.74	43.33	**3.53**	yes	AT1G09240.1	Nicotianamine Synthase 3	
Glyma18g43580	3.39	38.61	**3.51**	yes	AT1G66350.1	RGA-Like 1	
Glyma10g29190	11.65	129.14	**3.47**	yes	AT1G56220.1	Dormancy/Auxin Associated Family Protein	
Glyma08g20090	6.31	69.93	**3.47**	yes	AT1G60940.1	Snf1-Related Protein Kinase 2.10	
Glyma11g20020	1.49	15.96	**3.42**	yes	AT4G05160.1	Amp-Dependent Synthetase And Ligase Family Protein	
Glyma16g27760	1.15	11.66	**3.34**	yes	AT1G71710.2	Dnase I-Like Superfamily Protein	
Glyma19g28970	6.13	61.95	**3.34**	yes	AT5G18130.1		UNKNOWN
Glyma06g03420	7.88	79.37	**3.33**	yes	AT1G19530.1		UNKNOWN
Glyma14g06400	2.67	26.68	**3.32**	yes	AT5G05600.1	2-Oxoglutarate (2og) And Fe(Ii)-Dependent Oxygenase Superfamily Protein	
Glyma17g02600	3.36	33.17	**3.30**	yes	AT3G06390.1	Uncharacterised Protein Family (Upf0497)	
Glyma12g23150	4.55	44.71	**3.30**	yes	AT4G27450.1	Aluminium Induced Protein With Ygl And Lrdr Motifs	
Glyma14g38980	4.81	47.21	**3.29**	yes	AT5G56550.1	Oxidative Stress 3	
Glyma13g42340	1.51	14.55	**3.27**	yes	AT1G55020.1	Lipoxygenase 1	
Glyma06g05280	18.29	171.23	**3.23**	yes	AT1G10070.1	Branched-Chain Amino Acid Transaminase 2	
Glyma16g22650	24.67	226.76	**3.20**	yes	AT3G51430.1	Calcium-Dependent Phosphotriesterase Superfamily Protein	
Glyma19g25070	6.64	60.44	**3.19**	yes	AT5G20090.1	Uncharacterised Protein Family (Upf0041)	
Glyma06g17020	1.42	12.59	**3.15**	yes	AT5G51070.1	ClpAtpase	
Glyma03g37210	3.55	30.81	**3.12**	yes	AT1G02860.1	Spx (Syg1/Pho81/Xpr1) Domain-Containing Protein	
Glyma18g15570	2.44	20.73	**3.09**	yes	AT2G02870.1	Galactose Oxidase/Kelch Repeat Superfamily Protein	
Glyma01g06150	4.49	37.46	**3.06**	yes	AT1G69490.1	Nac-Like, Activated By Ap3/Pi	
Glyma01g42025	6.17	49.58	**3.01**	yes	AT1G62770.1	Plant Invertase/Pectin Methylesterase Inhibitor Superfamily Protein	
Glyma06g17910	5.52	44.25	**3.00**	yes	AT3G48530.1	Snf1-Related Protein Kinase Regulatory Subunit Gamma 1	
Glyma09g32370	2.32	18.45	**2.99**	yes	AT4G13250.1	Nad(P)-Binding Rossmann-Fold Superfamily Protein	
Glyma08g47520	9.05	70.47	**2.96**	yes	AT5G13180.1	Nac Domain Containing Protein 83	
Glyma05g21680	3.38	26.12	**2.95**	yes	AT4G37390.1	Auxin-Responsive Gh3 Family Protein	
Glyma06g35585	20.90	157.23	**2.91**	yes	AT5G53970.1	Tyrosine Transaminase Family Protein	
Glyma19g01470	6.19	46.32	**2.90**	yes	AT3G01420.1	Peroxidase Superfamily Protein	
Glyma07g33050	12.22	90.11	**2.88**	yes	AT2G15890.1	Maternal Effect Embryo Arrest 14	
Glyma08g39870	2.12	15.30	**2.85**	yes	AT1G60140.1	Trehalose Phosphate Synthase	
Glyma13g23170	26.02	186.88	**2.84**	yes	AT1G15690.1	Inorganic H Pyrophosphatase Family Protein	

Glyma05g27970	3.20	22.77	**2.83**	yes	AT1G13710.1	Cytochrome P450, Family 78, Subfamily A, Polypeptide 5	
Glyma13g00280	6.69	47.17	**2.82**	yes	AT4G25810.1	Xyloglucan Endo-transglycosylase 6	
Glyma02g12220	16.24	114.24	**2.81**	yes	AT1G69490.1	Nac-Like, Activated By Ap3/Pi	
Glyma10g42270	4.75	33.31	**2.81**	yes			UNKNOWN
Glyma02g40640	5.58	38.91	**2.80**	yes	AT5G16370.1	Acyl Activating Enzyme 5	
Glyma14g03410	6.31	43.29	**2.78**	yes	AT1G06570.1	Phytoene Desaturation 1	
Glyma08g16290	206.51	1403.67	**2.76**	yes			UNKNOWN
Glyma11g11640	5.27	35.68	**2.76**	yes	AT1G32700.1	Platz Transcription Factor Family Protein	
Glyma17g06090	9.87	66.45	**2.75**	yes	AT5G19890.1	Peroxidase Superfamily Protein	
Glyma20g28720	4.31	28.18	**2.71**	yes	AT3G22740.1	Homocysteine S-Methyltransferase 3	
Glyma15g07720	4.72	30.70	**2.70**	yes	AT4G27450.1	Aluminium Induced Protein With Ygl And Lrdr Motifs	
Glyma08g04110	5.08	32.99	**2.70**	yes	AT4G29100.1	Basic Helix-Loop-Helix (Bhlh) Dna-Binding Superfamily Protein	
Glyma04g01130	95.06	614.80	**2.69**	yes	AT1G20440.1	Cold-Regulated 47	
Glyma16g34320	5.47	35.20	**2.69**	yes	AT1G53345.1		UNKNOWN
Glyma17g04920	6.45	41.32	**2.68**	yes	AT4G33150.1	Lysine-KetoglutarateReductase/Saccharopine Dehydrogenase Bifunctional Enzyme	
Glyma10g41120	0.00	26.50	**2.66**	yes	AT2G01300.1		UNKNOWN
Glyma08g20220	12.39	77.80	**2.65**	yes	AT1G55020.1	Lipoxygenase 1	
Glyma15g01610	18.47	113.16	**2.62**	yes	AT1G80440.1	Galactose Oxidase/Kelch Repeat Superfamily Protein	
Glyma11g02980	10.78	64.38	**2.58**	yes	AT4G22920.1	Non-Yellowing 1	
Glyma04g01390	11.82	70.40	**2.57**	yes	AT2G21620.1	Adenine Nucleotide Alpha Hydrolases-Like Superfamily Protein	
Glyma17g05920	22.89	135.21	**2.56**	yes	AT1G17590.1	Nuclear Factor Y, Subunit A8	
Glyma01g11390	17.51	103.26	**2.56**	yes	AT4G17900.1	Platz Transcription Factor Family Protein	
Glyma07g02420	3.80	22.27	**2.55**	yes	AT2G24100.1		UNKNOWN
Glyma03g19260	25.17	147.46	**2.55**	yes	AT3G09390.1	Metallothionein 2a	
Glyma09g07100	5.46	31.90	**2.55**	yes	AT4G26140.2	Beta-Galactosidase 12	
Glyma17g11500	7.16	41.78	**2.54**	yes	AT5G11090.1	Serine-Rich Protein-Related	
Glyma12g35070	36.62	211.80	**2.53**	yes	AT4G27450.1	Aluminium Induced Protein With Ygl And Lrdr Motifs	
Glyma04g37140	11.32	65.07	**2.52**	yes	AT3G48530.1	Snf1-Related Protein Kinase Regulatory Subunit Gamma 1	
Glyma05g34810	4.31	24.68	**2.52**	yes	AT1G23200.1	Plant Invertase/Pectin Methylesterase Inhibitor Superfamily	
Glyma07g25390	16.31	92.87	**2.51**	yes	AT1G03400.1	2-Oxoglutarate (2og) And Fe(Ii)-Dependent Oxygenase Superfamily Protein	

Glyma13g39240	10.10	57.47	**2.51**	yes	AT2G27480.1	Calcium-Binding Ef-Hand Family Protein	
Glyma07g19060	9.77	55.00	**2.49**	yes	AT5G01600.1	Ferretin 1	
Glyma07g01660	14.65	82.02	**2.48**	yes	AT1G14520.1	Myo-Inositol Oxygenase 1	
Glyma01g38870	6.71	37.29	**2.47**	yes	AT4G31940.1	Cytochrome P450, Family 82, Subfamily C, Polypeptide 4	
Glyma04g05280	5.49	30.37	**2.47**	yes	AT2G25930.1	Hydroxyproline-Rich Glycoprotein Family Protein	
Glyma07g30880	4.31	23.45	**2.44**	yes	AT1G11260.1	Sugar Transporter 1	
Glyma16g04440	7.88	42.44	**2.43**	yes	AT5G18130.1		UNKNOWN
Glyma06g12701	11.40	60.34	**2.40**	yes	AT1G21450.1	Scarecrow-Like 1	
Glyma17g01510	9.20	48.43	**2.40**	yes	AT1G62280.1	Slac1 Homologue 1	
Glyma15g11680	6.37	33.54	**2.40**	yes	AT1G62300.1	WRKY Family Transcription Factor	
Glyma15g40790	18.37	95.47	**2.38**	yes	AT2G32190.1		UNKNOWN
Glyma04g04270	7.97	40.61	**2.35**	yes	AT5G07850.1	Hxxxd-Type Acyl-Transferase Family Protein	
Glyma03g07972	23.44	118.13	**2.33**	yes			UNKNOWN
Glyma08g41020	8.60	42.75	**2.31**	yes	AT2G02870.1	Galactose Oxidase/Kelch Repeat Superfamily Protein	
Glyma13g44400	5.69	27.79	**2.29**	yes	AT1G15740.1	Leucine-Rich Repeat Family Protein	
Glyma08g23590	5.49	26.53	**2.27**	yes	AT4G30780.1		UNKNOWN
Glyma16g27900	14.44	69.24	**2.26**	yes	AT4G33420.1	Peroxidase Superfamily Protein	
Glyma11g35800	18.04	84.79	**2.23**	yes	AT3G10985.1	Senescence Associated Gene 20	
Glyma13g36110	6.98	32.75	**2.23**	yes	AT4G31940.1	Cytochrome P450, Family 82, Subfamily C, Polypeptide 4	
Glyma09g00820	8.18	38.15	**2.22**	yes	AT1G62300.1	WRKYFamily Transcription Factor	
Glyma17g37020	8.95	41.69	**2.22**	yes	AT1G19530.1		UNKNOWN
Glyma09g22306	0.00	17.62	**2.18**	yes	AT1G70680.1	Caleosin-Related Family Protein	
Glyma08g16280	15.85	71.63	**2.18**	yes	AT2G32940.1	Argonaute Family Protein	
Glyma07g02180	18.87	82.55	**2.13**	yes	AT3G16170.1	Amp-Dependent Synthetase And Ligase Family Protein	
Glyma11g12650	39.54	170.10	**2.11**	yes			UNKNOWN
Glyma13g23161	0.00	11.57	**1.73**	yes	AT4G39140.1	Ring/U-Box Superfamily Protein	
Glyma08g15976	0.00	10.56	**1.64**	yes	AT2G44480.1	Beta Glucosidase 17	
Glyma17g17120	0.00	9.89	**1.57**	yes	AT5G65870.1	Phytosulfokine 5 Precursor	
Glyma13g24720	9.81	0.00	**-1.57**	yes	AT2G34700.1	Pollen Ole E 1 Allergen And Extensin Family Protein	

Glyma05g04380	10.39	0.00	**-1.62**	yes	AT1G12100.1	Bifunctional Inhibitor/Lipid-Transfer Protein/Seed Storage 2s Albumin Superfamily Protein	
Glyma17g06220	10.63	0.00	**-1.64**	yes	AT5G56970.1	Cytokinin Oxidase 3	
Glyma10g06390	11.82	0.00	**-1.75**	yes	AT3G09870.1	Saur-Like Auxin-Responsive Protein Family	
Glyma08g45281	14.87	0.00	**-1.99**	yes			UNKNOWN
Glyma18g50980	52.63	12.08	**-2.12**	yes	AT2G36780.1	Udp-Glycosyltransferase Superfamily Protein	
Glyma18g14490	17.95	0.00	**-2.20**	yes			UNKNOWN
Glyma13g16280	14.40	3.03	**-2.25**	yes	AT5G56890.1	Protein Kinase Superfamily Protein	
Glyma04g40000	148.43	31.00	**-2.26**	yes	AT2G40610.1	Expansin A8	
Glyma06g42080	57.08	11.35	**-2.33**	yes	AT5G54770.1	Thiazole Biosynthetic Enzyme, Chloroplast (Ara6) (Thi1) (Thi4)	
Glyma15g29880	18.25	3.57	**-2.36**	yes	AT5G14210.1	Leucine-Rich Repeat Protein Kinase Family Protein	
Glyma20g04150	22.97	4.40	**-2.38**	yes	AT1G14040.1	Exs (Erd1/Xpr1/Syg1) Family Protein	
Glyma08g00780	21.88	4.11	**-2.41**	yes	AT1G72160.1	Sec14p-Like Phosphatidylinositol Transfer Family Protein	
Glyma18g50970	32.84	6.00	**-2.45**	yes	AT5G15780.1	Pollen Ole E 1 Allergen And Extensin Family Protein	
Glyma14g01610	21.83	3.98	**-2.46**	yes	AT5G55930.1	Oligopeptide Transporter 1	
Glyma09g36220	61.81	11.22	**-2.46**	yes	AT3G08030.1	Protein Of Unknown Function, Duf642	
Glyma10g37380	17.97	3.25	**-2.47**	yes	AT5G52882.1	P-Loop Containing Nucleoside Triphosphate Hydrolases Superfamily Protein	
Glyma16g04750	31.05	5.57	**-2.48**	yes	AT5G18280.2	Apyrase 2	
Glyma15g10520	129.84	23.26	**-2.48**	yes	AT4G25050.1	Acyl Carrier Protein 4	
Glyma10g44360	213.13	38.13	**-2.48**	yes			UNKNOWN
Glyma02g43330	22.95	0.00	**-2.48**	yes		Nodulin-16 Precursor	
Glyma10g34700	14.70	2.62	**-2.49**	yes	AT1G66950.1	Pleiotropic Drug Resistance 11	
Glyma04g09740	74.39	12.95	**-2.52**	yes	AT1G09750.1	Eukaryotic Aspartyl Protease Family Protein	
Glyma02g43320	1213.36	206.72	**-2.55**	yes			Nodulin-16 Precursor
Glyma08g24610	13.02	2.21	**-2.56**	yes	AT5G14210.1	Leucine-Rich Repeat Protein Kinase Family Protein	
Glyma08g17190	63.38	10.70	**-2.57**	yes	AT2G18370.1	Bifunctional Inhibitor/Lipid-Transfer Protein/Seed Storage 2s Albumin Superfamily Protein	
Glyma13g38300	71.68	11.59	**-2.63**	yes	AT4G11290.1	Peroxidase Superfamily Protein	
Glyma18g02450	37.31	5.91	**-2.66**	yes	AT4G39640.1	Gamma-GlutamylTranspeptidase 1	
Glyma18g47820	44.67	6.53	**-2.77**	yes	AT4G12910.1	Serine Carboxypeptidase-Like 20	
Glyma15g15010	20.13	2.86	**-2.81**	yes	AT5G49460.1	Atp Citrate Lyase Subunit B 2	
Glyma07g14460	63.58	9.02	**-2.82**	yes	AT1G11680.1	Cytochrome P450 51g1	

Glyma15g10610	30.26	0.00	**-2.82**	yes	AT4G15560.1	Deoxyxylulose-5-Phosphate Synthase	
Glyma20g33710	127.97	16.95	**-2.92**	yes	AT5G10030.1	Tgacg Motif-Binding Factor 4	
Glyma10g40780	11.35	1.44	**-2.98**	yes	AT4G34220.1	Leucine-Rich Repeat Protein Kinase Family Protein	
Glyma20g27280	26.25	3.31	**-2.99**	yes	AT1G04820.1	Tubulin Alpha-4 Chain	
Glyma02g36580	202.82	25.09	**-3.02**	yes	AT5G57920.1	Early Nodulin-Like Protein 10	
Glyma11g21480	72.17	8.75	**-3.04**	yes	AT5G12380.1	Annexin 8	
Glyma08g14670	18.25	2.17	**-3.07**	yes	AT4G39800.1	Myo-Inositol-1-Phosphate Synthase 1	
Glyma05g22060	13.17	1.50	**-3.14**	yes	AT5G67360.1	Subtilase Family Protein	
Glyma02g43341	291.46	30.01	**-3.28**	yes			Nodulin-16 Precursor
Glyma13g11090	307.51	31.00	**-3.31**	yes	AT3G22142.1	Bifunctional Inhibitor/Lipid-Transfer Protein/Seed Storage 2s Albumin Superfamily Protein	
Glyma01g04545	57.37	0.00	**-3.64**	yes			UNKNOWN
Glyma18g52670	99.12	7.54	**-3.72**	yes	AT2G39518.1	Uncharacterised Protein Family (Upf0497)	
Glyma13g30950	26.55	1.94	**-3.77**	yes	AT2G33850.1		UNKNOWN
Glyma06g14850	101.89	7.08	**-3.85**	yes	AT2G40610.1	Expansin A8	
Glyma20g03988	28.72	1.97	**-3.86**	yes	AT1G25540.1	Phytochrome And Flowering Time Regulatory Protein (Pft1)	
Glyma07g17030	144.42	9.76	**-3.89**	yes	AT5G55410.2	Bifunctional Inhibitor/Lipid-Transfer Protein/Seed Storage 2s Albumin Superfamily Protein	
Glyma02g11820	84.97	4.98	**-4.09**	yes	AT2G46210.1	Fatty Acid/SphingolipidDesaturase	
Glyma03g08478	81.03	0.00	**-4.10**	yes			UNKNOWN
Glyma17g14270	38.35	2.11	**-4.18**	yes	AT1G04110.1	Subtilase Family Protein	
Glyma12g08990	17.76	0.96	**-4.22**	yes	AT5G22740.1	Cellulose Synthase-Like A02	
Glyma18g02230	136.96	6.95	**-4.30**	yes	AT5G10180.1	Slufate Transporter 2;1	
Glyma20g06290	320.49	14.61	**-4.46**	yes	AT3G22142.1	Bifunctional Inhibitor/Lipid-Transfer Protein/Seed Storage 2s Albumin Superfamily Protein	
Glyma06g45700	88.94	4.00	**-4.48**	yes	AT4G15210.1	Beta-Amylase 5	
Glyma17g06740	13.66	0.58	**-4.55**	yes	AT4G30020.1	Pa-Domain Containing Subtilase Family Protein	
Glyma17g08110	281.86	11.85	**-4.57**	yes	AT4G31840.1	Early Nodulin-Like Protein 15	
Glyma10g37080	894.07	33.67	**-4.73**	yes			4-Alpha-Glucanotransferase
Glyma10g35871	1092.42	38.21	**-4.84**	yes			Auxin Down Regulated Protein (Elongation Factor 1?)
Glyma20g33450	25.93	0.77	**-5.08**	yes	AT2G07050.1	Cycloartenol Synthase 1	
Glyma08g21420	601.82	17.72	**-5.09**	yes	AT4G25150.1	HAD Superfamily, Subfamily Iiib Acid Phosphatase	

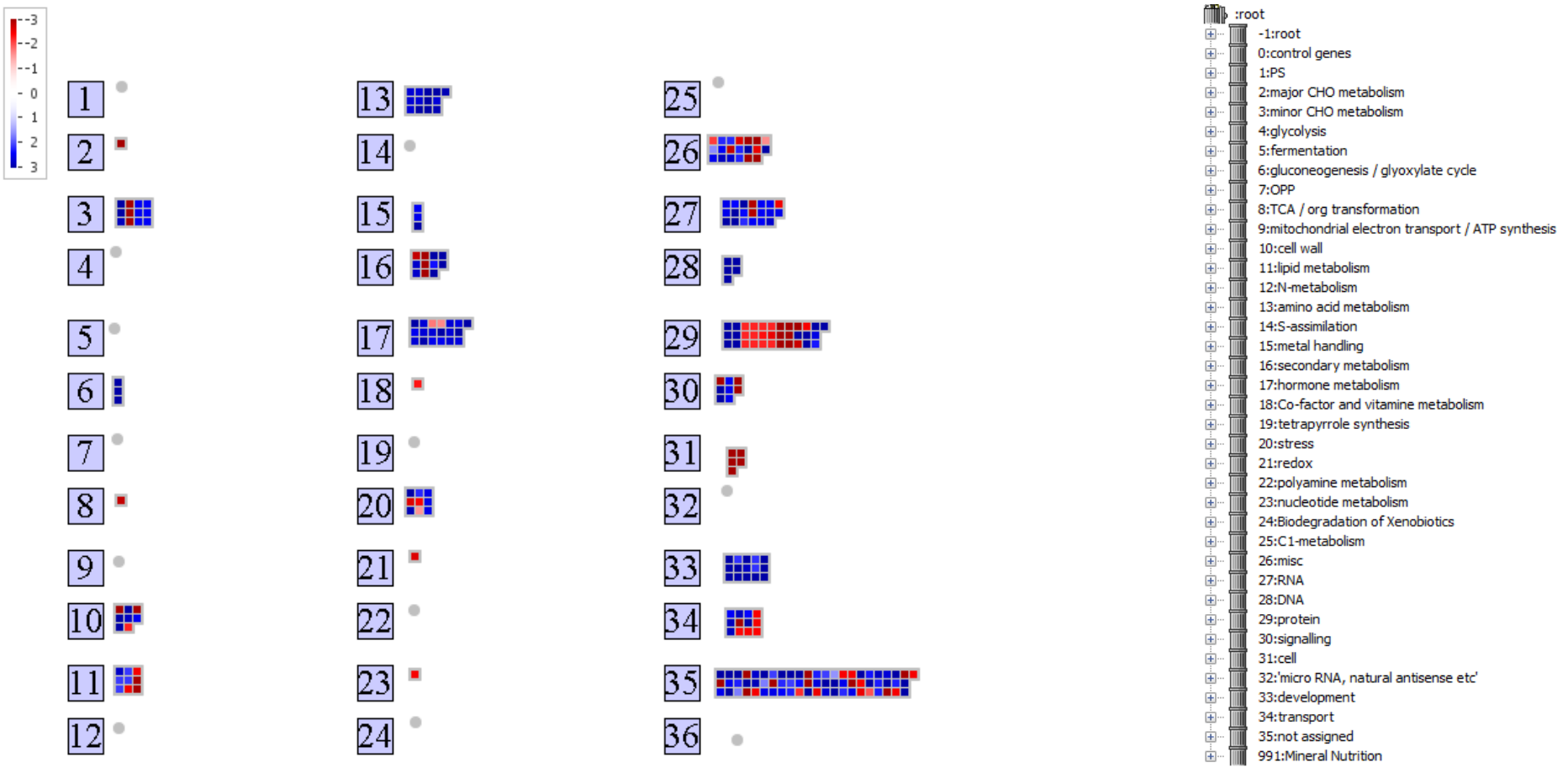


Figure 3.15 Bin maps generated by MapMan groups of differentially regulated genes between 8 and 14 weeks of nodule development. All genes shown had significant ($p \leq 0.05$) change in expression between the two respective time points. The degree of change is depicted based on the colour scale with red indicating down-regulation and blue indicating up-regulation.

Between 4 and 14 weeks of nodule development (Table 3.4), a total of 328 genes were significantly ($p \leq 0.05$) changed in expression between the time-points. Although similar gene expression patterns were found at the onset of senescence, the alpha vacuolar-processing enzyme Glyma14g10620 and the Kunitz-family trypsin protease inhibitor Glyma08g45520 had increased expression, but with Glyma08g21420 having a decreased expression. When comparing developing nodules (4 weeks) with senescing nodules (14 weeks), 207 genes were significantly ($p \leq 0.05$) up-regulated with the greatest significant ($p \leq 0.05$) change of nearly 6.5 $\log_2$-fold up-regulation for Glyma13g01140, a xyloglucan endo-transglucosylase/hydrolase family protein. This protein is proposed to be involved in the construction of cell walls of growing tissues and root hairs, as well as in adaptive changes during morphogenesis and in the root differentiation zone after cell expansion (Maris et al. 2009). The strictosidine synthase protein, Glyma16g22650, had a 6.0 $\log_2$-fold up-regulation. This protein is part of the calcium-dependent phospho-tri-esterase superfamily involved in the activity of strictosidine synthase during the flavonoid biosynthetic pathway, as well as during basal and inducible plant defence responses (Sohani et al. 2009). Other genes of interest detected included the gamma- (Glyma17g34900) and 2 beta- (Glyma05g04230 and Glyma17g14680) vacuolar-processing enzymes which had almost 4 and approximately 3 $\log_2$-fold expression increase, respectively. A further member of the cysteine protease superfamily, Glyma18g51533, also had a significant increased expression. Such members of the C48 subfamily of cysteine proteases are Small Ubiquitin-like Modifier-type (SUMO) proteases involved in SUMO modification during importing or exporting of protein to the nucleus during cell growth control (Ihara et al. 2007). The FTSH protease 8, Glyma15g17070, also had a 2 $\log_2$-fold expression increase. This metalloprotease family is involved in the biosynthesis of the thylakoid membrane, as well as the repair of photo-

damaged proteins in the thylakoid membrane, such as photosystem 2, following damage by photo-inhibition (Sakamoto et al. 2003, Zaltsman et al. 2005).

A total of 121 of genes were significantly ($p \leq 0.05$) down-regulated with the greatest significant ($p \leq 0.05$) change of 6.0 $\log_2$-fold down-regulation for Glyma08g21420, a member of the HAD superfamily, subfamily IIIB. This is an acid phosphatase protein. A 6.0 $\log_2$-fold down-regulation was found for Glyma10g35871 (unknown function). Other proteases that were down-regulated included the serine carboxypeptidase-like 51 (Glyma13g39600) and the serine carboxypeptidase-like 20 (Glyma18g47820), both had approximately 2.5 $\log_2$-fold expression decrease. Furthermore, the subtilase family proteins, Glyma14g06970 and Glyma17g14270, had $\log_2$-fold expression reductions of about 2.5 and 3.5, respectively. The Kunitz-family trypsin and protease inhibitor Glyma16g33710 had a further decrease in expression of approximately 2.7 $\log_2$-fold. The bin maps in Figure 3.16 show the distribution of these genes with differential gene regulation grouped into broad physiological functions.

Table 3.4 Differentially regulated genes after 4 and 14 weeks of nodule development after significance and FDR cut-offs have been applied. Proteases or protease inhibitors are indicated by bold letters in the Gene IDs column.

Gene ID	FPKM		Log_2-fold	Significant	TAIR	TAIR Define	Closest Blast Hit Annotation
	4w	14w					
Glyma13g01140	4.81	424.47	**6.46**	yes	AT5G57560.1	Xyloglucan Endo-transglucosylase/Hydrolase Family Protein	
Glyma16g22650	3.44	226.76	**6.04**	yes	AT3G51430.1	Calcium-Dependent Phosphotriesterase Superfamily Protein	
Glyma19g02580	1.56	82.63	**5.73**	yes	AT5G39820.1	NAC Domain Containing Protein 94	
Glyma06g46450	1.21	46.03	**5.26**	yes	AT2G32540.1	Cellulose Synthase-Like B4	
Glyma07g04810	3.28	114.78	**5.13**	yes	AT1G54070.1	Dormancy/Auxin Associated Family Protein	
Glyma04g19030	122.62	3526.55	**4.85**	yes	AT2G26670.1	Plant HaemOxygenase (Decyclizing) Family Protein	
Glyma16g31290	9.78	235.84	**4.59**	yes	AT3G61260.1	Remorin Family Protein	
Glyma13g34520	3.35	75.51	**4.49**	yes	AT1G78860.1	D-Mannose Binding Lectin Protein With Apple-Like Carbohydrate-Binding Domain	
Glyma12g35070	9.42	211.80	**4.49**	yes	AT4G27450.1	Aluminium Induced Protein With YGL And LRDR Motifs	
Glyma08g16290	63.87	1403.67	**4.46**	yes			UNKNOWN
Glyma17g01170	3.98	86.49	**4.44**	yes	AT4G17280.1	Auxin-Responsive Family Protein	
Glyma02g12220	5.29	114.24	**4.43**	yes	AT1G69490.1	NAC-Like, Activated By AP3/PI	
Glyma07g19060	2.57	55.00	**4.42**	yes	AT5G01600.1	Ferretin 1	
Glyma06g03420	3.90	79.37	**4.35**	yes	AT1G19530.1		UNKNOWN
Glyma14g03410	2.15	43.29	**4.33**	yes	AT1G06570.1	Phytoene Desaturation 1	
Glyma14g10620	5.68	109.64	**4.27**	yes	AT2G25940.1	Alpha-Vacuolar Processing Enzyme	
Glyma13g35550	3.30	61.63	**4.22**	yes	AT3G15500.1	NAC Domain Containing Protein 3	
Glyma13g39240	3.16	57.47	**4.18**	yes	AT2G27480.1	Calcium-Binding EF-Hand Family Protein	
Glyma04g08150	2.37	42.91	**4.18**	yes	AT4G31240.1	Protein Kinase C-Like Zinc Finger Protein	
Glyma11g02980	3.73	64.38	**4.11**	yes	AT4G22920.1	Non-Yellowing 1	
Glyma19g28476	4.54	78.27	**4.11**	yes	AT3G04070.2	NAC Domain Containing Protein 47	
Glyma15g07710	1.22	20.90	**4.09**	yes			Pf12143 - Protein Of Unknown Function (Duf_B2219), Pf00264 - Common Central Domain Of Tyrosinase, Pf12142 - Polyphenol Oxidase Middle Domain
Glyma13g35480	3.50	59.45	**4.09**	yes	AT4G27450.1	Aluminium Induced Protein With YGL And LRDR Motifs	
Glyma19g25070	3.56	60.44	**4.08**	yes	AT5G20090.1	Uncharacterised Protein Family (UPF0041)	

Glyma16g01390	4.10	68.24	**4.06**	yes	AT1G54070.1	Dormancy/Auxin Associated Family Protein	
Glyma19g28970	3.82	61.95	**4.02**	yes	AT5G18130.1		UNKNOWN
Glyma08g03680	1.61	25.19	**3.97**	yes	AT5G56550.1	Oxidative Stress 3	
Glyma08g45520	17.16	266.59	**3.96**	yes	AT1G17860.1	Kunitz Family Trypsin And Protease Inhibitor Protein	
Glyma03g06420	5.06	77.39	**3.93**	yes	AT5G01600.1	Ferretin 1	
Glyma08g04110	2.16	32.99	**3.93**	yes	AT4G29100.1	Basic Helix-Loop-Helix (Bhlh) DNA-Binding Superfamily Protein	
Glyma17g34900	41.10	608.95	**3.89**	yes	AT4G32940.1	Gamma Vacuolar Processing Enzyme	
Glyma03g07972	8.39	118.13	**3.82**	yes			UNKNOWN
Glyma07g33050	6.41	90.11	**3.81**	yes	AT2G15890.1	Maternal Effect Embryo Arrest 14	
Glyma16g04740	3.91	54.82	**3.81**	yes	AT3G04070.2	NAC Domain Containing Protein 47	
Glyma09g33650	5.29	74.05	**3.81**	yes	AT4G37870.1	PhosphoenolpyruvateCarboxykinase 1	
Glyma06g12701	4.34	60.34	**3.80**	yes	AT1G21450.1	SCARECROW-Like 1	
Glyma17g11781	1.63	22.07	**3.76**	yes			Leucine-Zipper-Like Protein Induced To Salt Stress
Glyma13g00280	3.53	47.17	**3.74**	yes	AT4G25810.1	Xyloglucan Endo-transglycosylase 6	
Glyma04g01130	46.33	614.80	**3.73**	yes	AT1G20440.1	Cold-Regulated 47	
Glyma09g34140	3.87	51.03	**3.72**	yes	AT1G32170.1	Xyloglucan Endo-transglucosylase/Hydrolase 30	
Glyma08g16280	5.48	71.63	**3.71**	yes	AT2G32940.1	Argonaute Family Protein	
Glyma18g07090	7.26	90.15	**3.63**	yes	AT3G47160.2	RING/U-Box Superfamily Protein	
Glyma08g04860	2.47	30.60	**3.63**	yes	AT1G14890.1	Plant Invertase/Pectin Methylesterase Inhibitor Superfamily Protein	
Glyma11g15180	7.77	93.51	**3.59**	yes	AT3G46130.1	MYB Domain Protein 48	
Glyma17g02600	2.76	33.17	**3.59**	yes	AT3G06390.1	Uncharacterised Protein Family (UPF0497)	
Glyma04g40640	2.32	27.46	**3.57**	yes	AT5G24470.1	Pseudo-Response Regulator 5	
Glyma15g42260	0.00	54.21	**3.57**	yes	AT2G31945.1		UNKNOWN
Glyma01g38870	3.18	37.29	**3.55**	yes	AT4G31940.1	Cytochrome P450, Family 82, Subfamily C, Polypeptide 4	
Glyma04g01390	6.18	70.40	**3.51**	yes	AT2G21620.1	Adenine Nucleotide Alpha Hydrolases-Like Superfamily Protein	
Glyma08g47520	6.19	70.47	**3.51**	yes	AT5G13180.1	NAC Domain Containing Protein 83	
Glyma15g08300	75.29	826.80	**3.46**	yes	AT1G28330.1	Dormancy-Associated Protein-Like 1	
Glyma16g04440	3.91	42.44	**3.44**	yes	AT5G18130.1		UNKNOWN
Glyma01g02215	4.78	51.91	**3.44**	yes			UNKNOWN
Glyma06g17910	4.09	44.25	**3.43**	yes	AT3G48530.1	SNF1-Related Protein Kinase Regulatory Subunit Gamma 1	
Glyma04g05280	2.86	30.37	**3.41**	yes	AT2G25930.1	Hydroxyproline-Rich Glycoprotein Family Protein	
Glyma02g40640	3.70	38.91	**3.39**	yes	AT5G16370.1	Acyl Activating Enzyme 5	

Glyma01g11390	9.88	103.26	**3.39**	yes	AT4G17900.1	PLATZ Transcription Factor Family Protein	
Glyma08g20090	6.84	69.93	**3.35**	yes	AT1G60940.1	SNF1-Related Protein Kinase 2.10	
Glyma06g03430	0.00	44.79	**3.32**	yes	AT1G19530.1		UNKNOWN
Glyma06g29530	95.31	948.88	**3.32**	yes	AT2G26670.1	Plant HaemOxygenase (Decyclizing) Family Protein	
Glyma07g25390	9.34	92.87	**3.31**	yes	AT1G03400.1	2-Oxoglutarate (2OG) And Fe(II)-Dependent Oxygenase Superfamily Protein	
Glyma11g11640	3.60	35.68	**3.31**	yes	AT1G32700.1	PLATZ Transcription Factor Family Protein	
Glyma06g35585	15.95	157.23	**3.30**	yes	AT5G53970.1	Tyrosine Transaminase Family Protein	
Glyma14g06160	3.72	36.64	**3.30**	yes	AT2G40300.1	Ferritin 4	
Glyma17g05920	14.06	135.21	**3.27**	yes	AT1G17590.1	Nuclear Factor Y, Subunit A8	
Glyma10g01690	4.26	40.69	**3.26**	yes	AT3G62550.1	Adenine Nucleotide Alpha Hydrolases-Like Superfamily Protein	
Glyma17g11500	4.42	41.78	**3.24**	yes	AT5G11090.1	Serine-Rich Protein-Related	
Glyma04g04270	4.30	40.61	**3.24**	yes	AT5G07850.1	HXXXD-Type Acyl-Transferase Family Protein	
Glyma11g16070	3.73	35.14	**3.24**	yes	AT5G13800.1	Pheophytinase	
Glyma05g27970	2.43	22.77	**3.23**	yes	AT1G13710.1	Cytochrome P450, Family 78, Subfamily A, Polypeptide 5	
Glyma17g14680	19.48	179.92	**3.21**	yes	AT1G62710.1	Beta Vacuolar Processing Enzyme	
Glyma18g43650	17.31	157.83	**3.19**	yes	AT5G01600.1	Ferretin 1	
Glyma08g26150	20.06	182.47	**3.19**	yes	AT1G72160.1	Sec14p-Like Phosphatidylinositol Transfer Family Protein	
Glyma04g02340	3.39	30.74	**3.18**	yes	AT5G42200.1	RING/U-Box Superfamily Protein	
Glyma17g11590	19.14	170.46	**3.15**	yes	AT1G15740.1	Leucine-Rich Repeat Family Protein	
Glyma05g33140	4.41	38.93	**3.14**	yes	AT5G63620.1	Groes-Like Zinc-Binding Alcohol Dehydrogenase Family Protein	
Glyma20g28720	3.20	28.18	**3.14**	yes	AT3G22740.1	Homocysteine S-Methyltransferase 3	
Glyma16g34320	4.02	35.20	**3.13**	yes	AT1G53345.1		UNKNOWN
Glyma18g15001	5.11	44.55	**3.13**	yes			UNKNOWN
Glyma07g00873	3.89	33.79	**3.12**	yes	AT1G55020.1	Lipoxygenase 1	
Glyma13g25273	9.89	85.69	**3.12**	yes	AT1G61120.1	Terpene Synthase 04	
Glyma08g14810	19.63	168.92	**3.11**	yes	AT5G65207.1		UNKNOWN
Glyma04g43000	2.76	23.54	**3.09**	yes	AT4G08290.2	Nodulin Mtn21 /Eama-Like Transporter Family Protein	
Glyma04g03330	4.99	42.51	**3.09**	yes	AT1G19530.1		UNKNOWN
Glyma17g11705	2.55	21.70	**3.09**	yes	AT1G15690.1	Inorganic H Pyrophosphatase Family Protein	
Glyma13g23170	22.37	186.88	**3.06**	yes	AT1G15690.1	Inorganic H Pyrophosphatase Family Protein	
Glyma04g40170	5.41	45.00	**3.06**	yes	AT3G10870.1	Methyl Esterase 17	
Glyma19g42060	32.18	264.71	**3.04**	yes	AT1G56220.3	Dormancy/Auxin Associated Family Protein	

Glyma04g08910	2.97	24.43	**3.04**	yes	AT4G32810.1	Carotenoid Cleavage Dioxygenase 8	
Glyma02g06990	0.00	35.90	**3.03**	yes	AT2G23620.1	Methyl Esterase 1	
Glyma19g39370	20.12	164.17	**3.03**	yes	AT1G48910.1	Flavin-Containing Monooxygenase Family Protein	
Glyma02g07130	11.48	93.69	**3.03**	yes	AT2G02990.1	Ribonuclease 1	
Glyma13g01570	2.85	22.46	**2.98**	yes	AT4G30420.1	Nodulin Mtn21 /Eama-Like Transporter Family Protein	
Glyma06g17020	1.61	12.59	**2.97**	yes	AT5G51070.1	ClpAtpase	
Glyma18g07220	2.20	17.15	**2.96**	yes	AT3G54140.1	Peptide Transporter 1	
Glyma03g19260	19.06	147.46	**2.95**	yes	AT3G09390.1	Metallothionein 2A	
Glyma08g22900	12.11	92.72	**2.94**	yes	AT1G15740.1	Leucine-Rich Repeat Family Protein	
Glyma15g10840	4.23	32.30	**2.93**	yes	AT3G23880.1	F-Box And Associated Interaction Domains-Containing Protein	
Glyma11g05030	18.14	138.49	**2.93**	yes	AT4G37680.1	Heptahelical Protein 4	
Glyma10g38080	5.91	45.01	**2.93**	yes	AT5G23750.2	Remorin Family Protein	
Glyma06g05280	22.51	171.23	**2.93**	yes	AT1G10070.1	Branched-Chain Amino Acid Transaminase 2	
Glyma19g43150	17.16	129.40	**2.91**	yes	AT4G14880.1	O-Acetylserine (Thiol) Lyase (OAS-TL) Isoform A1	
Glyma03g31210	5.91	44.40	**2.91**	yes	AT3G63210.1	Protein Of Unknown Function (DUF581)	
Glyma17g04920	5.51	41.32	**2.91**	yes	AT4G33150.1	Lysine-KetoglutarateReductase/Saccharopine Dehydrogenase Bifunctional Enzyme	
Glyma08g36641	40.03	299.54	**2.90**	yes	AT1G32700.1	PLATZ Transcription Factor Family Protein	
Glyma17g37020	5.61	41.69	**2.89**	yes	AT1G19530.1		UNKNOWN
Glyma16g27760	1.57	11.66	**2.89**	yes	AT1G71710.2	Dnase I-Like Superfamily Protein	
Glyma07g16420	36.50	271.07	**2.89**	yes	AT2G37750.1		UNKNOWN
Glyma13g25266	8.51	62.81	**2.88**	yes	AT1G61120.1	Terpene Synthase 04	
Glyma08g43330	25.46	187.88	**2.88**	yes	AT5G10770.1	Eukaryotic Aspartyl Protease Family Protein	
Glyma13g16770	8.87	63.85	**2.85**	yes	AT1G72830.1	Nuclear Factor Y, Subunit A3	
Glyma08g23590	3.72	26.53	**2.83**	yes	AT4G30780.1		UNKNOWN
Glyma13g00840	2.28	16.01	**2.81**	yes	AT4G30190.1	H(+)-Atpase 2	
Glyma15g11680	4.79	33.54	**2.81**	yes	AT1G62300.1	WRKY Family Transcription Factor	
Glyma09g41340	3.91	27.28	**2.80**	yes	AT5G58380.1	SOS3-Interacting Protein 1	
Glyma13g44400	4.00	27.79	**2.80**	yes	AT1G15740.1	Leucine-Rich Repeat Family Protein	
Glyma02g15430	22.39	155.22	**2.79**	yes	AT2G15890.1	Maternal Effect Embryo Arrest 14	
Glyma02g02560	24.69	171.17	**2.79**	yes	AT1G07040.1		UNKNOWN
Glyma05g04230	5.04	34.81	**2.79**	yes	AT1G62710.1	Beta Vacuolar Processing Enzyme	
Glyma14g36690	22.14	150.77	**2.77**	yes	AT3G49570.1	Response To Low Sulfur 3	

Glyma18g52070	15.46	104.99	**2.76**	yes	AT1G27980.1	Dihydrosphingosine Phosphate Lyase	
Glyma18g00590	12.54	84.22	**2.75**	yes	AT4G14030.1	Selenium-Binding Protein 1	
Glyma11g10340	3.74	25.10	**2.75**	yes	AT4G38810.2	Calcium-Binding EF-Hand Family Protein	
Glyma04g34550	2.93	19.52	**2.74**	yes	AT1G74780.1	Nodulin-Like / Major Facilitator Superfamily Protein	
Glyma08g03310	4.81	31.84	**2.73**	yes	AT2G19590.1	ACC Oxidase 1	
Glyma09g35680	6.99	45.80	**2.71**	yes	AT1G64970.1	Gamma-TocopherolMethyltransferase	
Glyma10g25710	3.53	23.10	**2.71**	yes	AT1G04620.1	Coenzyme F420 Hydrogenase Family / Dehydrogenase, Beta Subunit Family	
Glyma08g20220	11.89	77.80	**2.71**	yes	AT1G55020.1	Lipoxygenase 1	
Glyma18g15570	3.17	20.73	**2.71**	yes	AT2G02870.1	Galactose Oxidase/Kelch Repeat Superfamily Protein	
Glyma04g35560	3.32	21.52	**2.70**	yes	AT1G23550.1	Similar To RCD One 2	
Glyma15g01610	17.49	113.16	**2.69**	yes	AT1G80440.1	Galactose Oxidase/Kelch Repeat Superfamily Protein	
Glyma12g00390	21.84	140.74	**2.69**	yes	AT1G72160.1	Sec14p-Like Phosphatidylinositol Transfer Family Protein	
Glyma11g12980	63.66	408.12	**2.68**	yes	AT2G16060.1	Hemoglobin 1	
Glyma18g46500	1.74	11.10	**2.67**	yes	AT4G11610.1	C2 Calcium/Lipid-Binding Plant Phosphoribosyltransferase Family Protein	
Glyma13g06930	4.62	29.22	**2.66**	yes	AT5G38820.1	Transmembrane Amino Acid Transporter Family Protein	
Glyma07g08011	5.71	36.02	**2.66**	yes	AT1G01770.1		UNKNOWN
Glyma10g41120	0.00	26.50	**2.66**	yes	AT2G01300.1		UNKNOWN
Glyma09g00820	6.11	38.15	**2.64**	yes	AT1G62300.1	WRKY Family Transcription Factor	
Glyma04g14680	5.51	33.82	**2.62**	yes	AT3G45900.1	Ribonuclease P Protein Subunit P38-Related	
Glyma13g40180	9.36	57.16	**2.61**	yes	AT5G60570.1	Galactose Oxidase/Kelch Repeat Superfamily Protein	
Glyma04g37140	10.65	65.07	**2.61**	yes	AT3G48530.1	SNF1-Related Protein Kinase Regulatory Subunit Gamma 1	
Glyma12g16571	0.00	25.43	**2.61**	yes	AT1G11530.1	C-Terminal Cysteine Residue Is Changed To A Serine 1	
Glyma18g53954	10.01	60.70	**2.60**	yes	AT5G13180.1	NAC Domain Containing Protein 83	
Glyma02g43040	16.64	100.62	**2.60**	yes	AT2G40300.1	Ferritin 4	
Glyma10g33030	44.32	267.67	**2.59**	yes	AT3G54700.1	Phosphate Transporter 1;7	
Glyma11g12560	4.99	29.72	**2.57**	yes	AT1G63850.1	BTB/POZ Domain-Containing Protein	
Glyma15g16810	6.32	37.59	**2.57**	yes	AT1G30900.1	Vacuolar Sorting Receptor 6	
Glyma13g42340	2.46	14.55	**2.57**	yes	AT1G55020.1	Lipoxygenase 1	
Glyma07g02420	3.77	22.27	**2.56**	yes	AT2G24100.1		UNKNOWN
Glyma02g44380	29.99	173.10	**2.53**	yes	AT2G26980.4	CBL-Interacting Protein Kinase 3	
Glyma07g03200	8.23	47.21	**2.52**	yes	AT1G15740.1	Leucine-Rich Repeat Family Protein	
Glyma17g04220	5.42	30.92	**2.51**	yes	AT1G48040.1	Protein Phosphatase 2C Family Protein	

Glyma19g29350	6.10	34.34	**2.49**	yes	AT5G39890.1	Protein Of Unknown Function (DUF1637)	
Glyma12g29630	6.14	34.49	**2.49**	yes	AT5G60570.1	Galactose Oxidase/Kelch Repeat Superfamily Protein	
Glyma09g07100	5.75	31.90	**2.47**	yes	AT4G26140.2	Beta-Galactosidase 12	
Glyma02g15370	7.18	39.73	**2.47**	yes	AT3G19000.2	2-Oxoglutarate (2OG) And Fe(II)-Dependent Oxygenase Superfamily Protein	
Glyma04g05550	8.56	47.39	**2.47**	yes	AT5G11950.1	Putative Lysine Decarboxylase Family Protein	
Glyma19g34060	7.20	39.20	**2.44**	yes	AT3G63210.1	Protein Of Unknown Function (DUF581)	
Glyma06g42980	2.47	13.43	**2.44**	yes	AT2G30990.1	Protein Of Unknown Function (DUF688)	
Glyma11g09120	12.86	69.88	**2.44**	yes	AT4G35560.1	Transducin/WD40 Repeat-Like Superfamily Protein	
Glyma08g12235	0.00	22.15	**2.44**	yes			UNKNOWN
Glyma13g35220	6.24	33.57	**2.43**	yes	AT1G53035.1		UNKNOWN
Glyma19g01470	8.61	46.32	**2.43**	yes	AT3G01420.1	Peroxidase Superfamily Protein	
Glyma08g08280	3.71	19.94	**2.43**	yes	AT1G36990.1		UNKNOWN
Glyma15g23420	15.93	85.32	**2.42**	yes	AT1G29400.1	MEI2-Like Protein 5	
Glyma06g14020	7.62	40.80	**2.42**	yes	AT2G46780.1	RNA-Binding (RRM/RBD/RNP Motifs) Family Protein	
Glyma01g04730	5.84	31.14	**2.41**	yes	AT1G27100.1	Actin Cross-Linking Protein	
Glyma09g30320	12.54	66.54	**2.41**	yes	AT3G17810.1	Pyrimidine 1	
Glyma13g27060	34.63	183.66	**2.41**	yes	AT1G62620.1	Flavin-Binding Monooxygenase Family Protein	
Glyma18g46560	15.33	79.66	**2.38**	yes			UNKNOWN
Glyma04g33280	11.23	58.17	**2.37**	yes	AT1G74440.1	Protein Of Unknown Function (DUF962)	
Glyma05g35880	19.82	100.88	**2.35**	yes	AT4G29120.1	6-Phosphogluconate Dehydrogenase Family Protein	
Glyma20g03790	8.02	40.62	**2.34**	yes	AT3G11810.1		UNKNOWN
Glyma10g35550	8.09	40.98	**2.34**	yes	AT2G39220.1	PATATIN-Like Protein 6	
Glyma13g39450	12.52	63.16	**2.33**	yes	AT4G33790.1	Jojoba Acyl CoaReductase-Related Male Sterility Protein	
Glyma17g10520	11.31	56.63	**2.32**	yes	AT5G62740.1	SPFH/Band 7/PHB Domain-Containing Membrane-Associated Protein Family	
Glyma16g01610	17.06	84.50	**2.31**	yes	AT3G14280.1		UNKNOWN
Glyma08g41020	8.74	42.75	**2.29**	yes	AT2G02870.1	Galactose Oxidase/Kelch Repeat Superfamily Protein	
Glyma18g51533	3.52	17.15	**2.29**	yes	AT1G09730.1	Cysteine Proteinases Superfamily Protein	
Glyma08g03540	13.48	65.47	**2.28**	yes	AT2G19810.1	CCCH-Type Zinc Finger Family Protein	
Glyma04g42090	6.79	32.89	**2.28**	yes	AT1G21450.1	SCARECROW-Like 1	
Glyma08g04370	19.46	94.30	**2.28**	yes	AT3G24503.1	Aldehyde Dehydrogenase 2C4	
Glyma06g04060	5.16	24.91	**2.27**	yes	AT4G37080.2	Protein Of Unknown Function, DUF547	
Glyma15g08480	10.91	52.30	**2.26**	yes	AT4G28500.1	NAC Domain Containing Protein 73	

Glyma15g17070	7.49	35.76	**2.26**	yes	AT1G06430.1	FTSH Protease 8	
Glyma06g08100	22.01	104.29	**2.24**	yes	AT5G24890.1		UNKNOWN
Glyma07g05110	17.39	81.22	**2.22**	yes	AT3G14280.1		UNKNOWN
Glyma17g10490	12.23	56.94	**2.22**	yes	AT3G01470.1	Homeobox 1	
Glyma09g30370	29.91	139.13	**2.22**	yes	AT5G37600.1	Glutamine Synthase Clone R1	
Glyma14g27290	6.67	30.56	**2.20**	yes	AT1G21450.1	SCARECROW-Like 1	
Glyma17g10050	61.75	280.94	**2.19**	yes	AT5G14920.1	Gibberellin-Regulated Family Protein	
Glyma04g08050	27.69	124.85	**2.17**	yes	AT5G24890.1		UNKNOWN
Glyma18g09000	5.60	23.59	**2.07**	yes	AT3G60970.1	Multidrug Resistance-Associated Protein 15	
Glyma10g36760	0.00	15.71	**2.05**	yes	AT5G61890.1	Integrase-Type DNA-Binding Superfamily Protein	
Glyma17g08890	0.00	15.20	**2.01**	yes	AT5G60910.1	AGAMOUS-Like 8	
Glyma08g15993	0.00	14.26	**1.95**	yes	AT2G44480.1	Beta Glucosidase 17	
Glyma13g01150	0.00	14.25	**1.95**	yes	AT5G57560.1	Xyloglucan Endo-transglucosylase/Hydrolase Family Protein	
Glyma13g08060	0.00	14.10	**1.93**	yes	AT2G14095.1		UNKNOWN
Glyma02g17450	0.00	13.37	**1.88**	yes			UNKNOWN
Glyma20g23180	0.00	13.01	**1.85**	yes	AT2G37430.1	C2H2 And C2HC Zinc Fingers Superfamily Protein	
Glyma13g23161	0.00	11.57	**1.73**	yes	AT4G39140.1	RING/U-Box Superfamily Protein	
Glyma11g01990	0.00	11.02	**1.68**	yes	AT4G23880.1		UNKNOWN
Glyma14g09051	0.00	11.01	**1.68**	yes	AT2G43290.1	Calcium-Binding EF-Hand Family Protein	
Glyma03g28234	0.00	9.90	**1.58**	yes	AT2G32560.1	F-Box Family Protein	
Glyma17g17120	0.00	9.89	**1.57**	yes	AT5G65870.1	Phytosulfokine 5 Precursor	
Glyma16g01570	0.00	9.65	**1.55**	yes			UNKNOWN
Glyma09g12860	0.00	9.15	**1.50**	yes	AT5G25240.1		UNKNOWN
Glyma06g08860	11.32	0.00	**-1.71**	yes	AT4G28940.1	Phosphorylase Superfamily Protein	
Glyma20g32470	11.48	0.00	**-1.72**	yes			Extensin-Like Protein
Glyma12g35431	12.01	0.00	**-1.77**	yes	AT5G53880.1		UNKNOWN
Glyma03g28740	12.09	0.00	**-1.77**	yes			UNKNOWN
Glyma07g19360	12.14	0.00	**-1.78**	yes	AT1G58170.1	Disease Resistance-Responsive (Dirigent-Like Protein) Family Protein	
Glyma10g37370	38.42	9.00	**-2.09**	yes	AT5G24030.1	SLAC1 Homologue 3	
Glyma09g10154	16.45	0.00	**-2.10**	yes			UNKNOWN
Glyma19g03490	27.31	6.27	**-2.12**	yes	AT3G29320.1	GlycosylTransferase, Family 35	
Glyma20g23350	42.85	9.70	**-2.14**	yes	AT1G50380.1	ProlylOligopeptidase Family Protein	

Glyma13g16280	13.49	3.03	**-2.15**	yes	AT5G56890.1	Protein Kinase Superfamily Protein	
Glyma08g45820	169.28	36.62	**-2.21**	yes	AT5G48490.1	Bifunctional Inhibitor/Lipid-Transfer Protein/Seed Storage 2S Albumin Superfamily Protein	
Glyma15g34720	54.95	11.72	**-2.23**	yes	AT4G34131.1	UDP-GlucosylTransferase 73B3	
Glyma05g32150	42.89	9.13	**-2.23**	yes	AT5G64700.1	Nodulin Mtn21 /Eama-Like Transporter Family Protein	
Glyma20g02400	123.57	25.41	**-2.28**	yes	AT2G35120.1	Single Hybrid Motif Superfamily Protein	
Glyma11g17160	19.54	0.00	**-2.29**	yes	AT1G65680.1	Expansin B2	
Glyma08g41940	49.82	10.15	**-2.30**	yes	AT1G06660.1		UNKNOWN
Glyma19g42610	60.48	12.29	**-2.30**	yes	AT5G27380.1	Glutathione Synthetase 2	
Glyma06g17860	28.70	5.74	**-2.32**	yes	AT5G50400.1	Purple Acid Phosphatase 27	
Glyma02g10240	229.19	45.83	**-2.32**	yes	AT2G39530.1	Uncharacterised Protein Family (UPF0497)	
Glyma13g36930	49.46	9.73	**-2.35**	yes	AT2G20520.1	FASCICLIN-Like Arabinogalactan 6	
Glyma08g41640	20.46	0.00	**-2.35**	yes			UNKNOWN
Glyma12g10850	20.48	0.00	**-2.35**	yes	AT1G05260.1	Peroxidase Superfamily Protein	
Glyma10g35080	3401.55	667.31	**-2.35**	yes			UNKNOWN
Glyma12g10300	31.27	6.10	**-2.36**	yes	AT2G32530.1	Cellulose Synthase-Like B3	
Glyma20g22525	23.33	4.55	**-2.36**	yes	AT1G09460.1	Carbohydrate-Binding X8 Domain Superfamily Protein	
Glyma11g37360	41.65	8.08	**-2.37**	yes	AT3G04120.1	Glyceraldehyde-3-Phosphate Dehydrogenase C Subunit 1	
Glyma10g37960	34.10	6.52	**-2.39**	yes	AT5G23860.1	Tubulin Beta 8	
Glyma09g36220	59.09	11.22	**-2.40**	yes	AT3G08030.1	Protein Of Unknown Function, DUF642	
Glyma20g29850	44.01	8.32	**-2.40**	yes	AT3G48990.1	AMP-Dependent Synthetase And Ligase Family Protein	
Glyma07g14460	48.00	9.02	**-2.41**	yes	AT1G11680.1	Cytochrome P450 51g1	
Glyma07g05580	29.50	5.52	**-2.42**	yes	AT1G32900.1	UDP-Glycosyltransferase Superfamily Protein	
Glyma17g17850	40.96	7.59	**-2.43**	yes	AT5G67360.1	Subtilase Family Protein	
Glyma11g05680	38.62	7.14	**-2.44**	yes	AT4G34138.1	UDP-GlucosylTransferase 73B1	
Glyma13g39600	46.41	8.39	**-2.47**	yes	AT2G27920.1	Serine Carboxypeptidase-Like 51	
Glyma04g40000	172.86	31.00	**-2.48**	yes	AT2G40610.1	Expansin A8	
Glyma06g48260	53.42	9.51	**-2.49**	yes	AT4G24010.1	Cellulose Synthase Like G1	
Glyma02g47770	23.18	0.00	**-2.49**	yes			Extensin-Like Protein
Glyma07g17440	18.12	3.19	**-2.50**	yes	AT2G38060.1	Phosphate Transporter 4;2	
Glyma18g47820	37.30	6.53	**-2.51**	yes	AT4G12910.1	Serine Carboxypeptidase-Like 20	
Glyma10g35521	45.78	7.98	**-2.52**	yes	AT2G37170.1	Plasma Membrane Intrinsic Protein 2	
Glyma02g43320	1188.93	206.72	**-2.52**	yes			Nodulin-16 Precursor

Glyma13g24490	223.21	38.47	**-2.54**	yes	AT1G07400.1	HSP20-Like Chaperones Superfamily Protein	
Glyma02g11530	38.64	6.46	**-2.58**	yes	AT4G34050.1	S-Adenosyl-L-Methionine-Dependent Methyltransferases Superfamily Protein	
Glyma08g48190	11.84	1.95	**-2.60**	yes	AT1G67040.1		UNKNOWN
Glyma11g11020	18.34	2.97	**-2.62**	yes	AT3G19820.1	Cell Elongation Protein / DWARF1 / DIMINUTO (DIM)	
Glyma13g44950	21.30	3.45	**-2.63**	yes	AT3G21240.1	4-Coumarate:Coa Ligase 2	
Glyma08g12710	24.27	3.91	**-2.63**	yes	AT1G60500.1	Dynamin Related Protein 4C	
Glyma16g18480	162.71	26.06	**-2.64**	yes	AT2G02100.1	Low-Molecular-Weight Cysteine-Rich 69	
Glyma20g28230	80.72	12.75	**-2.66**	yes	AT5G26340.1	Major Facilitator Superfamily Protein	
Glyma18g04640	12.27	1.93	**-2.67**	yes	AT5G37630.1	ARM Repeat Superfamily Protein	
Glyma13g37080	396.76	61.28	**-2.69**	yes	AT5G53970.1	Tyrosine Transaminase Family Protein	
Glyma17g13420	60.63	9.34	**-2.70**	yes	AT3G26330.1	Cytochrome P450, Family 71, Subfamily B, Polypeptide 37	
Glyma20g29840	20.26	3.07	**-2.72**	yes	AT5G23860.1	Tubulin Beta 8	
Glyma11g10830	139.77	21.14	**-2.72**	yes	AT3G63380.1	Atpase E1-E2 Type Family Protein / HaloacidDehalogenase-Like Hydrolase Family Protein	
Glyma15g11930	130.88	19.73	**-2.73**	yes	AT1G05010.1	Ethylene-Forming Enzyme	
Glyma14g06970	168.63	25.42	**-2.73**	yes	AT5G59190.1	Subtilase Family Protein	
Glyma08g00780	27.34	4.11	**-2.73**	yes	AT1G72160.1	Sec14p-Like Phosphatidylinositol Transfer Family Protein	
Glyma16g33710	70.93	10.59	**-2.74**	yes	AT1G17860.1	Kunitz Family Trypsin And Protease Inhibitor Protein	
Glyma10g37380	21.91	3.25	**-2.75**	yes	AT5G52882.1	P-Loop Containing Nucleoside Triphosphate Hydrolases Superfamily Protein	
Glyma06g01380	64.84	9.60	**-2.76**	yes	AT4G39210.1	Glucose-1-Phosphate Adenylyltransferase Family Protein	
Glyma11g21001	43.70	6.43	**-2.76**	yes	AT1G65870.1	Disease Resistance-Responsive (Dirigent-Like Protein) Family Protein	
Glyma15g15950	8.87	1.29	**-2.78**	yes	AT3G06880.2	Transducin/WD40 Repeat-Like Superfamily Protein	
Glyma10g44360	264.03	38.13	**-2.79**	yes			UNKNOWN
Glyma08g19290	114.60	16.55	**-2.79**	yes	AT2G22590.1	UDP-Glycosyltransferase Superfamily Protein	
Glyma04g08100	13.83	1.99	**-2.80**	yes	AT4G31590.1	Cellulose-Synthase-Like C5	
Glyma10g44370	1203.10	171.61	**-2.81**	yes			Hypothetical Protein - Prupe_Ppb015435mg
Glyma13g38700	13.31	1.87	**-2.83**	yes	AT3G44050.1	P-Loop Containing Nucleoside Triphosphate Hydrolases Superfamily Protein	
Glyma18g50970	42.93	6.00	**-2.84**	yes	AT5G15780.1	Pollen Ole E 1 Allergen And Extensin Family Protein	
Glyma13g38300	83.00	11.59	**-2.84**	yes	AT4G11290.1	Peroxidase Superfamily Protein	
Glyma08g11330	31.74	4.35	**-2.87**	yes	AT4G15550.1	Indole-3-Acetate Beta-D-Glucosyltransferase	
Glyma08g24610	16.22	2.21	**-2.87**	yes	AT5G14210.1	Leucine-Rich Repeat Protein Kinase Family Protein	
Glyma14g03580	625.30	83.16	**-2.91**	yes			Acid Phosphatase
Glyma18g12660	38.32	5.06	**-2.92**	yes	AT1G78570.1	Rhamnose Biosynthesis 1	

Glyma08g25950	137.66	18.15	**-2.92**	yes	AT3G14690.1	Cytochrome P450, Family 72, Subfamily A, Polypeptide 15	
Glyma15g15010	22.66	2.86	**-2.98**	yes	AT5G49460.1	ATP Citrate Lyase Subunit B 2	
Glyma06g42080	90.52	11.35	**-3.00**	yes	AT5G54770.1	Thiazole Biosynthetic Enzyme, Chloroplast (ARA6) (THI1) (THI4)	
Glyma18g49240	23.45	2.92	**-3.01**	yes	AT3G29590.1	HXXXD-Type Acyl-Transferase Family Protein	
Glyma04g09740	104.45	12.95	**-3.01**	yes	AT1G09750.1	Eukaryotic Aspartyl Protease Family Protein	
Glyma13g11090	253.25	31.00	**-3.03**	yes	AT3G22142.1	Bifunctional Inhibitor/Lipid-Transfer Protein/Seed Storage 2S Albumin Superfamily Protein	
Glyma18g14490	36.64	0.00	**-3.06**	yes			UNKNOWN
Glyma10g06390	38.70	0.00	**-3.13**	yes	AT3G09870.1	SAUR-Like Auxin-Responsive Protein Family	
Glyma05g22060	13.19	1.50	**-3.14**	yes	AT5G67360.1	Subtilase Family Protein	
Glyma10g07190	24.12	2.73	**-3.15**	yes	AT1G22060.1		UNKNOWN
Glyma12g11820	71.28	8.01	**-3.15**	yes	AT2G20840.1	Secretory Carrier Membrane Protein (SCAMP) Family Protein	
Glyma10g40780	13.47	1.44	**-3.23**	yes	AT4G34220.1	Leucine-Rich Repeat Protein Kinase Family Protein	
Glyma11g09920	28.61	2.93	**-3.29**	yes	AT5G12470.1	Protein Of Unknown Function (DUF3411)	
Glyma07g32050	171.56	17.53	**-3.29**	yes	AT1G53540.1	HSP20-Like Chaperones Superfamily Protein	
Glyma20g27280	32.86	3.31	**-3.31**	yes	AT1G04820.1	Tubulin Alpha-4 Chain	
Glyma18g50980	122.37	12.08	**-3.34**	yes	AT2G36780.1	UDP-Glycosyltransferase Superfamily Protein	
Glyma02g36580	255.92	25.09	**-3.35**	yes	AT5G57920.1	Early Nodulin-Like Protein 10	
Glyma10g34700	27.50	2.62	**-3.39**	yes	AT1G66950.1	Pleiotropic Drug Resistance 11	
Glyma08g13960	11.65	1.10	**-3.40**	yes	AT3G26380.1	Melibiase Family Protein	
Glyma20g33710	183.58	16.95	**-3.44**	yes	AT5G10030.1	TGACG Motif-Binding Factor 4	
Glyma02g10770	12.62	1.16	**-3.44**	yes	AT3G28040.1	Leucine-Rich Receptor-Like Protein Kinase Family Protein	
Glyma17g14270	23.04	2.11	**-3.45**	yes	AT1G04110.1	Subtilase Family Protein	
Glyma07g17030	112.17	9.76	**-3.52**	yes	AT5G55410.2	Bifunctional Inhibitor/Lipid-Transfer Protein/Seed Storage 2S Albumin Superfamily Protein	
Glyma08g14670	25.07	2.17	**-3.53**	yes	AT4G39800.1	Myo-Inositol-1-Phosphate Synthase 1	
Glyma06g14850	83.26	7.08	**-3.56**	yes	AT2G40610.1	Expansin A8	
Glyma16g04750	66.72	5.57	**-3.58**	yes	AT5G18280.2	Apyrase 2	
Glyma04g11400	35.22	2.85	**-3.63**	yes	AT4G08685.1	Pollen Ole E 1 Allergen And Extensin Family Protein	
Glyma02g11820	63.78	4.98	**-3.68**	yes	AT2G46210.1	Fatty Acid/SphingolipidDesaturase	
Glyma02g11440	34.37	2.61	**-3.72**	yes	AT1G49740.1	PLC-Like Phosphodiesterases Superfamily Protein	
Glyma20g06290	199.30	14.61	**-3.77**	yes	AT3G22142.1	Bifunctional Inhibitor/Lipid-Transfer Protein/Seed Storage 2S Albumin Superfamily Protein	
Glyma20g03988	27.13	1.97	**-3.78**	yes	AT1G25540.1	Phytochrome And Flowering Time Regulatory Protein (PFT1)	
Glyma11g21480	123.86	8.75	**-3.82**	yes	AT5G12380.1	Annexin 8	

Glyma08g46520	95.75	5.97	**-4.00**	yes	AT5G06900.1	Cytochrome P450, Family 93, Subfamily D, Polypeptide 1	
Glyma20g12150	33.59	2.06	**-4.03**	yes	AT5G21105.1	Plant L-Ascorbate Oxidase	
Glyma12g08990	18.05	0.96	**-4.24**	yes	AT5G22740.1	Cellulose Synthase-Like A02	
Glyma13g30950	39.06	1.94	**-4.33**	yes	AT2G33850.1		UNKNOWN
Glyma04g36520	35.73	1.60	**-4.48**	yes	AT5G63180.1	Pectin Lyase-Like Superfamily Protein	
Glyma18g52670	175.30	7.54	**-4.54**	yes	AT2G39518.1	Uncharacterised Protein Family (UPF0497)	
Glyma17g06740	14.23	0.58	**-4.61**	yes	AT4G30020.1	PA-Domain Containing Subtilase Family Protein	
Glyma01g16140	61.57	2.45	**-4.65**	yes	AT1G65680.1	Expansin B2	
Glyma17g08110	323.96	11.85	**-4.77**	yes	AT4G31840.1	Early Nodulin-Like Protein 15	
Glyma15g23830	227.84	7.24	**-4.98**	yes			UNKNOWN
Glyma10g37080	1087.17	33.67	**-5.01**	yes			4-Alpha-Glucanotransferase
Glyma18g02230	230.54	6.95	**-5.05**	yes	AT5G10180.1	Slufate Transporter 2;1	
Glyma03g08478	170.62	0.00	**-5.13**	yes			UNKNOWN
Glyma06g45700	197.06	4.00	**-5.62**	yes	AT4G15210.1	Beta-Amylase 5	
Glyma10g35871	2200.89	38.21	**-5.85**	yes			Auxin Down Regulated Protein (Elongation Factor 1?)
Glyma08g21420	1222.71	17.72	**-6.11**	yes	AT4G25150.1	HAD Superfamily, Subfamily IIIB Acid Phosphatase	

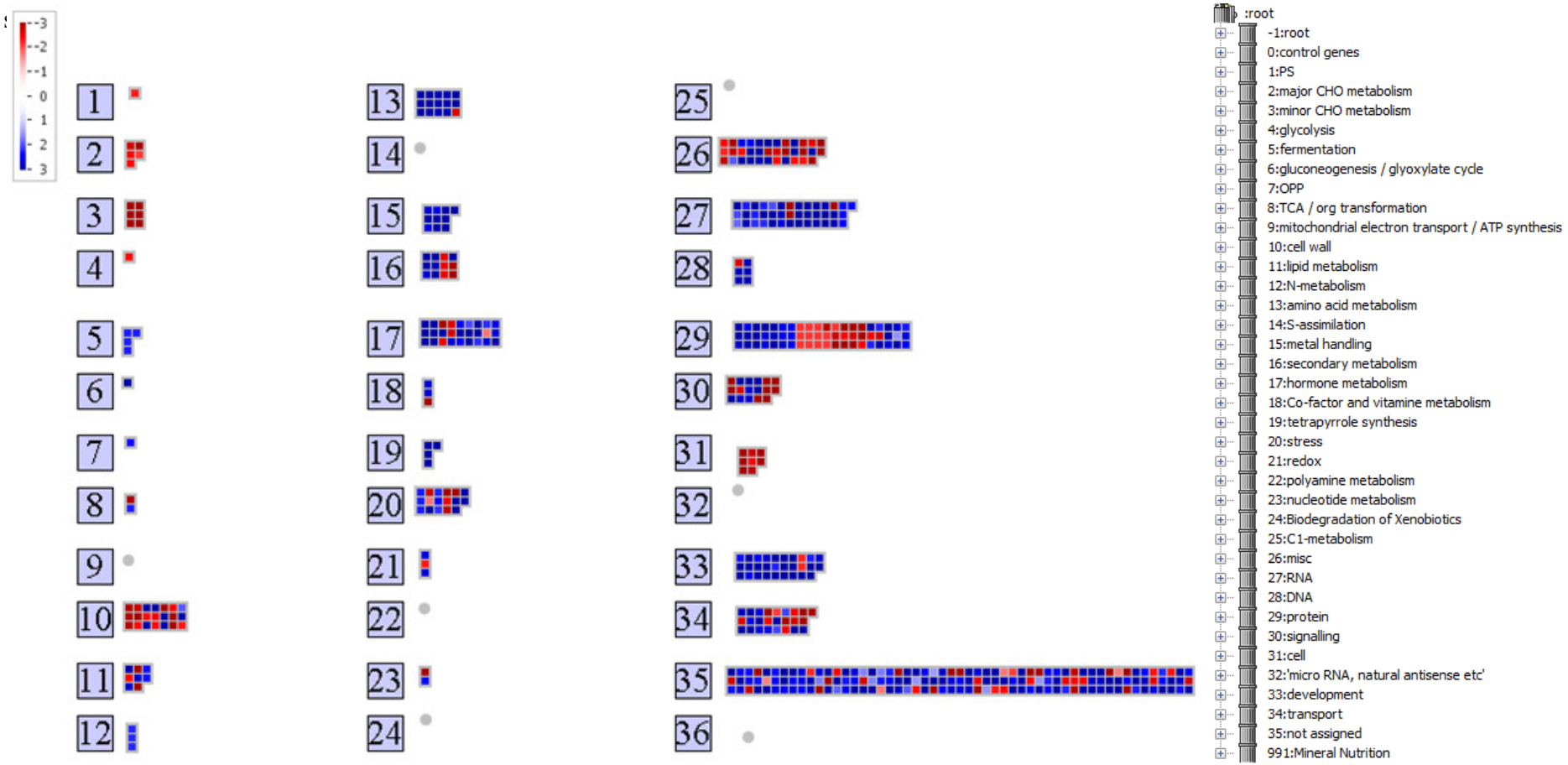


Figure 3.16 Bin maps generated by MapMan groups of differentially regulated genes between 4 and 14 weeks of nodule development. All genes shown had significant ($p \leq 0.05$) change in expression between the two respective time points. The degree of change is depicted based on the colour scale with red indicating down-regulation and blue indicating up-regulation.

3.2 Cystatin identification and classification

All expressed nodule cystatins were identified from the RNA-Seq data. When the oryzacystatin-I (PDB: 1EQK_A) was used for comparison as a model cystatin, 25 cystatin sequences were identified in the assembled genome; of these 20 were non-redundant sequences (Appendix C, Table A.5). When a phylogenetic analysis of cystatins was carried out by comparing different I25 cystatin subfamilies (Figure 3.17), Glyma13g04250 and Glyma20g08800, transcribed in nodules during nodule development and senescence, had high similarity to group A cystatins (*V. unguiculata* cystatin, OCI, HvCPI-1 and HvCPI-2) (Martinez, et al. 2009). Glyma13g04250 was further paralogous to Glyma14g04250 with identical location, but on a different chromosome. Also, the two cystatins Glyma13g25870 and Glyma15g36180 were highly similar to Cystatin B (At3g12490) and HvCPI-4 (group A) and Glyma05g28250 was further highly similar to group B cystatins (cystatin 2 (At2g31980), HvCPI-5 and HvCPI-9). They also contained a C-terminal extension with a SNSL amino acid motif enabling them to inhibit legumain C13 cysteine proteases (Martinez et al. 2007b). Finally, Glyma15g12211, which was the most abundant cystatin in nodules, was similar to group C (subgroup C1) cystatins (Monellin cystatin (At5g47550), HvCPI-6 and HvCPI-8).

All cystatin sequences were also analysed for the presence of signal peptides indicating their possible cellular localisation (Appendix D, Table A.7). Glyma05g28250, Glyma07g39590 and Glyma13g25870 might be localised in the secretory pathway, whereas Glyma13g04250, Glyma14g04250 and Glyma20g08800 are localised to any location, except the chloroplast, mitochondrion or secretory pathway. Localisation of Glyma15g36180 was not reliable and the cystatin could be located in either the mitochondrion or the secretory pathway.

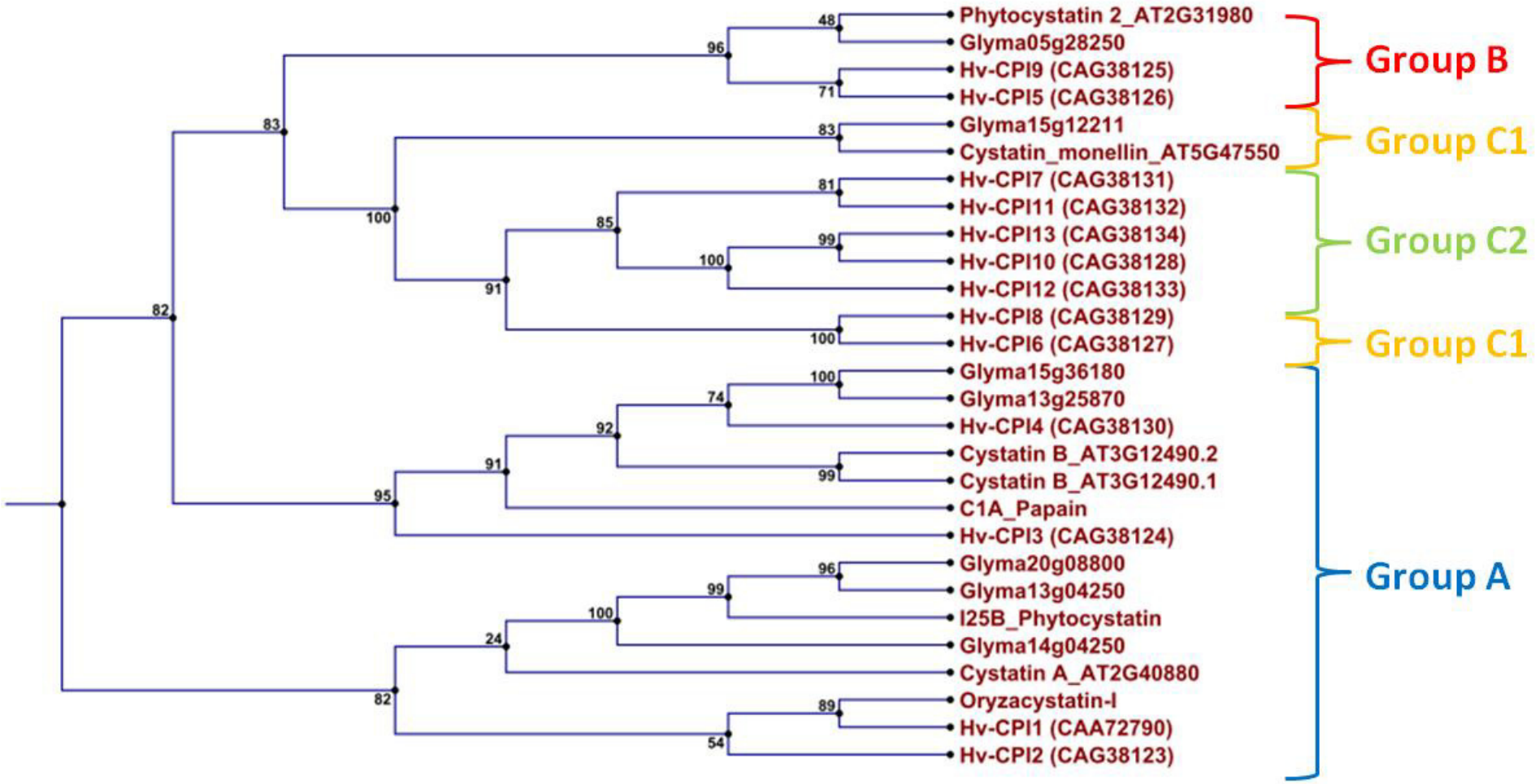


Figure 3.17 Mapping of transcribed soybean nodule cystatins to different I25 cystatin subfamilies. The phylogenetic trees were generated with the CLC Main Workbench v6.7.1 using the Neighbour Joining algorithm.

3.3 Cysteine protease identification and classification

A total of 99 cysteine protease sequences with homology (1E≤-1.0) to the model cysteine protease papain (E.C.3.4.22.2) were further identified from the soybean genome assembly (Appendix C, Table A.6). Several sequences were alleles, paralogous and orthologous of other cysteine protease gene sequences. From these, 79 non-redundant cysteine protease gene sequences were identified which had similarity to members of eight different cysteine protease sub-families (Richau et al. 2012). Seven sub-families were distinguished from the expression data and confidently five functional groups were identified (Figure 3.18). However, none of the identified soybean cysteine proteases clustered with papain (subfamily XCP1). Cysteine proteases with cathepsin-L-like activity included Glyma04g03090 (closely clustering with subfamily RD21), as well as the two proteases Glyma14g09440 and Glyma17g35720 (similar to subfamily RD21 members). The C-terminal granulin domain was confirmed, characteristic of the RD21 subfamily, in these cysteine proteases. Glyma04g04400 (cathepsin-L-like activity) had highest similarity to RDL2 (Arabidopsis gene At3g19400) and closely clustered with the RD21 subfamily members. Finally, Glyma04g36470 and Glyma06g18390 (cathepsin-L-like activity) were highly similar to members of the SAG12 subfamily despite absence of the additional C amino acid in the CGCCWAFS motif.

Seven proteases with cathepsin-F-like activity (Glyma04g03020, Glyma06g03050, Glyma10g35100, Glyma11g12130, Glyma12g04340, Glyma14g40670, and Glyma17g37400) were highly similar to subfamily RD19 members. However, the ERFNAQ motif (instead of the ERFNIN motif in the pro-domain) characteristic of the RD19 subfamily, was absent. Glyma08g12340, which had no significant similarity to any specific subfamily,

was closest to the two subfamilies RD19 or CTB3 with the C-terminal granulin domain. Further cysteine proteases with cathepsin-H-like activity included Glyma09g08100, Glyma15g19580 and Glyma17g05670, which had high similarity to AALP and ALP2. The three proteases also had an N-terminal NPIR vacuolar-targeting signal and other RD21 subfamily motifs (except that the ATC motif was lacking in Glyma09g08100). Although Glyma03g38520 and Glyma19g41120 had similarity to this subfamily, they contained an ECGIE motif in the C terminus, characteristic of subfamily CTB3.

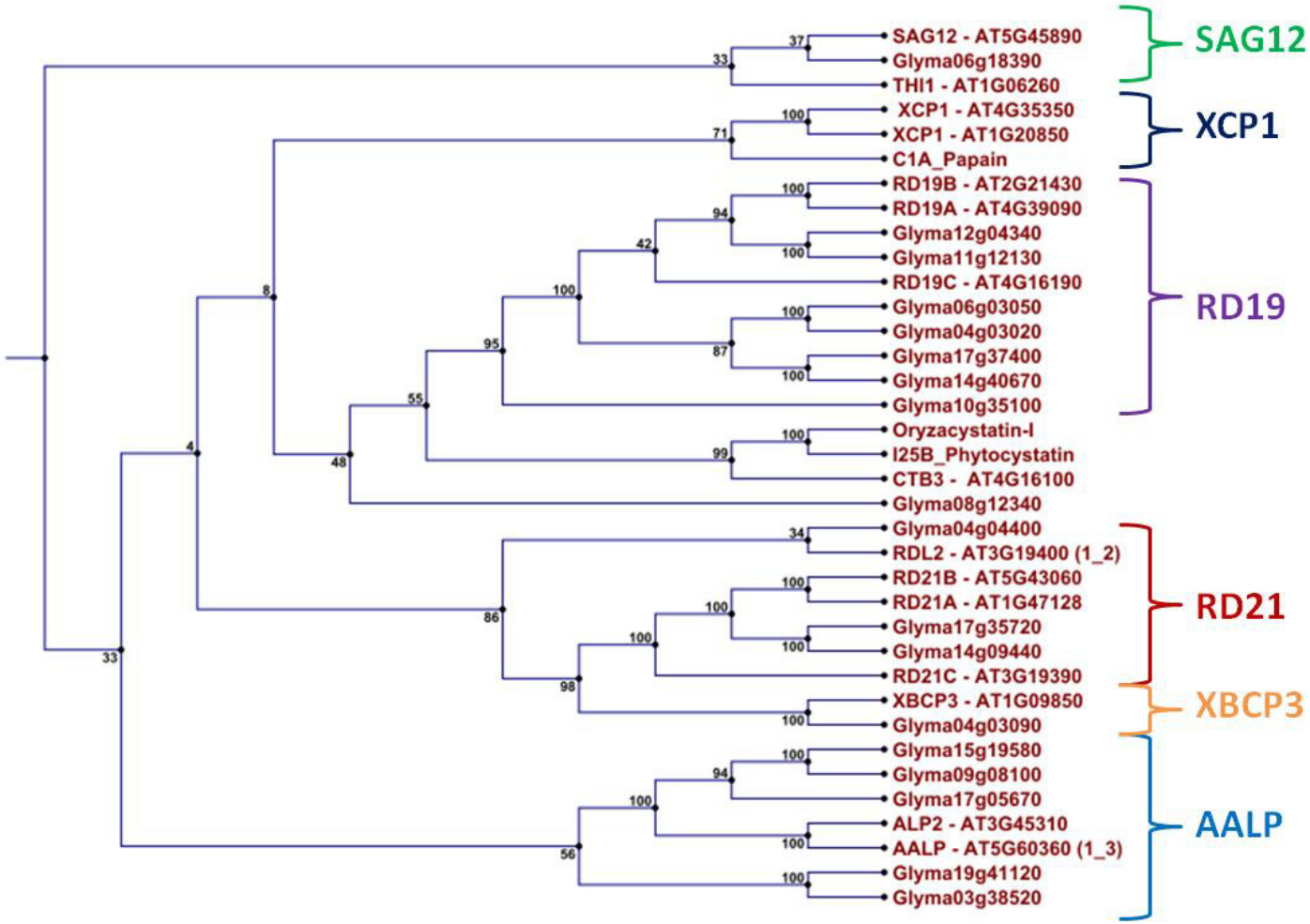


Figure 3.18 Mapping of transcribed cysteine proteases to sub-families and functional groups with similarity to the C1 cysteine protease papain. The phylogenetic tree was generated with the CLC Main Workbench v6.7.1 using the Neighbour Joining algorithm.

3.4 Cystatin transcription

The nodule cystatin and cysteine protease transcriptome was then investigated at various time points (4, 8 and 14 weeks) of soybean nodule development and senescence (Figure 3.19.A). The time point at 4 weeks represented initial nodule development, 8 weeks mature nodules actively fixing nitrogen, and 14 weeks senescing nodules. FPKM data were first compared with previous published data available online at SoySeq database (http://soybase.org/soyseq/) on the SoyBase website (Severin et al. 2010) where various tissue types were analysed 20-25 days after inoculation (comparable to the 4 weeks data). From a total of 20 putative soybean cystatins identified with the model I25B cystatin OC-I, only seven cystatins were transcriptionally active in nodules (Figure 3.19A). Glyma13g04250 and Glyma20g08800 had highest expression after 4 weeks but their expression decreased when nodules aged (Figure 3.19A). In contrast, transcription of Glyma05g28250, Glyma15g12211 (the most abundant cystatin) and Glyma15g36180 increased in the later stages of nodule development (Figure 3.19A), although none of these cystatins had statistically significant ($p \leq 0.05$) transcriptional changes. The two remaining cystatins, Glyma13g25870 and Glyma14g04250, did either not change (Glyma13g25870) or expression was below, or close to, the detectable threshold level (Glyma14g04250). RNA-Seq data were also validated by RT-qPCR where tested transcripts were selected on the basis of being representative for each investigated gene family. Determination of relative fold-expression of transcripts during development confirmed the RNA-Seq data indicating the fidelity of our RNA-Seq analysis approach (Figure 3.20).

3.5 Cysteine protease transcription

From the initial 99 putative cysteine protease sequences homologous to the model C1 cysteine protease papain, 18 cysteine proteases were transcriptionally active in nodules during at least one time point (Figure 3.19B). Glyma15g19580 (cathepsin-H like activity) was the most abundant cysteine protease in 4 weeks old nodules with Glyma17g37400 (cathepsin-F like activity) the most abundant at 14 weeks. Transcription of the majority of cysteine proteases increased with the onset of senescence, with five cysteine proteases (Glyma04g04400, Glyma08g12340, Glyma10g35100, Glyma11g12130 and Glyma17g05670) highly expressed in 4 and 8 weeks old nodules, but did not substantially increase with the onset of senescence. None of the cysteine protease transcription changed significantly ($p \geq 0.05$) except Glyma06g18390, with a very low relative abundance, which changed ($p \leq 0.05$) due to senescence (Figure 3.19B).

VPE protease (C13 cysteine proteases) transcription was also investigated (Figure 3.19C). These proteases resemble mammalian caspases. VPE transcription significantly increased during nodule senescence and transcription of four sequences (Glyma05g04230, Glyma14g10620, Glyma17g14680, Glyma17g34900) significantly ($p \leq 0.05$) increased (4.0 $\log_2$-fold change) for Glyma14g10620 and Glyma17g34900, with Glyma17g34900 having the largest increase in transcription or during senescence (Figure 3.19C). From the seven VPE gene sequences identified in the genome, only Glyma16g07190 was not transcribed during nodule development.

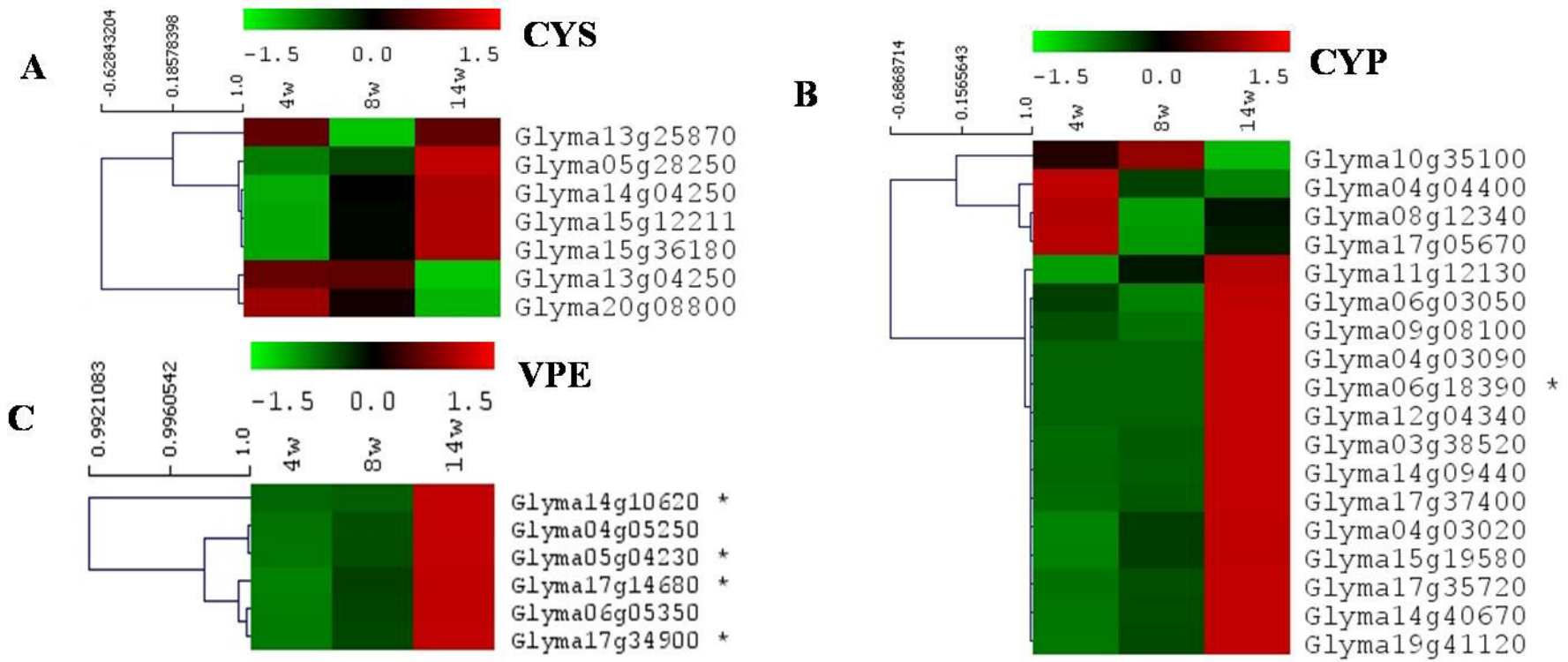


Figure 3.19 (A) Expression of cystatins (CYS) (B) cysteine proteases (CYP) and (C) vacuolar processing enzymes (VPE) in 4, 8 and 14 weeks old nodules expressed as FPKM (transcript abundances in fragments per kilobase of exon per million fragments mapped). Colour scale represents transcription for each time point normalized by subtracting the mean/median of three values from each individual value for each gene reduced by SD/RMS. Yellow indicates lowest value, orange middle and red highest FPKM value. *indicates significant change ($p \leq 0.05$) in transcription between individual time points. Multi-experiment viewer (MeV v4.8.1) software package was applied to graphically represent data (Saeed et al. 2003).

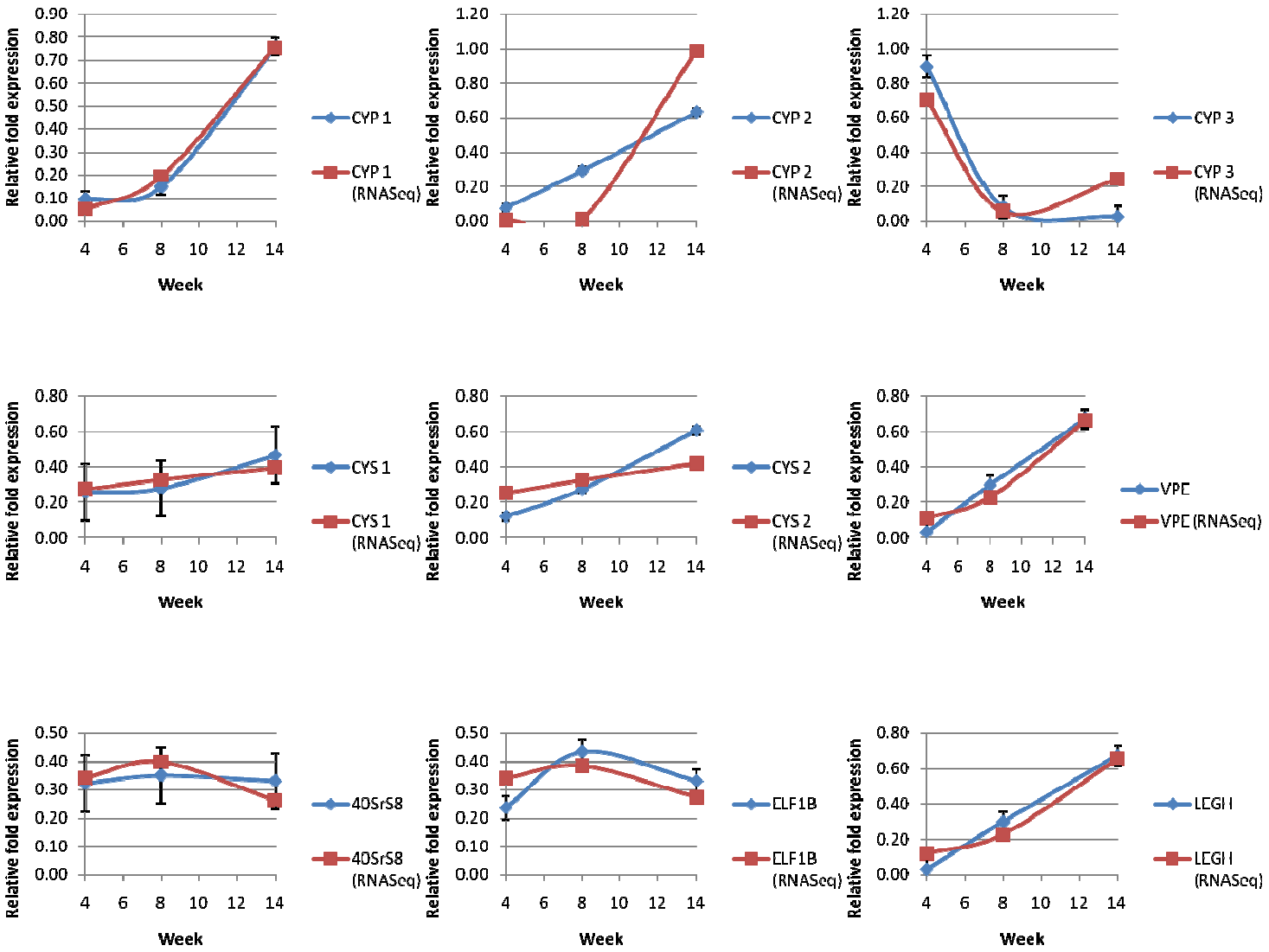


Figure 3.20 Relative expression measured by RT-qPCR of soybean cysteine proteases, cystatins, leghemoglobin and a VPE at each time point (4, 8 and 14 weeks) and corresponding FPKM abundance estimates derived from RNA-Seq mapping.

3.6 Cystatin inhibition strength and specificity

Cysteine protease activity measurements were carried out with nodule extracts to determine potency of transcribed cystatins. Fluorometric interaction assays were applied with commercially available cathepsin-L or cathepsin-B as well as isolated nodule protein extracts representing the total proteolytic complement active in nodules to establish a preferential binding profile for each cystatin. Cystatins transcribed in nodules had generally stronger affinity for cathepsin-L compared to cathepsin-B, with Glyma13g27980 and Glyma14g04250 equally effective in blocking both cathepsin-L and -B activities (Table3.5). Further, Glyma15g36180 inhibited cathepsin-L, but was unable to inhibit cathepsin-B, even when an inhibitor concentration of 1 mM was used. In contrast, cystatins not transcriptionally active in nodules showed higher inhibition rates of cathepsin-L, with Glyma18g12240 inhibiting both cathepsin-L and -B. Glyma14g04260's second domain and both domains of Glyma14g04291 were further unable to inhibit cathepsin-B, even at a concentration of 1 mM (Table 3.5).

Table 3.5 Inhibition (%) of protease activity by actively and non-actively transcribed cystatins during nodule life-span.

Cystatin	Cathepsin L-like activity (% inhibition)				Cathepsin B-like activity (% inhibition)			
	4 weeks	**8 weeks**	**14 weeks**	**p ≤ 0.05**	**4 weeks**	**8 weeks**	**14 weeks**	**p ≤ 0.05**
Positive control (E64)	50.3±1.1	26.4±5.0	31.9±4.5	*ac	37.2±2.3	NI	NI	*ac
OCI (1 µM)	47.4±1.3	28.2±2.3	22.7±7.3	*ac	44.9±3.8	NI	NI	*ac
Actively transcribed								
Glyma05g28250	36.1±0.5	31.5±0.9	30.6±0.4	ns	32.8±1.4	32.8±1.4	NI	*bc
Glyma13g04250	26.4±0.9	NI	29.7±1.8	*ab	27.6±2.3	27.6±2.3	24.9±3.2	*ab
Glyma13g27980	33.2±2.3	NI	NI	*ac	42.0±0.2	42.0±0.2	NI	*ac
Glyma14g04250	NI	NI	21.9±1.6	*bc	NI	NI	NI	ns
Glyma15g36180	49.9±5.3	28.4±3.1	NI	*abc	48.7±4.5	48.7±4.5	NI	*ac
Glyma20g08800	NI	NI	NI	ns	NI	NI	32.5±3.2	*ab
Non-actively transcribed								
Glyma04g10360	38.6±2.9	32.0±3.9	39.0±3.5	ns	35.3±5.5	30.9±5.5	28.6±5.8	ns
Glyma07g39590	47.5±3.2	39.1±9.5	51.3±5.1	*b	42.3±5.3	26.9±8.7	34.0±2.9	*a
Glyma08g11210	43.6±3.8	28.2±1.8	33.5±4.3	*abc	42.1±4.4	NI	NI	*ac
Glyma14g04260 (1^{st} domain)	58.9±1.1	37.8±4.9	36.2±3.3	*ac	46.4±1.8	NI	NI	*ac
Glyma14g04260 (2^{nd} domain)	36.6±4.9	NI	NI	*ac	39.8±5.6	NI	NI	*ac
Glyma14g04291 (1^{st} domain)	42.1±3.3	NI	NI	*ac	30.9±5.6	NI	NI	*ac
Glyma14g04291 (2^{nd} domain)	40.8±8.3	NI	NI	*ac	28.6±8.4	NI	NI	*ac
Glama18g12240	54.0±2.6	43.1±1.9	51.5±3.7	*a	36.6±5.8	28.3±3.9	22.4±7.4	*c

*significant change at $p \leq 0.05$; a (significant change from 4 to 8 weeks); b (significant change from 8 to 14 weeks); c (significant change from 4 to 14 weeks); ns (not significant); NI represents inhibition ≥ 20%. Blank values for Cathepsin L-like activity and Cathepsin B-like activity was 0.5 ± 0.7 FU/sec and 0.0 ± 0.3 FU/sec, respectively. The negative control values for Cathepsin L-like activity and Cathepsin B-like activity was 42.5 ± 1.6 FU/sec and 28.2 ± 0.8 FU/sec, respectively.

Cystatin potency was also measured against various nodule extracts (Table 3.5). The model rice cystatin OC-I was first applied, as well as the cysteine protease inhibitor E64. OC-I and E64 both prevented cathepsin-L-like activity in 4 weeks old nodules but were less efficient against extracts derived from 8 and 14 weeks old nodules (Table 3.5). Both inhibitors also prevented cathepsin-B-like activity in an extract of 4-week old nodules. OC-I and E64 potency were then compared with the potency of various recombinant soybean cystatins either actively transcribed or non-active in nodules (Table 3.5). Cystatins tested were generally more active against extracts from younger nodules (Table 3.5). Five of the cystatins actively transcribed in nodules blocked cysteine protease activity in nodule extracts. However, only Glyma05g2850 inhibited cathepsin-L-like activity in nodule extracts from all three time points (4, 8, and 14 weeks) and cathepsin-B-like activity in extracts derived from 4 and 8-week old nodules. The most potent cystatin among the expressed cystatins was Glyma15g36180. Potency of this cystatin was comparable to OC-I and E64 when either cathepsin–L or B activity was measured in an extract derived from 4-week old nodules.

When cystatins not actively transcribed during nodule development were tested, they were generally more active against nodule extracts than cystatins actively transcribed in nodules (Table 3.6). All non-transcribed cystatins had potency comparable to OC-I and E64 when tested against an extract derived from 4-week old nodules. Among them, Glyma14g04260 domain 1 and Glyma18g12240 had highest inhibition of all tested cystatins with 58.9 % and 54 % inhibition, respectively. Three cystatin (Glyma04g10360, Glyma07g39590 and Glyma18g12240) inhibited cathepsin-L, as well as cathepsin-B like activity, in extracts derived from all three time points. Glyma07g39590 was further the most potent of all tested cystatin against cathepsin-L and -B activity in an extract derived from 14-week old senescent nodules.

Table 3.6 Expression and inhibitory potency of cystatins against proteases from different aged nodules.

Cystatin	Expression	Cat-L inhibition	Cat-B inhibition
Active			
4 weeks			
Glyma05g28250	+ (22.65)	+ (36.1%)	+ (32.8%)
Glyma13g04250	+ (97.58)	+ (26.4%)	+ (27.6%)
Glyma14g04250	(-) (2.14)	(-)	(-)
Glyma15g36180	+ (26.34)	++ (49.9%)	++ (48.7%)
Glyma20g08800	+ (85.83)	(-)	(-)
14 weeks			
Glyma05g28250	+ (39.78)	+ (30.6%)	(-)
Glyma13g04250	+ (63.86)	+ (29.7%)	(+) (24.9%)
Glyma14g04250	(+) (12.38)	(+) (21.9%)	(-)
Glyma15g36180	+ (55.64)	(-)	(-)
Glyma20g08800	+ (56.25)	(-)	(-)
Non-active			
4 weeks			
Glyma04g10360	(-) (0)	+ (38.6%)	+ (35.3%)
Glyma07g39590	(-) (2.09)	+ (47.5%)	+ (42.3%)
Glyma08g11210	(-) (0)	+ (43.6%)	+ (42.1%)
Glyma18g12240	(-) (0.28)	++ (54.0%)	+ (36.6%)
Glyma13g27980	(-) (0)	+ (33.2%)	+ (42.0%)
14 weeks			
Glyma04g10360	(-) (0)	+ (39.0%)	+ (28.6%)
Glyma07g39590	(-) (1.23)	++ (51.3%)	+ (34.0%)
Glyma08g11210	(-) (0)	+ (33.5%)	(-)
Glyma18g12240	(-) (0.58)	++ (51.5%)	(+) (22.4%)
Glyma13g27980	(-) (0)	(-)	(-)

++ strong, + medium. (+) low and (–) no cystatin expression/or activity (tested up to 1mM). Expression indicated as measured FPKM abundances and activity indicated as % inhibition.

3.7 Cystatin, cysteine protease localisation and nodule histology

The relationship between the gene expression and biochemical data was confounding. I speculated that different transcripts were derived from different layers of the nodules, due to a macro-dissection approach of the entire nodule, as appose to micro-dissection approach of specific regions of the nodule tissue. The same problem would apply to the biochemical assays, therefore, the localisation of the cystatins and cysteine proteases expressed at the different time points of nodule development were investigated by a histological and immune-histochemical approach. Microtomic slices of nodules from each respective time point were prepared and mounted on microscope slides and used in both histological staining as well as immuno-histochemical detection. The histochemical stains applied included crystal violet, safranin-O and fast green, which were used either individually or in conjunction with each other as counterstains. These were applied to firstly obtain optimal resolution of cellular structures of the nodules, at each respective time point, prior to detection of target proteins by immunohistochemistry. Each stain is specific to different cellular structures, allowing nodule tissue types and physiological anatomy to be differentiated. Crystal violet stains cells violet and is typically applied in determining chromatin and nucleoli in plant tissue (Figure 3.21). Safranin-O stains the nuclei, chromosomes, lignified and cutinized cell walls red (Figure 3.22). The fast green stains the cytoplasm and cellulose cell walls in plant cells green (Figure 3.23). The different nodule structures that can be distinguished include: the outer cortex (OC) from figure A for Figures 3.21 – 3.25, late symbiotic zone (LS) and early symbiotic zone (ES), the vascular tissue is indicated by the balled arrows and vascular bundles indicated by arrows, the nodule parenchyma is indicated by an asterisk (*) and intercellular air spaces can be seen at the junctions of cells at the symbiotic zones of almost all figure sets B – D for Figures 3.21 – 3.25. The cell nuclei can also be seen in figures B – D for figures 3.21 – 3.25.

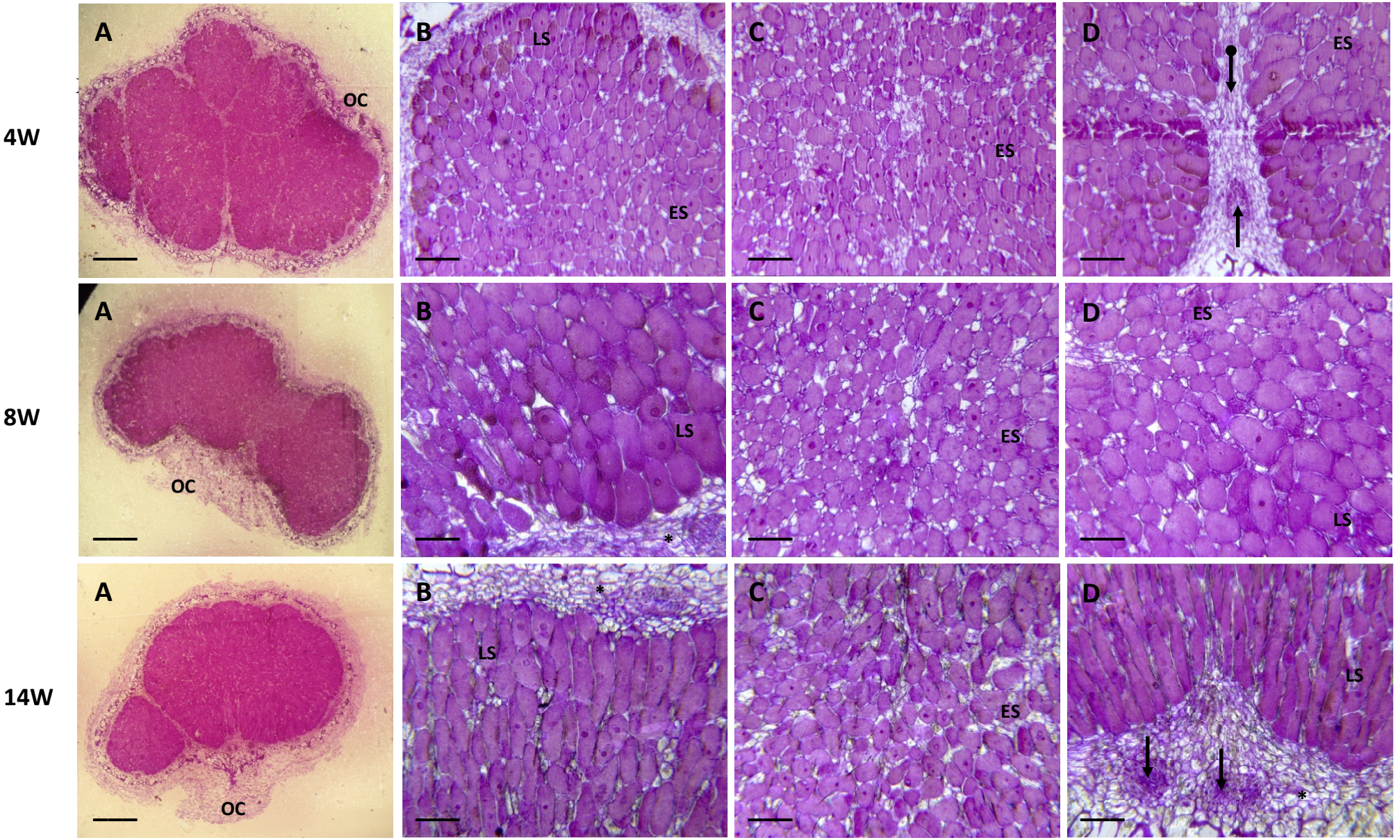


Figure 3.21 Nodule cross-sections stained with crystal violet, A. Overview picture taken with dissection microscope, bar represents 100 µm, B – D. Pictures taken with bright-field microscope from different regions of nodule tissue, bar represents 20 µm.

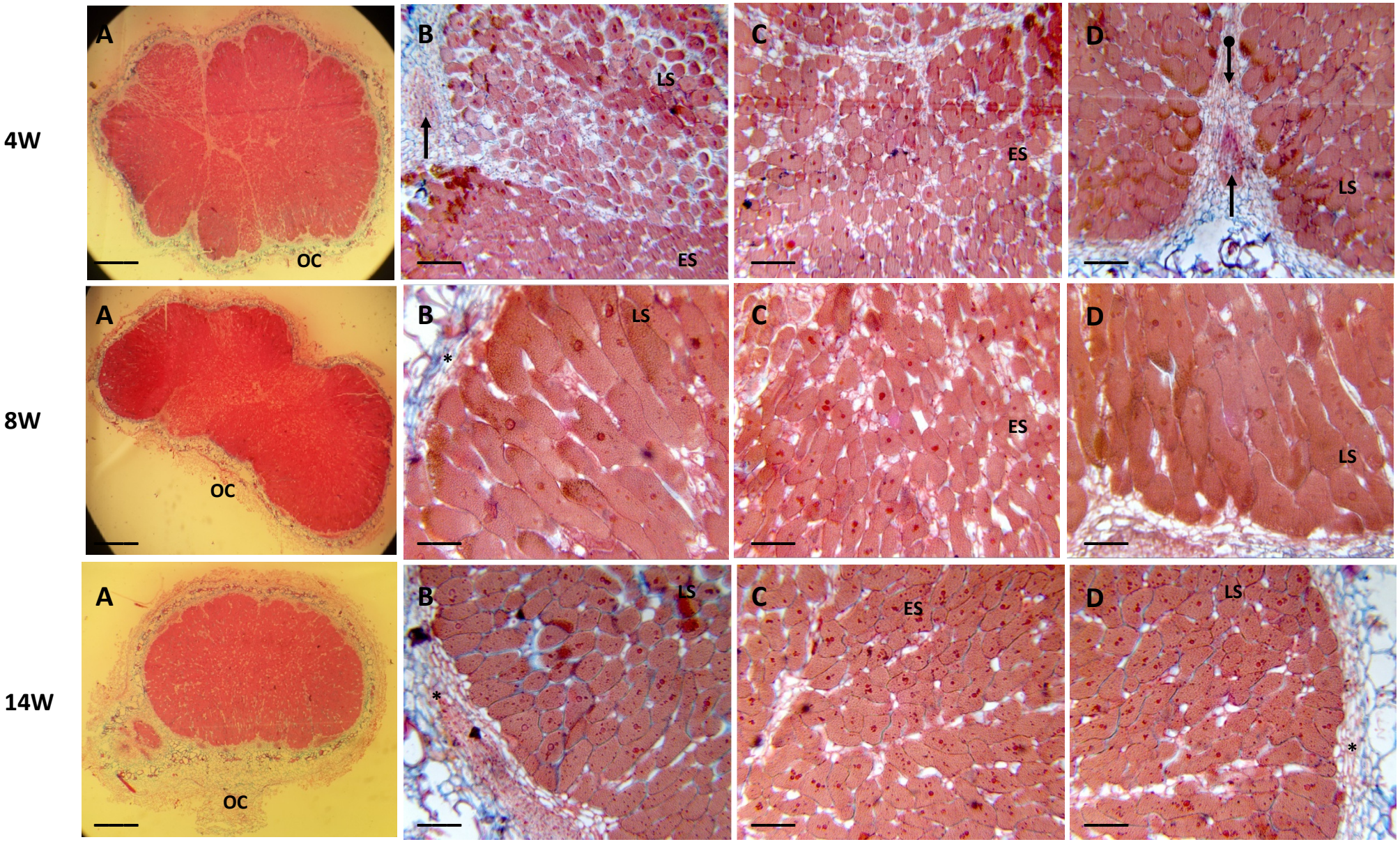


Figure 3.22 Nodule cross-sections stained with safranin-O, A. Overview picture taken with dissection microscope, bar represents 100 µm, B – D. Pictures taken with bright-field microscope from different regions of nodule tissue, bar represents 20 µm.

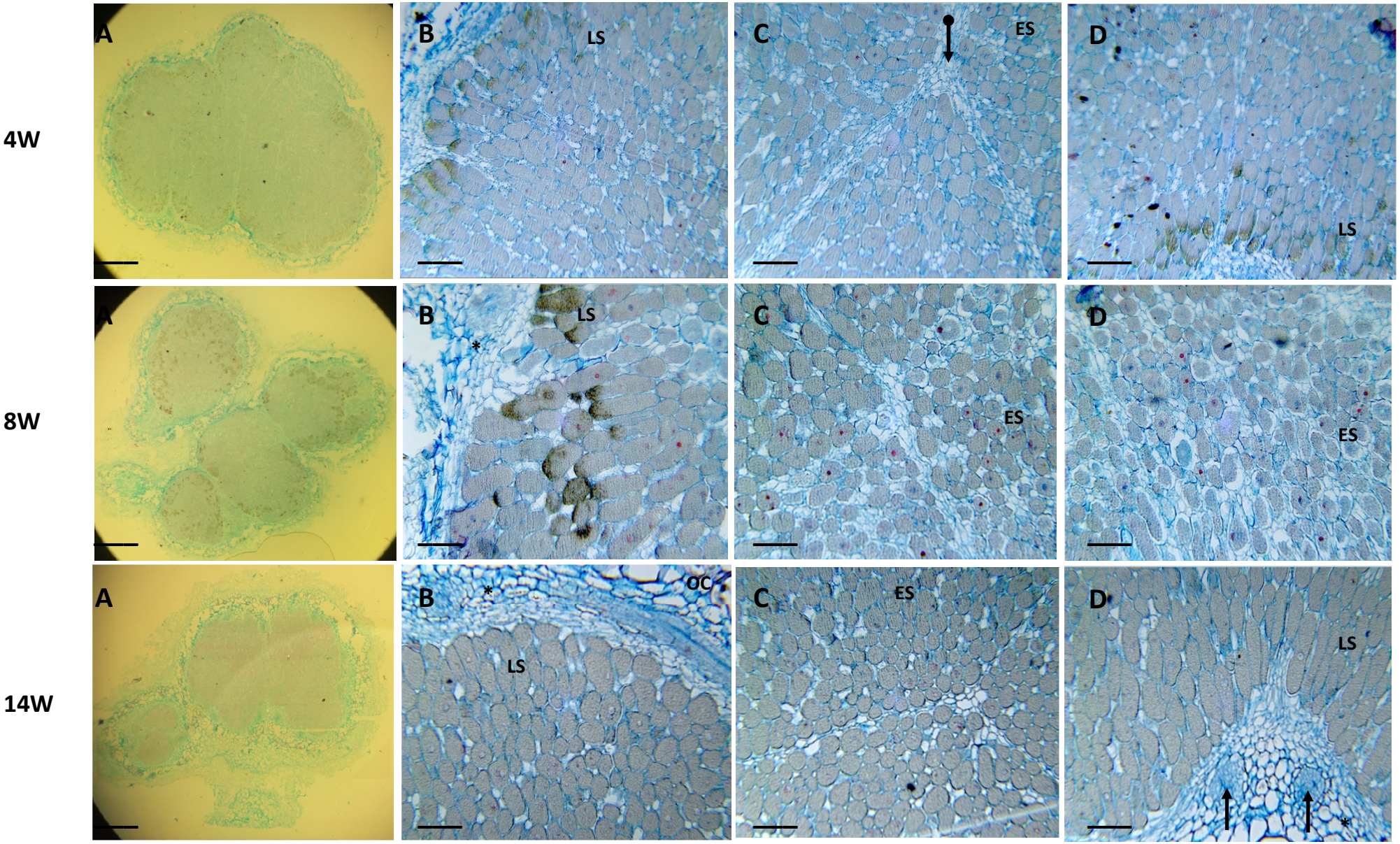


Figure 3.23 Nodule cross-sections stained with fast green, A. Overview picture taken with dissection microscope, bar represents 100 µm, B – D. Pictures taken with bright-field microscope from different regions of nodule tissue, bar represents 20 µm.

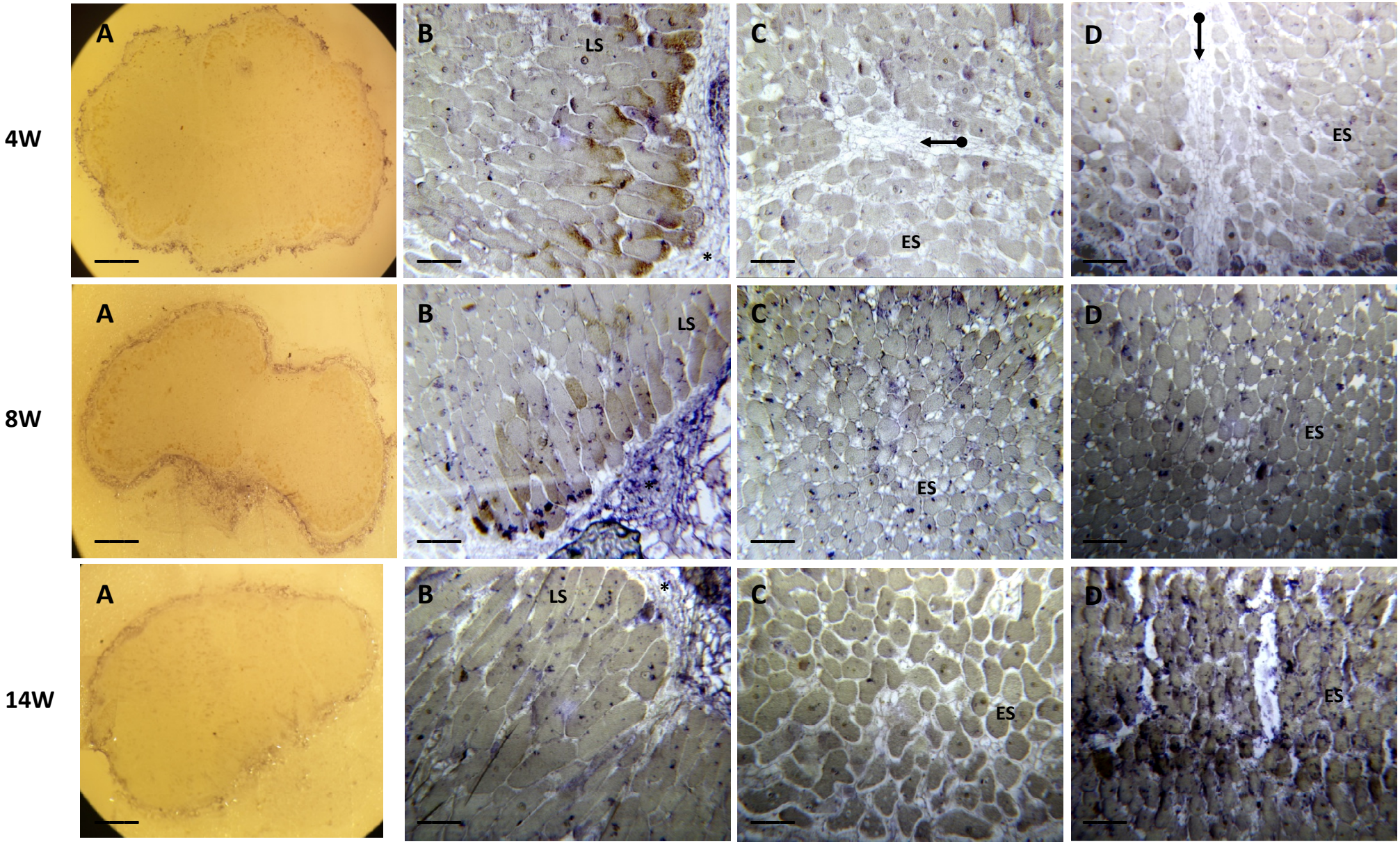


Figure 3.24 Nodule cross-sections after antibody detection of cysteine proteases using the R79 antibody and staining with alkaline phosphatase, A. Overview picture taken with dissection microscope, bar represents 100 µm, B – D. Pictures taken with bright-field microscope from different regions of nodule tissue, bar represents 20 µm.

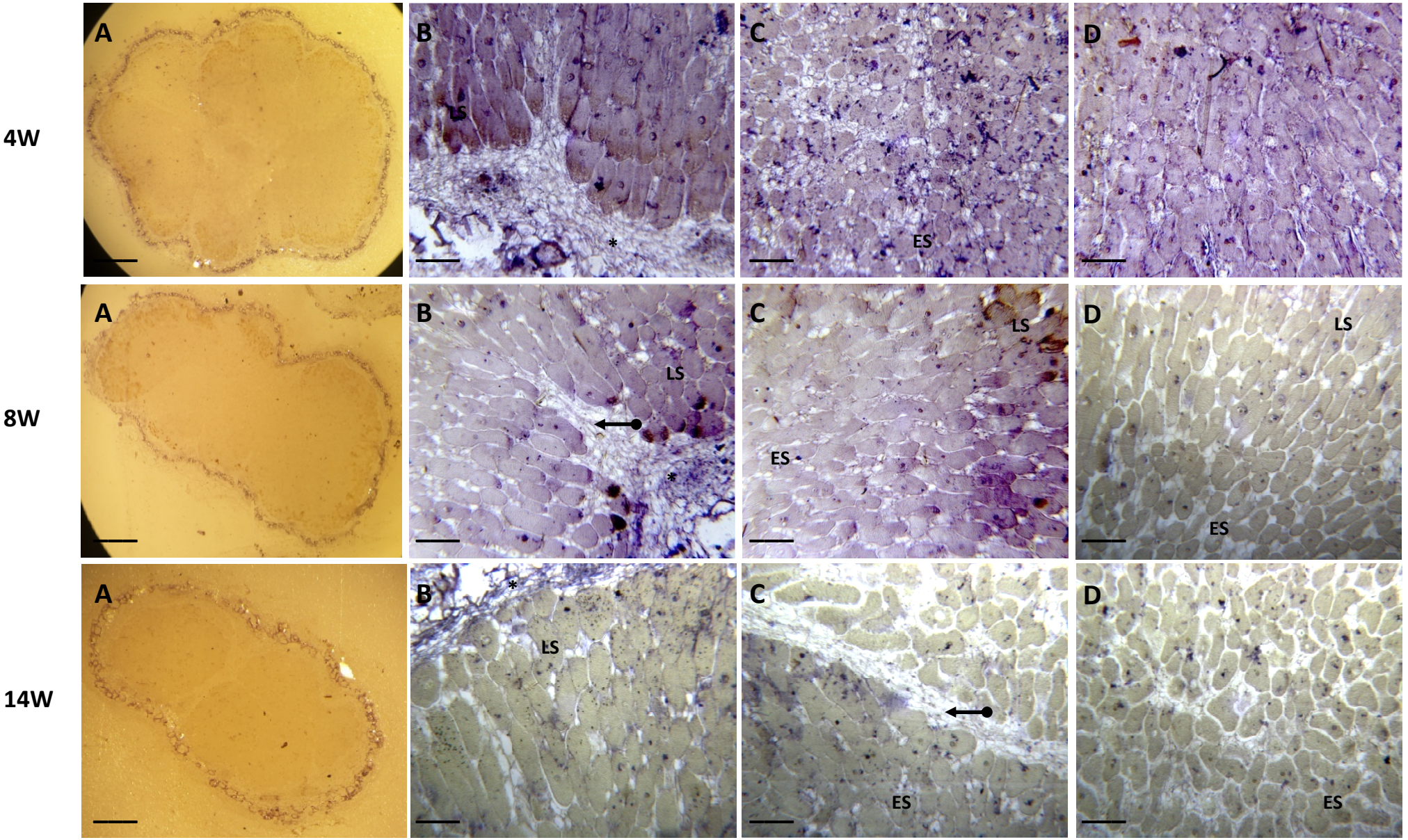


Figure 3.25 Nodule cross-sections after antibody detection of cystatins using the OCI antibody and staining with alkaline phosphatase, A. Overview picture taken with dissection microscope, bar represents 100 µm, B – D. Pictures taken with bright-field microscope from different regions of nodule tissue, bar represents 20 µm.

Soybean nodules are determinate in nature, meaning that their meristematic activity is lost after formation. However, nodule maturation is largely based on cellular expansion leading to nodule growth (Vorster et al. 2013). This can be seen from comparing the late symbiotic zone from each of the different time points, 4 weeks compared to 8 and 14 weeks, in Figures 3.21 – 3.25.

For the antibody detection of cysteine proteases and cystatins (Figures 3.24 and 3.25) it was expected that the cysteine proteases and cystatins active in the nodules would be differently localised at each of the respective time-points. The study by Kardailsky and Brewin (1996) found transcripts of the cysteine protease in pea nodules (PsCYP1) to be localised to the infected region of the apical region associated of the meristematic activity. The subsequent study by Vincent and Brewin (2000) further investigated the subcellular localization by antibody probes and found protein accumulation in the vacuoles and vesicles of the endomembrane system. This same cysteine protease (Figure 3.24) antibody was applied, as an antiserum, raised against a recombinant PsCYP15A protein in New Zealand white rabbit (Vincent and Brewin 2000). To date, no study has investigated the localisation of cystatins in the nodules of legumes. The cystatin (Figure 3.25) antibody used was the antiserum raised against a recombinant oryzacystatin-I (OCI) protein in rabbit (Van der Vyver et al. 2003). No positive detection was made with the antibodies that were available. However, artefacts from the colorimetric detection are visible.

CHAPTER 4

DISCUSSION

This study set about investigating the processes of nodule development and nodule senescence in a globally important crop plant, soybean. During the study a large amount of transcriptomic expression data was generated from different nodule developmental stages. Although the members of the cysteine protease and cysteine protease inhibitor (cystatins) gene families were particularly investigated in this study, other gene(s) and gene families could be of equal or greater importance in the processes of nodule development and nodule senescence, however subsequent studies are required.

4.1 Nodule senescence

Establishment of the RNA-Seq database for nodule development also provided the benefit to identify, besides cysteine proteases and cystatins, expression of other nodule genes. This included, for example, genes specifically associated with the onset of senescence. In general, a simple comparison of legume species with non-legume species, such as Arabidopsis, is not reliable for this type of functional studies when complex symbiotic nitrogen fixation, unique for nodules, is involved (Benedito et al. 2008). Benedito et al. (2008) previously reported a massive difference in transcript abundance of genes when developing nodules and other organs were compared in particular with regard to regulatory activity at transcriptional and/or post-translational level. Legume-specific genes included over 300 cysteine cluster proteins (CCPs), 63 proline-rich proteins (PRPs) and 21 glycine-rich proteins, which were

exclusively, or preferentially, expressed during nodule development having nodule specific functions. Those genes were, however, absent from genomes of several non-legume species.

In this PhD study, selection and identification of genes that significantly changed during nodule development were based on expression changes of at least 2-fold during nodule development. This included, for example, pathogenesis (*PR*)-related defence genes as well as genes involved in glycolysis, carbon fixation, nitrogen metabolism and SNF. This partly confirmed already published data (El Yahyaoui et al. 2004). Expression of genes involved in the TCA cycle as well as certain genes with functions in the anti-oxidative system significantly changed during nodule development and specifically at onset of nodule senescence. Very likely due to *Rhizobium* infection, *PR* genes were generally up-regulated in earlier stages of development, despite that during very early plant-*Rhizobium* interaction such *PR* genes have been found to be repressed to allow *Rhizobium* infection (Maunoury et al. 2010). *PR* genes significantly changing in expression during development, with expression decreasing during senescence, include, for example, the thaumatin-like osmotin 34 protein (Glyma05g38130), the glycosyl hydrolase family 81 protein (Glyma08g26935), which is part of the beta-glucan elicitor receptor, the acidic endochitinase (Glyma12g25740) and also the MLO protein, which not only limits the cell-wall restricted oxidative burst but also regulates salicylic acid and *PR* genes (Piffanelli et al. 2002, Colebatch et al. 2004). Only a jasmonic acid carboxyl methyltransferase (Glyma09g38930), involved in the hormonal control of nodulation (Costanzo, et al. 2012), was active in senescent nodules and, interestingly, also a trypsin inhibitor (Glyma08g45520). This inhibitor was significantly (4 $\log_2$-fold) up-regulated in senescent nodules. Lievens et al. (2004) previously reported that a Kunitz-type trypsin inhibitor was expressed in the nodule primordium but during early stages of nodulation in *Sesbania rostrate* with maximal expression after two days with inhibitor

transcripts also found in the central nodule tissue, which is part of the fixation zone. However, contrary to the result in the PhD study with Glyma08g45520, expression of the Kunitz-type trypsin inhibitor decreased 12 days after nodulation.

Genes involved in nitrogen metabolism and SNF changing significantly during development include, for example, the glutamine synthase (Glyma09g30370). This synthase contributes to ammonia metabolism. The gene was greatly up-regulated at onset of senescence, while gene homologs were down-regulated. It is possible that the function of this single transcript might still be required in the assimilation process of residual nitrogen fixed during the onset of senescence. However, down-regulation might be expected when the symbiotic interaction is terminated during nodule senescence. The expression of these transcripts should be further investigated as an interesting avenue of investigation for further studies. Other genes down-regulated due to senescence are genes coding for leghemoglobin, required in SNF (Downie 2005), and down-regulation was confirmed in this PhD study. Bacterial nitrogenase requires low levels of oxygen and high levels of ATP produced from bacterial respiration. Leghemoglobin regulates the transfer and presence of oxygen in the nodule (Appleby 1984). Five leghemoglobin genes were active in soybean nodules and, as expected, expression of the majority of genes (Glyma10g34260, Glyma10g34280, Glyma10g34290, and Glyma20g33290) was strongly down-regulated during senescence. Down-regulated leghemoglobins consisted of symbiotic leghemoglobins, contributing to regulating oxygen levels in the nodules. However, expression of one leghemoglobin gene (Glyma11g12980) increased during senescence. This single up-regulated leghemoglobin was a non-symbiotic leghemoglobin. Such non-symbiotic leghemoglobins possibly detoxify nitric oxide, which influences the induction of plant defence responses (Downie 2005, Ott et al. 2005).

Genes involved in glycolysis, carbon fixation, and also the TCA cycle were, according to the RNA-Seq analysis, expressed when the nodule matured, but were generally down-regulated during senescence when the symbiotic plant-*Rhizobium* relationship terminates. Lodwig and Poole (2003) previously reported that the plant supplies the bacteroid with dicarboxylic acids, such as malate and succinate, from the sucrose breakdown. In this PhD study certain exceptions were, however, found. Examples were the gene for PEPC (Glyma9g33650), highly up-regulated in mature nodules, and the genes coding for the glyoxylate cycle key enzymes malate synthase (Glyma17g13730) and iso-citrate lysase (Glyma06g45950 and Glyma12g10780), which all have a role in nodule metabolism. Expression of malate synthase increased during senescence and the two iso-citrate lysases were also highly expressed during senescence. Phosphoenolpyruvate carboxylase (PEPC, EC 4.1.1.31) plays a paramount role in nodule metabolism with PEPC providing substantial carbon for N_2-fixation and N assimilation (Lodwig and Poole 2003). Work carried out by Johnson, et al. (1966) already demonstrated that malate synthetase is active in extracts of rhizobia from soybean nodules, but any iso-citrate lyase activity was lacking, very likely requiring plant nodule iso-citrate lyase. However, why these genes are either increasing expression or are highly expressed during senescence when symbiosis breaks down requires further investigation. Genes with changed expression during nodule development also included sucrose invertase (Glyma04g00750) beta-amylase (Glyma06g45700) and the glucose-6-phosphate transmembrane transporter protein (Glyma15g11270). The majority of sucrose synthase homologs were strongly down-regulated during senescence possibly due to termination of the plant-*Rhizobium* interaction. Significant increases in expression of genes associated with oxidative stress and antioxidant capacity were also not found by RNA-Seq analysis. For example, the glutathione synthase 2 (Glyma19g42610) gene, expressed during nodule development, was down-regulated during senescence (>2-fold). Also, the dehydroascorbate

reductase 2 (Glyma20g38440), highly expressed at 4 weeks, was steadily down-regulated afterwards. This dehydroascorbate reductase gene subsequently declined in expression with the onset of senescence, possibly indicating involvement of such enzymes in nodule anti-oxidative system to ultimately provide the correct environment for PCD to occur.

4.2 Cysteine proteases and cystatins

The Phytozome database (www.phytozome.net) currently contains over 300 cystatin-like sequences from the Viridiplantae kingdom, 706 C1 cysteine protease sequences and 362 C13 cysteine protease (VPE-type) sequences. The recent release of the complete soybean genome (Schmutz et al. 2010) as well as the release of a RNA-Seq atlas of genes expressed in 14 different soybean tissues including nodules (Severin et al. 2010) now allows accurate studies to determine the cystatin and cysteine protease classes expressed in nodules. This also allows investigating if endogenous cystatins preferentially interact with specific target cysteine proteases in nodules. Due to this recently established knowledge, this PhD study was aimed to provide a particular insight into the cysteine protease- cystatin interaction by first identifying and characterizing expression of members of the cystatin and cysteine protease gene families in soybean nodules and secondly studying their interaction and determining their possible localization in the nodule cells. In this study 20 non-redundant cystatins and 79 non-redundant cysteine proteases were identified through homology searches in the soybean genomic database consisting of both actively and non-actively transcribed gene sequences during nodule development. Nodule gene transcription profiles established with the RNA-Seq technique (McIntyre et al. 2011) allowed the expression of all oryzacystatin I (OCI)-like cystatins, papain-like cysteine proteases, as well as vacuole VPE-type cysteine proteases, to be determined. This was done for selected time points throughout the lifespan of determinate

soybean crown nodules, from nodule development to senescence. A changing expression profile for each of these gene families was also confirmed with the onset of senescence supporting the original set working hypothesis for this PhD study of a balanced interplay between individual cysteine proteases and cystatins at work during nodule development and senescence.

Eight cystatin genes were identified to be actively expressed in soybean nodules using RNA-Seq. Since a macro-dissection of crown nodule tissue was carried out, and not a micro-dissection with selected tissues, RNA-Seq data represented gene transcription profiles of the entire nodule. This also contained transcripts from areas surrounding the senescing nodule cortex. When gene transcription in nodules was compared with already published RNA-Seq data from various other tissue types (Severin et al. 2010), none of the identified nodule cystatins was uniquely transcribed at onset of senescence or during senescence. Further, only Glyma05g28250 was actively transcribed and inhibited cathepsin-L-like activity in nodule extracts at all three selected time points. This particular cystatin very likely plays a maintenance role and regulates cysteine protease activity throughout nodule development and senescence. The other actively transcribed cystatins were only capable of inhibiting specific types of cysteine proteases with cathepsin-L or -B activity at specific time points. Cathepsin-B is a member of the peptidase C1 family and this cysteine protease is required for PCD involved in hypersensitive response providing plant disease resistance (Gilroy et al. 2007). Transcription of cystatins Glyma05g28250, Glyma15g12211, Glyma15g36180 increased significantly during onset of senescence with concurrent co-induction of several cysteine proteases. These three cystatins possibly regulate proteolysis when nodules senesce and undergo PCD and are therefore involved in maintaining nitrogen fixation, due to limiting protease activity, in determinate soybean nodules during senescence.

In this PhD study, there was also an interest to determine in which families and functional groups nodule cystatins and cysteine proteases belong as well as the cystatin substrate preference by testing recombinant expressed and purified cystatins *in vitro* with various cysteine protease-containing extracts. Cystatins are part of subfamily B of the I25 cystatin family and in cereals they can be divided into various functional groups (A, B and C) with most cystatins belonging to groups A and C (Martinez et al. 2009). Group A cystatins, which efficiently inhibit cathepsin L-like cysteine-proteases, are preferentially expressed in dry and germinating seeds. Group C1 cystatins, which are potent inhibitors of C1A peptidases, are mostly expressed in developing seed endosperms. However, any classification solely based on phylogenetic analysis might not accurately provide information about a particular function (Martinez et al. 2009). Several nodule cystatins, almost equally transcribed during nodule development and senescence, had high similarity to group A cystatins which regulate endogenous enzymes involved in the mobilization of stored proteins upon germination (Arai et al. 2002, Kiyosaki et al. 2007, Martinez et al. 2009).

The nodule group A cystatin cluster also contained two cystatins, Glyma13g25870 and Glyma15g36180, with a C-terminal extension. Such an extension was also found in Glyma05g28250, highly similar to group B cystatins. These cystatins with a carboxy-terminal extension contain a SNSL amino acid motif and inhibit cysteine proteases of the legumain C13 family (VPEs) (Martinez et al. 2007b). The consistent transcription of these proteins during nodule development and increase of transcription was found for Glyma15g36180 and Glyma05g28250 when nodules senesce. This indicates that they are very likely produced to tightly control cell disruption and activation of any cysteine proteases that may compromise nitrogen fixation. VPE proteases resemble mammalian caspases and they contribute to the senescence process and PCD by contributing to the collapse of the vacuole membrane with

release of proteases into the cell (Hara-Nishimura et al. 2005). There is also evidence that VPEs play a regulatory role activating pre-proteases by post-translational modification, leading to maturation and proteolytic activity upon removal on the I19 inhibitory domain (Roberts et al. 2012). Cysteine proteases, expressed as pre-proteins, consist of an I29 inhibitor domain preventing non-specific activity (Wiederanders 2003). In this PhD study, transcription of the entire set of nodule VPE cysteine proteases strongly increased coinciding with the progression of senescence. VPEs are, therefore, predominantly transcribed in senescent nodules and might play an important role in the activation of cysteine proteases. These activated cysteine proteases finally degrade both the bacteroids and nodule cells (Vincent et al. 1980, Malik et al. 1981, Pladys and Vance 1993, Kardailsky and Brewin 1996, Vincent and Brewin 2000) and also correlates with decrease in both nitrogenase activity (Pfeiffer et al. 1983) and crown nodule biomass as well as decrease in nodule number (Vorster et al. 2013).

Glyma15g12211 was the most abundant nodule cystatin with similarity to group C1 cystatins. It was almost 4-times higher transcribed than all other actively transcribed nodule cystatins. Glyma15g12211 was further identical to the previously described Glyma15g12210 (Severin et al. 2010), which is highly transcribed in both nodules and other tissues. In cereals, group C1 cystatins are preferentially expressed in seeds, particularly in developing endosperms, and these cystatins are potent inhibitors of C1A peptidases (Martinez et al. 2009). Future research is, however, needed to identify the specific target cysteine proteases and it is preferentially expressed in nodules. In addition to actively transcribed cystatins, non-actively transcribed cystatins were also identified in this PhD study in nodules. When expressed *in vitro*, these non-actively transcribed cystatins had a much broader range of inhibitory activity against the nodule proteolytic complement than actively transcribed cystatins. They had nearly double

the inhibitory capacity than actively transcribed cystatins against cysteine protease activity with cathepsin-L-like activity and were partially also more active against cysteine proteases with cathepsin-B-like activity. These non-actively transcribed cystatins might therefore have other specific functions and are only activated under certain biotic and abiotic stress conditions to prevent premature nodule death, which might be investigated in future studies.

Besides cystatins, 18 cysteine proteases were also actively transcribed during nodule development. The identified cysteine proteases were functionally diverse belonging to different proteolytic sub-families. In general, cysteine proteases cluster into different subfamilies (Richau et al. 2012) based on activity and substrate specificity, and cysteine proteases closest to papain clustered with subfamily XCP1 represented by *Arabidopsis thaliana* genes At1g20850 and At4g35350. Cysteine proteases with cathepsin-L-like activity further closely cluster with subfamily RD21 consisting of RD21A (*A. thaliana* gene At1g47128), RD21B (At5g43060) and RD21C (At3g19390). A C-terminal granulin domain is characteristic of the RD21 subfamily (Richau et al. 2012). Cysteine proteases with cathepsin-L-like activity can further cluster with the SAG12 subfamily (Richau et al. 2012) and cysteine proteases with cathepsin-F-like activity cluster with subfamily RD19. Members of this subfamily RD19A (At4g39090), RD19B (At2g21430) and RD19C (At4g16190) and RD19 have a characteristic ERFNAQ motif in the pro-domain (Richau et al. 2012). Further, cysteine proteases with cathespsin-H-like activity clustered with members of the AALP (At5g60360) and ALP2 (At3g45310) subfamilies. In this PhD study, transcript abundance of cysteine proteases at early and mature stages of nodule development was relatively constant, with different cysteine proteases (cathepsin-B-, F-, L- and H-like activity) contributing towards the overall proteolytic activity. Most tested nodule cystatins had preferential affinity to cathepsin L-like cysteine proteases. At the onset of senescence, however, cysteine protease

transcript abundance was nearly doubled and correlated with the progression of senescence. However, only transcription of Glyma06g18390, which was very lowly transcribed, changed significantly in transcription levels due to the onset of senescence. This cysteine protease is homologous to the senescence-related SAG12 (At5g45890). In a previous and comprehensive transcriptomics study in *Medicago truncatula* to understand nodule senescence, four cysteine protease genes, highly homologous to SAG12, were also the most abundant transcripts (Van de Velde et al. 2006). Future research has to determine the reason why soybean determinate nodules and *Medicago* indeterminate nodules differ greatly in the number of senescence-related cysteine proteases expressed.

To analyse any endogenous cystatin function in nodules, it is further crucial to identify their possible endogenous target cysteine proteases. Only little is so far known about any possible *in vivo* interaction between cystatins and their target proteases (Diaz-Mendoza et al. 2014). Cystatin and cysteine proteases sequences were therefore analysed in this study for the presence of signal peptides to obtain possible information about cellular localisation of expressed cysteine proteases and cystatins which also included a histochemical analysis. Based on signal peptide analysis, cystatins Glyma05g28250, Glyma07g39590 and Glyma13g25870 are likely localised in the secretory pathway. The ER has a vast protein storage protein capacity and relatively low proteolytic activity and cystatins might contribute to low proteolytic activity in this organelle (Erwin Ivessa et al. 1999). Nodule cystatins, such as Glyma05g28250, Glyma07g39590 and Glyma13g25870 with signal peptides, are possibly able to interact with C1A proteases that also have secretion sequences in their gene sequences (Grudkowska and Zagdanska 2004). Chrispeels and Raikhel (1992) further predicted that the open reading frame of two cysteine proteases contain a putative vacuolar-targeting signal. Such signals were also found for Glyma04g36470 and Glyma06g18390, despite that,

Glyma06g18390 was explicitly expressed during nodule senescence. This raises the question of the particular Glyma06g18390 function when targeted to the vacuole. One possible answer might be that this vacuolar-targeting signal directs these cysteine proteases to the bacteroid-containing symbiosome compartment for bacteroid protein degradation to affect peribacteriod membrane stability with rupture of the peribacteriod membrane then eliminating the microbial partner (Pladys and Vance 1993).

Finally, there was also an interest in this study to determine the interaction affinities between selected actively and non-actively transcribed cystatins during nodule development. Investigation provided first information about the relative activities and specificities of both expressed and non-expressed soybean cystatin genes. An overlap was found in this study in activities and specificities of expressed and non-expressed cystatin genes raising the question of whether non-transcribed cystatins provide a reservoir for responses to particular environmental conditions. This study also hypothesized that a strong correlation in expression of the protease and the inhibitor would enable identifying the interaction between these two components during nodule development. Not much support for the originally set hypothesis was found and the number of active proteins from each respective gene family was greater than expected. Also, cystatins can interact to various degrees with different proteases. Sugawara et al. (2002) previously reported that serine- and cysteine protease inhibitors were down-regulated in senescing organs when cysteine proteases were up-regulated. This finding could not be confirmed in this PhD study because only a macro-dissection, and not a micro-dissection, was carried out, which allowed measurement of transcripts from various nodule layers with different age. Furthermore antibodies designed and produced for each specific cysteine protease and cystatin active at various time points during nodule developmental were not available during the PhD study. Use of an available cysteine protease antiserum raised

either against a recombinant PsCYP15A protein (Vincent and Brewin 2000) or the rice cystatin (OCI) (Van der Vyver et al. 2003) did not detect any expressed proteins in microtomic slices of nodules. It cannot be excluded, however, that either a very low specificity of applied antibodies raised against non-soybean proteins or that amounts of expressed proteins being too low for detection have resulted in this negative result. A repeat of this histochemical study should therefore be envisaged with more specific antibodies and also determining any cross-reaction by simple gel-based Western blotting analysis. Applying bright-field microscopy might not have been adequate to detect the localisation of these proteins. The use of electron microscopy will be more adequate. Furthermore, the detection limits of the colorimetric technique used might not have been sufficient and perhaps an immune-gold technique would work better.

4.3 Conclusions and future work

In this PhD study, the phylogenetic relationship and transcription of individual components of the cysteine protease–cystatin system in soybean nodules during nodule development and senescence were successfully characterised. Several of these components, including the legumain-type cysteine proteases, had coordinated transcription during nodule senescence. This strongly indicates their direct involvement in nodule senescence. The study has thereby created substantial new knowledge about types of cystatin and cysteine proteases expressed, the timing of their coordinated expression, as well as establishing the inhibitory activity of actively as well as non-transcribed nodule cystatins. Since little is currently known about the *in vivo* interaction between the two components of the cysteine protease-cystatin system, their interaction might be investigated in the future in greater detail by *in vitro* assays with both purified recombinant cystatins and purified cysteine proteases. Such future studies might also

include testing if particularly stress-induced premature nodule senescence, for example due to drought or cold treatment, can be prevented when a cystatin-based protection strategy is applied. Prolonging of the period of active nitrogen fixation, by delaying nodule senescence and regulating cysteine protease activity, could ultimately provide the benefit of better soybean growth and improved soybean yield potentially contributing to future food security. Unfortunately, performing a first localization study applying an immuno-histochemical analysis did not result in any successful cellular detection of cysteine proteases or cystatins. These experiments have to be repeated preferably using more specific antisera. Although this PhD study predominantly focussed on cysteine proteases and cystatins, establishment of a RNA-Seq database for expressed genes during nodule development also provided for a catalogue of genes that significantly change their expression during nodule development. This included genes involved in glycolysis, carbon fixation, and also the TCA cycle which can act as indicators for *Rhizobium*-plant interaction during the whole process of nodule development. Besides allowing in the future more detailed studies of particular function of genes during nodule senescence, the established RNA-Seq database will also allow further comparative gene expression studies, for example comparing gene expression during natural senescence with expression during premature stress-induced senescence. Any detection of uniqueness of gene expression under specific environmental conditions might be ultimately useful in the future for plant performance improvement applying either a GMO strategy or marker-assisted plant selection.

CHAPTER 5

LITERATURE CITED

Al-Karaki GN. 2000. Morphological and Yield Traits of Wild Legume (*Tetragonolobus palaestinus* Boiss) Populations. *Journal of Agronomy and Crop Science*, 184: 267-270.

Alesandrini F, Frendo P, Puppo A, Hérouart D. 2003a. Isolation of a molecular marker of soybean nodule senescence. *Plant Physiology and Biochemistry*, 41: 727-732.

Alesandrini F, Mathis R, Van de Sype G, Herouart D, Puppo A. 2003b. Possible roles for a cysteine protease and hydrogen peroxide in soybean nodule development and senescence. *New Phytologist*, 158: 131-138.

Anon. Available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Anon. A Companion to Plant Physiology, Fifth edition: Topic 12.1. Development of a Root Nodule. Available at http://5e.plantphys.net/images/ch12/wt1201a_s.png.

Appleby CA. 1984. Leghemoglobin and *Rhizobium* respiration. *Annual Review of Plant Physiology*, 35: 443-478.

Arai S, Matsumoto I, Emori Y, Abe K. 2002. Plant Seed Cystatins and Their Target Enzymes of Endogenous and Exogenous Origin. *Journal of Agricultural and Food Chemistry*, 50: 6612-6617.

Babiychuk E, Bouvier-Nave P, Compagnon V, Suzuki M, Muranaka T, Van Montagu M, Kushnir S, Schaller H. 2008. Allelic mutant series reveal distinct functions for *Arabidopsis* cycloartenol synthase 1 in cell viability and plastid biogenesis. *Proceedings of the National Academy of Sciences*, 105: 3163-3168.

Bateman HG, Clark JH. 2000. Soybean Based Feeds for Dairy Cows. *Illinois Livestock Trail Dairy Cattle*.

Beers EP, Jones AM, Dickerman AW. 2004. The S8 serine, C1A cysteine and A1 aspartic protease families in *Arabidopsis*. *Phytochemistry*, 65: 43-58.

Beers EP, Woffenden BJ, Zhao C. 2000. Plant proteolytic enzymes: possible roles during programmed cell death. *Plant Molecular Biology*, 44: 399-415.

Belenghi B, Acconcia F, Trovato M, Perazzolli M, Bocedi A, Polticelli F, Ascenzi P, Delledonne M. 2003. AtCYS1, a cystatin from *Arabidopsis thaliana*, suppresses hypersensitive cell death. *European Journal of Biochemistry*, 270: 2593-2604.

Benchabane M, Schluter U, Vorster J, Goulet MC, Michaud D. 2010. Plant cystatins. *Biochimie*, 92: 1657-1666.

Benedito VA, Torres-Jerez I, Murray JD, Andriankaja A, Allen S, Kakar K, Wandrey M, Verdier J, Zuber H, Ott T, *et al.* 2008. A gene expression atlas of the model legume *Medicago truncatula*. *The Plant Journal*, 55: 504-513.

Bergersen F. 1997. Regulation of nitrogen fixation in infected cells of leguminous root nodules in relation to O2 supply. *Plant and Soil*, 191: 189-203.

Beyene G, Foyer CH, Kunert KJ. 2006. Two new cysteine proteinases with specific expression patterns in mature and senescent tobacco (*Nicotiana tabacum* L.) leaves. *Journal of Experimental Botany*, 57: 1431-1443.

Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A. 2010. Manipulation of FASTQ data with Galaxy. *Bioinformatics*, 26: 1783-1785.

Bottari A, Capocchi A, Galleschi L, Jopova A, Saviozzi F. 1996. Asparaginyl endopeptidase during maturation and germination of durum wheat. *Physiologia Plantarum*, 97: 475-480.

Bradford MM. 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry*, 72: 248-254.

Burns MJ, Nixon GJ, Foy CA, Harris N. 2005. Standardisation of data from real-time quantitative PCR methods - evaluation of outliers and comparison of calibration curves. *BMC Biotechnology*, 5: 31.

Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL, *et al.* 2009. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clinical chemistry*, 55: 611-622.

Cabello P, Roldán MD, Castillo F, Moreno-Vivián C. 2009. Nitrogen Cycle. In: Schaechter M editor. *Encyclopedia of Microbiology (Third Edition)*. Oxford: Academic Press. p. 299-321.

Callis J. 1995. Regulation of Protein Degradation. *Plant Cell*, 7: 845-857.

Chrispeels MJ, Raikhel NV. 1992. Short peptide domains target proteins to plant vacuoles. *Cell*, 68: 613-616.

Chrispeels MJ, Sadava DE. 2003. *Plants, genes, and crop biotechnology*. Jones & Bartlett Learning.

Chu MH, Liu KL, Wu HY, Yeh KW, Cheng YS. 2011. Crystal structure of tarocystatin-papain complex: implications for the inhibition property of group-2 phytocystatins. *Planta*, 234: 243-254.

Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38: 1767-1771.

Colebatch G, Desbrosses G, Ott T, Krusell L, Montanari O, Kloska S, Kopka J, Udvardi MK. 2004. Global changes in transcription orchestrate metabolic differentiation during symbiotic nitrogen fixation in *Lotus japonicus*. *The Plant Journal*, 39: 487-512.

Costanzo ME, Andrade A, del Carmen Tordable M, Cassan F, Abdala G. 2012. Production and function of jasmonates in nodulated roots of soybean plants inoculated with *Bradyrhizobium japonicum*. *Archives of Microbiology*, 194: 837-845.

Crespi M, Galvez S. 2000. Molecular Mechanisms in Root Nodule Development. *Journal of Plant Growth Regulation*, 19: 155-166.

Dakora FD, Keya SO. 1997. Contribution of legume nitrogen fixation to sustainable agriculture in Sub-Saharan Africa. *Soil Biology and Biochemistry*, 29: 809-817.

Dalton DA, Post CJ, Langeberg L. 1991. Effects of ambient oxygen and of fixed nitrogen on concentrations of glutathione, ascrobate, and associated enzymes in soybean root nodules. *Plant Physiology*, 96: 812-818.

Day DA, Kaiser BN, Thomson R, Udvardi MK, Moreau S, Puppo A. 2001. Nutrient transport across symbiotic membranes from legume nodules. *Functional Plant Biology*, 28: 669-676.

Demirevska K, Simova-Stoilova L, Fedina I, Georgieva K, Kunert K. 2010. Response of Oryzacystatin I Transformed Tobacco Plants to Drought, Heat and Light Stress. *Journal of Agronomy and Crop Science*, 196: 90-99.

Department of Agriculture, Forestry and Fisheries. 2013. Grain market quarterly No.1.

Devienne-Barret F, Justes E, Machet JM, Mary B. 2000. Integrated Control of Nitrate Uptake by Crop Growth Rate and Soil Nitrate Availability under Field Conditions. *Annals of Botany*, 86: 995-1005.

Diaz-Mendoza M, Velasco-Arroyo B, Gonzalez-Melendi P, Martinez M, Diaz I. 2014. C1A cysteine protease-cystatin interactions in leaf senescence. *Journal of Experimental Botany*, 65: 3825-3833.

Dovring F. 1974. Soybeans. *Scientific American*, 230: 14-&.

Downie JA. 2005. Legume haemoglobins: symbiotic nitrogen fixation needs bloody nodules. *Current Biology*, 15: R196-198.

Driscoll BT, Finan TM. 1993. NAD+-dependent malic enzyme of *Rhizobium meliloti* is required for symbiotic nitrogen fixation. *Molecular Microbiology*, 7: 865-873.

Du Z, Zhou X, Ling Y, Zhang Z, Su Z. 2010. AgriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Research, 38: 64-70.

Dupont L, Hérouart D, Alloing G, Hopkins J, Pierre O, Frendo P, Msehli SE. 2012. *The Legume Root Nodule: From Symbiotic Nitrogen Fixation to Senescence*. INTECH Open Access Publisher.

Duranti M, Gius C. 1997. Legume seeds: protein content and nutritional value. *Field Crops Research*, 53: 31-45.

El Yahyaoui F, Kuster H, Ben Amor B, Hohnjec N, Puhler A, Becker A, Gouzy J, Vernie T, Gough C, Niebel A*, et al.* 2004. Expression profiling in *Medicago truncatula* identifies more than 750 genes differentially expressed during nodulation, including many potential regulators of the symbiotic program. *Plant Physiology*, 136: 3159-3176.

Erwin Ivessa N, Kitzmüller C, de Virgilio M. 1999. Endoplasmic-reticulum-associated protein degradation inside and outside of the endoplasmic reticulum. *Protoplasma*, 207: 16-23.

Espinosa-Victoria D, Vance CP, Graham PH. 2000. Host Variation in Traits Associated with Crown Nodule Senescence in Soybean. *Crop Science*, 40: 103-109.

Fan S-G, Guo-Jiang W. 2005. Characteristics of plant proteinase inhibitors and their applications in combating phytophagous insects. *Botanical Bulletin of Academia Sinica*, 46.

Fenta BA, Driscoll SP, Kunert KJ, Foyer CH. 2012. Characterization of Drought-Tolerance Traits in Nodulated Soya Beans: The Importance of Maintaining Photosynthesis and Shoot

Biomass Under Drought-Induced Limitations on Nitrogen Metabolism. *Journal of Agronomy and Crop Science*, 198: 92-103.

Ferguson BJ, Mathesius U. 2003. Signaling interactions during nodule development. *Journal of Plant Growth Regulation*, 22: 47-72.

Food and Agriculture Organization of The United Nations Statistics Division. 2014. Food and Nutrition in Numbers 2014.

Foyer CH, Noctor G. 2005a. Oxidant and antioxidant signalling in plants: a re-evaluation of the concept of oxidative stress in a physiological context. *Plant, Cell and Environment*, 28: 1056-1071.

Foyer CH, Noctor G. 2005b. Redox homeostasis and antioxidant signaling: a metabolic interface between stress perception and physiological responses. *Plant Cell*, 17: 1866-1875.

Gilroy EM, Hein I, van der Hoorn R, Boevink PC, Venter E, McLellan H, Kaffarnik F, Hrubikova K, Shaw J, Holeva M*, et al.* 2007. Involvement of cathepsin B in the plant disease resistance hypersensitive response. *The Plant Journal*, 52: 1-13.

Goulet MC, Dallaire C, Vaillancourt LP, Khalf M, Badri AM, Preradov A, Duceppe MO, Goulet C, Cloutier C, Michaud D. 2008. Tailoring the specificity of a plant cystatin toward herbivorous insect digestive cysteine proteases by single mutations at positively selected amino acid sites. *Plant Physiology*, 146: 1010-1019.

Grudkowska M, Zagdanska B. 2004. Multifunctional role of plant cysteine proteinases. *Acta Biochimica Polonica*, 51: 609-624.

Grummer RR, Luck ML, Barmore JA. 1994. Lactational Performance of Dairy Cows Fed Raw Soybeans, with or Without Animal By-product Proteins, or Roasted Soybeans. *Journal of Dairy Science*, 77: 1354-1359.

Hara-Nishimura I, Hatsugai N, Nakaune S, Kuroyanagi M, Nishimura M. 2005. Vacuolar processing enzyme: an executor of plant cell death. *Current Opinion in Plant Biology*, 8: 404-408.

Henderson PJ. 1972. A linear equation that describes the steady-state kinetics of enzymes and subcellular particles interacting with tightly bound inhibitors. *Biochemical Journal*, 127: 321-333.

Hirsch AM. 1992. Tansley Review No. 40. Developmental Biology of Legume Nodulation. *New Phytologist*, 122: 211-237.

Howard JB, Rees DC. 1996. Structural Basis of Biological Nitrogen Fixation. *Chem Rev*, 96: 2965-2982.

Huffaker RC. 1990. Proteolytic activity during senescence of plants. *New Phytologist*, 116: 199-231.

Hungria M, Franchini JC, Campo RJ, Crispino CC, Moraes JZ, Sibaldelli RNR, Mendes IC, Arihara J. 2006. Nitrogen nutrition of soybean in Brazil: Contributions of biological N2 fixation and N fertilizer to grain yield. *Canadian Journal of Plant Science*, 86: 927-939.

Ihara M, Koyama H, Uchimura Y, Saitoh H, Kikuchi A. 2007. Noncovalent binding of small ubiquitin-related modifier (SUMO) protease to SUMO is necessary for enzymatic activities and cell growth. *Journal of Biological Chemistry*, 282: 16465-16475.

Interagency Agricultural Projections Committee. 2014. USDA Agricultural Projections to 2023. In: Agriculture USDo editor.

Ishler V, Varga G. 2000. Soybeans and soybean byproducts for dairy cattle. *Dairy and Animal Science*. College of Agricultural Sciences, Cooperative Extension, Department of Dairy and Animal Science: The Pennsylvania State University, USA, 2000; 1 – 12.

James O. 2010. Introduction. *Genetics, Genomics, and Breeding of Soybean*: Science Publishers. p. 1-18.

Jensen WA. 1962. *Botanical histochemistry: principles and practice*. San Francisco: W.H. Freeman.

Johansen DA. 1940. *Plant Microtechnique*. New York; London: McGraw-Hill Book Company, Inc.

Johnson GV, Evans HJ, Ching T. 1966. Enzymes of the glyoxylate cycle in *Rhizobia* and nodules of legumes. *Plant Physiology*, 41: 1330-1336.

Kardailsky IV, Brewin NJ. 1996. Expression of cysteine protease genes in pea nodule development and senescence. *Molecular Plant-Microbe Interactions Journal*, 9: 689-695.

Kaschuk G, Yin X, Hungria M, Leffelaar PA, Giller KE, Kuyper TW. 2012. Photosynthetic adaptation of soybean due to varying effectiveness of N2 fixation by two distinct *Bradyrhizobium japonicum* strains. *Environmental and Experimental Botany*, 76: 1-6.

Keyser H, Li F. 1992. Potential for increasing biological nitrogen fixation in soybean. In: Ladha JK, George T, Bohlool BB editors. *Biological Nitrogen Fixation for Sustainable Agriculture*. *Developments in Plant and Soil Sciences*: Springer Netherlands. p. 119-135.

Kiggundu A, Goulet MC, Goulet C, Dubuc JF, Rivard D, Benchabane M, Pepin G, van der Vyver C, Kunert K, Michaud D. 2006. Modulating the proteinase inhibitory profile of a plant cystatin by single mutations at positively selected amino acid sites. *The Plant Journal*, 48: 403-413.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14: R36.

King CA, Purcell LC. 2001. Soybean Nodule Size and Relationship to Nitrogen Fixation Response to Water Deficit. *Crop Science*, 41: 1099.

Kiyosaki T, Matsumoto I, Asakura T, Funaki J, Kuroda M, Misaka T, Arai S, Abe K. 2007. Gliadain, a gibberellin-inducible cysteine proteinase occurring in germinating seeds of wheat,

Triticum aestivum L., specifically digests gliadin and is regulated by intrinsic cystatins. *FEBS Journal*, 274: 1908-1917.

Koch K. 2004. Sucrose metabolism: regulatory mechanisms and pivotal roles in sugar sensing and plant development. *Current Opinion in Plant Biology*, 7: 235-246.

Koh S, Wiles AM, Sharp JS, Naider FR, Becker JM, Stacey G. 2002. An oligopeptide transporter gene family in *Arabidopsis*. *Plant Physiology*, 128: 21-29.

Ladrera R, Marino D, Larrainzar E, Gonzalez EM, Arrese-Igor C. 2007. Reduced carbon availability to bacteroids and elevated ureides in nodules, but not in shoots, are involved in the nitrogen fixation response to early drought in soybean. *Plant Physiology*, 145: 539-546.

Laemmli UK. 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, 227: 680-685.

Lee H, Hur CG, Oh CJ, Kim HB, Pakr SY, An CS. 2004. Analysis of the root nodule-enhanced transcriptome in soybean. *Molecules and Cells*, 18: 53-62.

Li Y, Zhou L, Chen D, Tan X, Lei L, Zhou J. 2008. A nodule-specific plant cysteine proteinase, AsNODF32, is involved in nodule senescence and nitrogen fixation activity of the green manure legume *Astragalus sinicus*. *New Phytologist*, 180: 185-192.

Lievens S, Goormachtig S, Holsters M. 2004. Nodule-enhanced protease inhibitor gene: emerging patterns of gene expression in nodule development on *Sesbania rostrata*. *Journal of Experimental Botany*, 55: 89-97.

Lim PO, Woo HR, Nam HG. 2003. Molecular genetics of leaf senescence in *Arabidopsis*. *Trends in Plant Science*, 8: 272-278.

Liu K. 1997. Chemistry and Nutritional Value of Soybean Components. *Soybeans*: Springer US. p. 25-113.

Lodwig E, Poole P. 2003. Metabolism of *Rhizobium* Bacteroids. *Critical Reviews in Plant Sciences*, 22: 37-78.

Lohman KN, Gan S, John MC, Amasino RM. 1994. Molecular analysis of natural leaf senescence in *Arabidopsis thaliana*. *Physiologia Plantarum*, 92: 322-328.

Long SR. 1996. *Rhizobium* symbiosis: nod factors in perspective. *Plant Cell*, 8: 1885-1898.

Long SR. 2001. Genes and signals in the *Rhizobium*-legume symbiosis. *Plant Physiology*, 125: 69-72.

Lucas MM, Van de Sype G, Hérouart D, Hernández MJ, Puppo A, de Felipe MR. 1998. Immunolocalization of ferritin in determinate and indeterminate legume root nodules. *Protoplasma*, 204: 61-70.

Luyten E, Vanderleyden J. 2000. Survey of genes identified in *Sinorhizobium meliloti* spp., necessary for the development of an efficient symbiosis. *European Journal of Soil Biology*, 36: 1-26.

Major IT, Constabel CP. 2008. Functional analysis of the Kunitz trypsin inhibitor family in poplar reveals biochemical diversity and multiplicity in defense against herbivores. *Plant Physiology*, 146: 888-903.

Malik NS, Pfeiffer NE, Williams DR, Wagner FW. 1981. Peptidohydrolases of Soybean Root Nodules: Identification, separation, and partial characterization of enzymes from bacteroid-free extracts. *Plant Physiology*, 68: 386-392.

Manen JF, Simon P, Van Slooten JC, Osterås M, Frutiger S, Hughes GJ. 1991. A nodulin specifically expressed in senescent nodules of winged bean is a protease inhibitor. *Plant Cell*, 3: 259-270.

Maris A, Suslov D, Fry SC, Verbelen JP, Vissenberg K. 2009. Enzymic characterization of two recombinant xyloglucan endotransglucosylase/hydrolase (XTH) proteins of *Arabidopsis* and their effect on root growth and cell wall extension. *Journal of Experimental Botany*, 60: 3959-3972.

Martineau B, McBride KE, Houck CM. 1991. Regulation of metallocarboxypeptidase inhibitor gene expression in tomato. *Molecular Genetics and Genomics*, 228: 281-286.

Martinez DE, Bartoli CG, Grbic V, Guiamet JJ. 2007a. Vacuolar cysteine proteases of wheat (*Triticum aestivum* L.) are common to leaf senescence induced by different factors. *Journal of Experimental Botany*, 58: 1099-1107.

Martinez M, Cambra I, Carrillo L, Diaz-Mendoza M, Diaz I. 2009. Characterization of the entire cystatin gene family in barley and their target cathepsin L-like cysteine-proteases, partners in the hordein mobilization during seed germination. *Plant Physiology*, 151: 1531-1545.

Martinez M, Diaz-Mendoza M, Carrillo L, Diaz I. 2007b. Carboxy terminal extended phytocystatins are bifunctional inhibitors of papain and legumain cysteine proteinases. *FEBS Letters*, 581: 2914-2918.

Massonneau A, Condamine P, Wisniewski JP, Zivy M, Rogowsky PM. 2005. Maize cystatins respond to developmental cues, cold stress and drought. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1729: 186-199.

Matamoros MA, Baird LM, Escuredo PR, Dalton DA, Minchin FR, Iturbe-Ormaetxe I, Rubio MC, Moran JF, Gordon AJ, Becana M. 1999. Stress-induced legume root nodule senescence. Physiological, biochemical, and structural alterations. *Plant Physiology*, 121: 97-112.

Maunoury N, Redondo-Nieto M, Bourcy M, Van de Velde W, Alunni B, Laporte P, Durand P, Agier N, Marisa L, Vaubert D*, et al.* 2010. Differentiation of symbiotic cells and endosymbionts in *Medicago truncatula* nodulation are coupled to two transcriptome-switches. *PLoS ONE*, 5: e9519.

McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, Young LJ, Nuzhdin SV. 2011. RNA-seq: technical variability and sampling. *BMC Genomics*, 12: 293.

Michaud D, Nguyen-Quoc B, Yelle S. 1993. Selective inhibition of Colorado potato beetle cathepsin H by oryzacystatins I and II. *FEBS Letters*, 331: 173-176.

Mortier V, Fenta BA, Martens C, Rombauts S, Holsters M, Kunert K, Goormachtig S. 2011. Search for nodulation-related CLE genes in the genome of *Glycine max*. *Journal of Experimental Botany*, 62: 2571-2583.

National Soybean Research Laboratory. 2014. National Soybean Research Laboratory: Soy Benefits. Available at http://www.nsrl.uiuc.edu/soy_benefits.html.

Noodén LD, Guiamét JJ, John I. 1997. Senescence mechanisms. *Physiologia Plantarum*, 101: 746-753.

NordiQC. 2005. Epitope retrieval: Demasking antigens. Available at http://www.nordiqc.org/Techniques/Epitope_retrieval.htm.

O'Brien TP, McCully ME. 1981. *The Study of Plant Structure: Principles and Selected Methods*. Melbourne, Aus: Termarcarphi Pty, Ltd.

Oh CJ, Lee H, Kim HB, An CS. 2004. Isolation and characterization of a root nodule-specific cysteine proteinase cDNA from soybean. *Journal of Plant Biology*, 47: 216-220.

Oldroyd GE, Murray JD, Poole PS, Downie JA. 2011. The rules of engagement in the legume-Rhizobial symbiosis. *Annual Review of Genetics*, 45: 119-144.

Oliveira AS, Xavier-Filho J, Sales MP. 2003. Cysteine proteinases and cystatins. *Brazilian Archives of Biology and Technology*, 46: 91-104.

Ott T, van Dongen JT, Gunther C, Krusell L, Desbrosses G, Vigeolas H, Bock V, Czechowski T, Geigenberger P, Udvardi MK. 2005. Symbiotic leghemoglobins are crucial for nitrogen fixation in legume root nodules but not for general plant growth and development. *Current Biology*, 15: 531-535.

Peng YL, Shirano Y, Ohta H, Hibino T, Tanaka K, Shibata D. 1994. A novel lipoxygenase from rice. Primary structure and specific expression upon incompatible infection with rice blast fungus. *Journal of Biological Chemistry*, 269: 3755-3761.

Pereira PAA, Miranda BD, Attewell JR, Kmiecik KA, Bliss FA. 1993. Selection for increased nodule number in common bean (*Phaseolus vulgaris* L.). *Plant and Soil*, 148: 203-209.

Pfeiffer NE, Torres CM, Wagner FW. 1983. Proteolytic Activity in Soybean Root Nodules : Activity in Host Cell Cytosol and Bacteroids throughout Physiological Development and Senescence. *Plant Physiology*, 71: 797-802.

Piffanelli P, Zhou F, Casais C, Orme J, Jarosch B, Schaffrath U, Collins NC, Panstruga R, Schulze-Lefert P. 2002. The barley MLO modulator of defense and cell death is responsive to biotic and abiotic stress stimuli. *Plant Physiology*, 129: 1076-1085.

Pillay P, Schluter U, van Wyk S, Kunert KJ, Vorster BJ. 2014. Proteolysis of recombinant proteins in bioengineered plant cells. *Bioengineered*, 5: 15-20.

Pladys D, Vance CP. 1993. Proteolysis during Development and Senescence of Effective and Plant Gene-Controlled Ineffective Alfalfa Nodules. *Plant Physiology*, 103: 379-384.

Prins A, van Heerden PD, Olmos E, Kunert KJ, Foyer CH. 2008. Cysteine proteinases regulate chloroplast protein content and composition in tobacco leaves: a model for dynamic interactions with ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) vesicular bodies. *Journal of Experimental Botany*, 59: 1935-1950.

Puppo A, Groten K, Bastian F, Carzaniga R, Soussi M, Lucas MM, de Felipe MR, Harrison J, Vanacker H, Foyer CH. 2005. Legume nodule senescence: roles for redox and hormone signalling in the orchestration of the natural aging process. *New Phytologist*, 165: 683-701.

Quain MD, Makgopa ME, Marquez-Garcia B, Comadira G, Fernandez-Garcia N, Olmos E, Schnaubelt D, Kunert KJ, Foyer CH. 2014. Ectopic phytocystatin expression leads to

enhanced drought stress tolerance in soybean (*Glycine max*) and *Arabidopsis thaliana* through effects on strigolactone pathways and can also result in improved seed traits. *Plant Biotechnology Journal*, 12: 903-913.

Qureshi MI, Muneer S, Bashir H, Ahmad J, Iqbal M. 2010. Chapter 1 - Nodule Physiology and Proteomics of Stressed Legumes. In: Jean-Claude K, Michel D editors. *Advances in Botanical Research*: Academic Press. p. 1-48.

Rees DC, Akif Tezcan F, Haynes CA, Walton MY, Andrade S, Einsle O, Howard JB. 2005. *Structural basis of biological nitrogen fixation*, vol. 363.

Richau KH, Kaschani F, Verdoes M, Pansuriya TC, Niessen S, Stuber K, Colby T, Overkleeft HS, Bogyo M, Van der Hoorn RA. 2012. Subclassification and biochemical analysis of plant papain-like cysteine proteases displays subfamily-specific characteristics. *Plant Physiology*, 158: 1583-1599.

Roberts IN, Caputo C, Criado MV, Funk C. 2012. Senescence-associated proteases in plants. *Physiologia Plantarum*, 145: 130-139.

Roponen I. 1970. The Effect of Darkness on the Leghemoglobin Content and Amino Acid Levels in the Root Nodules of Pea Plants. *Physiologia Plantarum*, 23: 452-460.

Ryan CA. 2000. The systemin signaling pathway: differential activation of plant defensive genes. *Biochim Biophys Acta*, 1477: 112-121.

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, *et al.* 2003. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, 34: 374-378.

Sakamoto W, Zaltsman A, Adam Z, Takahashi Y. 2003. Coordinated regulation and complex formation of yellow variegated1 and yellow variegated2, chloroplastic FtsH metalloproteases involved in the repair cycle of photosystem II in *Arabidopsis* thylakoid membranes. *Plant Cell*, 15: 2843-2855.

Salas CE, Gomes MT, Hernandez M, Lopes MT. 2008. Plant cysteine proteinases: evaluation of the pharmacological activity. *Phytochemistry*, 69: 2263-2269.

Salvesen G, Nagase H. 1989. *Inhibition of proteolytic enzymes*. New York: IRL Press.

Sanginga N. 2003. Role of biological nitrogen fixation in legume based cropping systems; a case study of West Africa farming systems. *Plant and Soil*, 252: 25-39.

Schilmiller AL, Howe GA. 2005. Systemic signaling in the wound response. *Current Opinion in Plant Biology*, 8: 369-377.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J*, et al.* 2010. Genome sequence of the palaeopolyploid soybean. *Nature*, 463: 178-183.

Schultze M, Kondorosi A. 1998. Regulation of symbiotic root nodule development. *Annual Review of Genetics*, 32: 33-57.

Severin AJ, Woody JL, Bolon YT, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE*, et al.* 2010. RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome. *BMC Plant Biology*, 10: 160.

Shah VK, Brill WJ. 1977. Isolation of an iron-molybdenum cofactor from nitrogenase. *Proceedings of the National Academy of Sciences*, 74: 3249-3253.

Sheokand S, Dahiya P, Vincent JL, Brewin NJ. 2005. Modified expression of cysteine protease affects seed germination, vegetative growth and nodule development in transgenic lines of *Medicago truncatula*. *Plant Science*, 169: 966-975.

Siddiqi MR. 2000. *Tylenchida: parasites of plants and insects*. CABI.

Snyder HE, Kwon TW. 1987. Soybean Utilization. New York: Van Nostrand Reinhold Co.

Sohani MM, Schenk PM, Schultz CJ, Schmidt O. 2009. Phylogenetic and transcriptional analysis of a strictosidine synthase-like gene family in *Arabidopsis thaliana* reveals involvement in plant defence responses. *Plant Biology*, 11: 105-117.

Solomon M, Belenghi B, Delledonne M, Menachem E, Levine A. 1999. The involvement of cysteine proteases and protease inhibitor genes in the regulation of programmed cell death in plants. *Plant Cell*, 11: 431-444.

Stacey G. 2007. Chapter 10 - The Rhizobium-Legume Nitrogen-Fixing Symbiosis. In: Newton HBJFE editor. *Biology of the Nitrogen Cycle*. Amsterdam: Elsevier. p. 147-163.

Sugawara H, Shibuya K, Yoshioka T, Hashiba T, Satoh S. 2002. Is a cysteine proteinase inhibitor involved in the regulation of petal wilting in senescing carnation (*Dianthus caryophyllus* L.) flowers? *Journal of Experimental Botany*, 53: 407-413.

Swaraj K, Dhandi S, Sheokand S. 1995. Relationship between defense mechanism against activated oxygen species and nodule functioning with progress in plant and nodule development in *Cajanus cajan* L. Millsp. *Plant Science*, 112: 65-74.

Tajima T, Yamaguchi A, Matsushima S, Satoh M, Hayasaka S, Yoshimatsu K, Shioi Y. 2011. Biochemical and molecular characterization of senescence-related cysteine protease-cystatin complex from spinach leaf. *Physiologia Plantarum*, 141: 97-116.

Terpolilli JJ, Hood GA, Poole PS. 2012. Chapter 5 - What Determines the Efficiency of N2-Fixing Rhizobium-Legume Symbioses? In: Robert KP editor. *Advances in Microbial Physiology*: Academic Press. p. 325-389.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28: 511-515.

Van de Velde W, Guerra JC, De Keyser A, De Rycke R, Rombauts S, Maunoury N, Mergaert P, Kondorosi E, Holsters M, Goormachtig S. 2006. Aging in legume symbiosis. A molecular view on nodule senescence in *Medicago truncatula*. *Plant Physiology*, 141: 711-720.

Van der Hoorn RA, Leeuwenburgh MA, Bogyo M, Joosten MH, Peck SC. 2004. Activity profiling of papain-like cysteine proteases in plants. *Plant Physiology*, 135: 1170-1178.

Van der Vyver C, Schneidereit J, Driscoll S, Turner J, Kunert K, Foyer CH. 2003. Oryzacystatin I expression in transformed tobacco produces a conditional growth phenotype and enhances chilling tolerance. *Plant Biotechnology Journal*, 1: 101-112.

Van Heerden PDR, De Beer M, Mellet DJ, Maphike HS, Foit W. 2007. Growth media effects on shoot physiology, nodule numbers and symbiotic nitrogen fixation in soybean. *South African Journal of Botany*, 73: 600-605.

Vance CP, Heichel GH, Barnes DK, Bryan JW, Johnson LE. 1979. Nitrogen Fixation, Nodule Development, and Vegetative Regrowth of Alfalfa (*Medicago sativa* L.) following Harvest. *Plant Physiology*, 64: 1-8.

Vartapetian AB, Tuzhikov AI, Chichkova NV, Taliansky M, Wolpert TJ. 2011. A plant alternative to animal caspases: subtilisin-like proteases. *Cell Death & Differentiation*, 18: 1289-1297.

Vincent JL, Brewin NJ. 2000. Immunolocalization of a cysteine protease in vacuoles, vesicles, and symbiosomes of pea nodule cells. *Plant Physiology*, 123: 521-530.

Vincent JM, Newton WE, Orme-Johnson WH. 1980. Factors controlling the legume-*Rhizobium* symbiosis. *Nitrogen fixation. Volume 2: symbiotic associations and cyanobacteria.*: 103-129.

Vorster BJ, Schlüter U, du Plessis M, van Wyk SG, Makgopa ME, Ncube I, Quain MD, Kunert KJ, Foyer CH. 2013. The Cysteine Protease–Cysteine Protease Inhibitor System Explored in Soybean Nodule Development. *Agronomy*, 3: 550-570.

Walsh KB. 1990. Vascular transport and soybean nodule function. III: Implications of a continual phloem supply of carbon and water. *Plant, Cell and Environment*, 13: 893-901.

Warden CD, Yuan YC, Wu X. 2013. Optimal Calculation of RNA-Seq Fold-Change Values. *International Journal of Computational Bioinformatics and In Silico Modeling*, 2: 285 - 292.

Wiederanders B. 2003. Structure-function relationships in class CA1 cysteine peptidase propeptides. *Acta Biochimica Polonica*, 50: 691-713.

Zaltsman A, Ori N, Adam Z. 2005. Two types of FtsH protease subunits are required for chloroplast biogenesis and Photosystem II repair in *Arabidopsis*. *Plant Cell*, 17: 2782-2790.

Zhang N, Jones BL. 1995. Characterization of germinated barley endoproteolytic enzymes by two dimensional gel electrophoresis. *Journal of Cereal Science*, 21: 145–153.

CHAPTER 6

APPENDICES

6.1 APPENDIX A.

Table A.1 Tophat2 settings and parameters used for the RNA-Seq data analysis.

Input Parameter	Value
Is this library mate-paired?	paired
RNA-Seq FASTQ file, forward reads	37: 4.1
RNA-Seq FASTQ file, reverse reads	38: 4.2
Mean Inner Distance between Mate Pairs	90
Std. Dev for Distance between Mate Pairs	40
Report discordant pair alignments?	Yes
Use a built in reference genome or own from your history	history
Select the reference genome	42: Gmax_189.fa
TopHat settings to use	full
Max realign edit distance	1000
Max edit distance	2
Library Type	FR Unstranded
Final read mismatches	2
Use bowtie -n mode	No
Anchor length (at least 3)	8
Maximum number of mismatches that can appear in the anchor region of spliced alignment	0
The minimum intron length	70
The maximum intron length	500000
Allow indel search	Yes
Max insertion length.	3
Max deletion length.	3
Maximum number of alignments to be allowed	20
Minimum intron length that may be found during split-segment (default) search	50
Maximum intron length that may be found during split-segment (default) search	500000
Number of mismatches allowed in each segment alignment for reads mapped independently	2
Minimum length of read segments	25
Use Own Junctions	Yes
Use Gene Annotation Model	Yes
Gene Model Annotations	41: Gmax_189_gene_exons.gff3
Use Raw Junctions	No
Only look for supplied junctions	No

Use Coverage Search	No
Use Microexon Search	No
Do Fusion Search	No
Set Bowtie2 settings	No
Specify read group?	no

Table A.2 Cufflinks settings and parameters used for the RNA-Seq data analysis.

Input Parameter	Value
SAM or BAM file of aligned RNA-Seq reads	110: 4w Tophat2 on data 38, data 41, and others: accepted_hits
Max Intron Length	300000
Min Isoform Fraction	0.1
Pre MRNA Fraction	0.15
Perform quartile normalization	Yes
Use Reference Annotation	Use reference annotation
Reference Annotation	41: Gmax_189_gene_exons.gff3
Perform Bias Correction	No
Use multi-read correct	Yes
Use effective length correction	Yes
Global model (for use in Trackster)	No dataset

Table A.3 Cuffdiff settings and parameters used for the RNA-Seq data analysis.

Input Parameter	Value
Transcripts	1: Gmax_189_gene_exons.gff3
Name	4w
Add replicate	11: 4w Tophat2 on data 38, data 41, and others: accepted_hits
Name	8w
Add replicate	12: 8w Tophat2 on data 23, data 26, and others: accepted_hits
Name	14w
Add replicate	13: 14w Tophat2 on data 38, data 41, and others: accepted_hits
Library normalization method	classic-fpkm
Dispersion estimation method	blind
False Discovery Rate	0.05
Min Alignment Count	5
Use multi-read correct	Yes
Perform Bias Correction	No
Include Read Group Datasets	No
Set Additional Parameters? (not recommended for paired-end reads)	No

6.2 APPENDIX B.

Table A.4 Primer sets to amplify of target transcripts.

Description		Phytozome ID		5'/3'	Amplicon (bp)
CYP	**1**	Glyma04g36470	F	TCTTTGGTTCTTGGAGTGGC	145
			R	CGTTAAACCGCTTGTGCTTG	
	2	Glyma06g18390	F	ACCAACCTGTTTCTGTAGCC	149
			R	CCAGTAACTAGTCCCATCAACAG	
	3	Glyma08g12340	F	CTCCCTTGCTATGTCCAGTAAC	132
			R	TGGAACCTCTTTGCCTTCTC	
CYS	**4**	Glyma05g28250	F	AGGATCTAGGGCGGTTCTC	149
			R	CGTGGCCGATATCTTCATGTAG	
	5	Glyma15g36180	F	TCTCAGAACAGCGTCCAAAC	94
			R	CCACCCTAGAAAACTCCAGAAG	
Leghemoglobin	**6**	Glyma11g12980	F	TGCTACCCATTTTAGAACCGG	136
			R	ATCATAAGCTTCTCCCCATGC	
VPE	**7**	Glyma05g04230	F	AACGGCTATGGAAACTACAGG	137
			R	GGTCTGGGATTTAACTCGTCTG	
40SrS8	**8**	Glyma08g43690	F	GCATTATGGCGTTGAGGTTG	144
			R	CGGTTCTGCTTTCGCTTTTC	
ELF1B	**9**	Glyma02g44460	F	GTGGTACGATGCTGTCTCTTC	72
			R	CCACTGAATCTTACCCCTTGAG	

6.3 APPENDIX C.

Table A.5 Cystatin sequences identified in soybean nodules by RNA-Seq analysis with similarity to oryzacystatin-I. * indicates cystatins transcriptionally active in nodules.

	FPKM			
CYS	**4w**	**8w**	**14w**	
Glyma04g10360	0	0	0	-
Glyma05g28250	23	26	40	*
Glyma07g39590	2	1	1	-
Glyma08g11210	0	0	0	
Glyma09g01360	0	0	0	-
Glyma09g16960	0	0	0	-
Glyma11g06850	0	0	0	-
Glyma11g36780	0	0	0	-
Glyma13g04250	98	97	64	*
Glyma13g25870	68	63	68	*
Glyma13g27980	0	0	0	-
Glyma14g04250	2	7	12	*
Glyma14g04260	0	0	1	-
Glyma14g04291	0	0	0	-
Glyma15g12211	100	163	234	*
Glyma15g36180	26	40	56	*
Glyma18g00690	0	0	0	-
Glyma18g12240	0	0	1	-
Glyma19g39400	0	0	0	-
Glyma20g08800	86	74	56	*

Table A.6 Cysteine protease sequences identified in soybean nodules by RNA-Seq analysis with similarity to papain. * indicates cysteine proteases transcriptionally active in nodules.

	FPKM			
CYP	**4w**	**8w**	**14w**	
Glyma0079s00280	0	0	0	-
Glyma0079s00290	0	0	0	-
Glyma0079s00300	0	0	0	-
Glyma02g15831	0	0	0	-
Glyma02g28980	0	0	0	-
Glyma03g38520	101	103	144	*
Glyma04g01630	0	0	0	-
Glyma04g01640	0	0	0	-
Glyma04g03020	172	186	235	*
Glyma04g03090	3	3	6	*
Glyma04g04400	9	5	4	*
Glyma04g36470	0	1	3	-
Glyma05g20930	1	1	1	-
Glyma05g29131	0	0	0	-
Glyma05g29181	0	0	0	-
Glyma06g01710	0	0	0	-
Glyma06g01730	0	0	1	-
Glyma06g03050	58	51	84	*
Glyma06g04540	0	0	0	-
Glyma06g18390	0	0	8	*
Glyma06g42485	0	0	0	-
Glyma06g42520	0	0	0	-
Glyma06g42545	0	0	0	-
Glyma06g42590	0	0	0	-
Glyma06g42610	0	0	0	-
Glyma06g42620	0	0	0	-
Glyma06g42630	0	0	0	-
Glyma06g42640	0	0	0	-
Glyma06g42661	0	0	0	-
Glyma06g42670	0	0	0	-
Glyma06g42750	0	0	0	-
Glyma06g42780	0	0	0	-
Glyma06g43090	0	0	0	-
Glyma06g43100	0	0	0	-
Glyma06g43160	0	0	0	-
Glyma06g43170	0	0	0	-
Glyma06g43251	0	0	0	-
Glyma06g43530	0	0	0	-

Glyma06g43540	0	0	2	-
Glyma07g27971	0	0	0	-
Glyma07g32650	0	0	0	-
Glyma08g12270	0	0	0	-
Glyma08g12340	38	2	17	*
Glyma09g08100	87	71	215	*
Glyma10g23650	0	0	0	-
Glyma10g35100	4	5	2	*
Glyma11g12130	2	4	7	*
Glyma11g20400	0	0	0	-
Glyma12g04340	3	3	6	*
Glyma12g08180	0	0	0	-
Glyma12g08200	0	0	0	-
Glyma12g14540	0	0	0	-
Glyma12g14781	0	0	0	-
Glyma12g14925	0	0	0	-
Glyma12g15120	0	0	0	-
Glyma12g15130	0	0	0	-
Glyma12g15690	0	0	0	-
Glyma12g15725	0	0	0	-
Glyma12g15740	0	0	0	-
Glyma12g15760	0	0	0	-
Glyma12g15780	0	0	0	-
Glyma12g15790	0	0	0	-
Glyma12g17421	0	0	0	-
Glyma12g33581	1	1	1	-
Glyma13g30182	0	0	0	-
Glyma14g09440	73	80	261	*
Glyma14g40670	80	106	272	*
Glyma15g08840	0	0	0	-
Glyma15g08951	0	0	0	-
Glyma15g19580	332	365	491	*
Glyma16g16290	0	0	0	-
Glyma17g05670	44	33	37	*
Glyma17g13530	0	0	3	-
Glyma17g18440	0	1	1	-
Glyma17g35720	96	131	418	*
Glyma17g37400	181	202	525	*
Glyma18g09380	0	0	0	-
Glyma19g41120	33	43	103	*
Glyma20g32454	0	0	0	-

6.4 APPENDIX D.

Table A.7 The predicted signal peptide data generated by TargetP, include the Name, Length of protein, Final NN scores of final prediction (cTP, mTP, SP and other), Prediction of localization (Loc), Reliability class (RC), TPlen (Predicted presequence length), Chloroplast (C), Mitochondrion (M), Secretory pathway (S) and any other location (-). The reliability classes are indicated as differences (diff) between the best second best prediction, expressed from high to low; 1: diff > 0.800, 2: 0.800 > diff > 0.600, 3: 0.600 > diff > 0.400, 4: 0.400 > diff > 0.200 and 5: 0.200 > diff.

	Name	Length	cTP	mTP	SP	other	Loc	RC	TPlen
CYP	Glyma0079s00280	343	0.001	0.169	0.963	0.029	S	2	27
	Glyma0079s00290	343	0.001	0.121	0.985	0.038	S	1	27
	Glyma0079s00300	352	0.001	0.163	0.974	0.024	S	1	22
	Glyma02g15831	220	0.041	0.24	0.034	0.877	_	2	-
	Glyma02g28980	103	0.155	0.095	0.098	0.844	_	2	-
	Glyma03g38520	357	0.005	0.044	0.978	0.038	S	1	28
	Glyma04g01630	349	0.013	0.01	0.997	0.018	S	1	24
	Glyma04g01640	349	0.015	0.019	0.992	0.017	S	1	24
	Glyma04g03020	366	0.01	0.032	0.949	0.178	S	2	19
	Glyma04g03090	439	0.001	0.063	0.988	0.032	S	1	27
	Glyma04g04400	367	0.003	0.033	0.993	0.051	S	1	24
	Glyma04g36470	362	0.007	0.038	0.974	0.069	S	1	22
	Glyma05g20930	366	0.006	0.027	0.987	0.077	S	1	20
	Glyma05g29131	347	0.06	0.076	0.439	0.601	_	5	-
	Glyma05g29181	178	0.019	0.087	0.72	0.432	S	4	17
	Glyma06g01710	350	0.008	0.078	0.969	0.027	S	1	25
	Glyma06g01730	350	0.017	0.01	0.995	0.024	S	1	25
	Glyma06g03050	366	0.008	0.032	0.959	0.174	S	2	21
	Glyma06g04540	333	0.001	0.034	0.965	0.165	S	2	24
	Glyma06g18390	362	0.008	0.036	0.969	0.079	S	1	22
	Glyma06g42485	306	0.012	0.068	0.918	0.03	S	1	26
	Glyma06g42520	339	0.004	0.155	0.936	0.012	S	2	25
	Glyma06g42545	148	0.019	0.141	0.923	0.022	S	2	26
	Glyma06g42590	338	0.004	0.089	0.973	0.029	S	1	26
	Glyma06g42610	338	0.004	0.089	0.973	0.029	S	1	26
	Glyma06g42620	342	0.002	0.392	0.949	0.015	S	3	26
	Glyma06g42630	339	0.004	0.155	0.936	0.012	S	2	25
	Glyma06g42640	342	0.015	0.106	0.903	0.012	S	2	25
	Glyma06g42661	406	0.002	0.14	0.959	0.061	S	1	26
	Glyma06g42670	336	0.002	0.146	0.991	0.013	S	1	26
	Glyma06g42750	342	0.009	0.111	0.913	0.016	S	1	25

Glyma06g42780	341	0.009	0.076	0.929	0.024	S	1	26
Glyma06g43090	343	0.001	0.163	0.974	0.024	S	1	22
Glyma06g43100	343	0.001	0.121	0.985	0.038	S	1	27
Glyma06g43160	352	0.001	0.163	0.974	0.024	S	1	22
Glyma06g43170	279	0.092	0.131	0.064	0.836	_	2	-
Glyma06g43251	245	0.002	0.043	0.989	0.015	S	1	27
Glyma06g43530	343	0.001	0.1	0.986	0.037	S	1	27
Glyma06g43540	343	0.001	0.203	0.976	0.021	S	2	22
Glyma07g27971	187	0.002	0.053	0.989	0.058	S	1	20
Glyma07g32650	340	0.006	0.156	0.914	0.052	S	2	26
Glyma08g12270	379	0.125	0.127	0.633	0.027	S	3	17
Glyma08g12340	362	0.004	0.121	0.98	0.012	S	1	26
Glyma09g08100	354	0.005	0.071	0.982	0.038	S	1	22
Glyma10g23650	421	0.092	0.104	0.143	0.921	_	2	-
Glyma10g35100	380	0.002	0.201	0.959	0.032	S	2	27
Glyma11g12130	363	0.003	0.017	0.977	0.164	S	1	22
Glyma11g20400	343	0.002	0.068	0.991	0.023	S	1	28
Glyma12g04340	365	0.003	0.034	0.98	0.135	S	1	22
Glyma12g08180	343	0.006	0.047	0.985	0.04	S	1	27
Glyma12g08200	313	0.002	0.045	0.991	0.052	S	1	26
Glyma12g14540	343	0.001	0.231	0.972	0.02	S	2	27
Glyma12g14781	231	0.086	0.344	0.031	0.638	_	4	-
Glyma12g14925	201	0.091	0.459	0.056	0.479	_	5	-
Glyma12g15120	343	0.001	0.092	0.989	0.03	S	1	27
Glyma12g15130	343	0.001	0.221	0.977	0.016	S	2	27
Glyma12g15690	337	0.003	0.07	0.984	0.026	S	1	26
Glyma12g15725	309	0.006	0.08	0.963	0.128	S	1	22
Glyma12g15740	287	0.134	0.065	0.364	0.678	_	4	-
Glyma12g15760	337	0.007	0.058	0.982	0.021	S	1	26
Glyma12g15780	337	0.007	0.058	0.982	0.021	S	1	26
Glyma12g15790	333	0.005	0.411	0.819	0.037	S	3	26
Glyma12g17421	138	0.412	0.076	0.088	0.691	_	4	-
Glyma12g33581	275	0.009	0.015	0.991	0.067	S	1	25
Glyma13g30182	517	0.002	0.175	0.982	0.017	S	1	25
Glyma14g09440	496	0.648	0.031	0.088	0.429	C	4	51
Glyma14g40670	367	0.001	0.057	0.991	0.056	S	1	21
Glyma15g08840	369	0.008	0.012	0.992	0.035	S	1	30
Glyma15g08951	373	0.002	0.069	0.989	0.042	S	1	20
Glyma15g19580	329	0.003	0.059	0.98	0.054	S	1	22
Glyma16g16290	366	0.011	0.019	0.984	0.094	S	1	22
Glyma17g05670	353	0.012	0.042	0.938	0.037	S	1	21
Glyma17g13530	379	0.094	0.036	0.951	0.095	S	1	40
Glyma17g18440	366	0.01	0.014	0.992	0.066	S	1	22
Glyma17g35720	446	0.04	0.036	0.961	0.114	S	1	30

	Glyma17g37400	284	0.002	0.066	0.991	0.043	S	1	19
	Glyma18g09380	285	0.084	0.78	0.007	0.256	M	3	53
	Glyma19g41120	356	0.006	0.075	0.935	0.039	S	1	27
	Glyma20g32454	255	0.173	0.121	0.066	0.824	_	2	-

	Name	Length	cTP	mTP	SP	other	Loc	RC	TPlen
CYS	Glyma04g10360	114	0.008	0.552	0.618	0.024	S	5	23
	Glyma05g28250	130	0.019	0.018	0.989	0.028	S	1	26
	Glyma07g39590	114	0.007	0.029	0.973	0.112	S	1	20
	Glyma08g11210	103	0.066	0.283	0.055	0.796	_	3	-
	Glyma09g01360	114	0.01	0.547	0.642	0.02	S	5	23
	Glyma09g16960	114	0.008	0.554	0.602	0.024	S	5	23
	Glyma11g06850	114	0.008	0.55	0.661	0.027	S	5	23
	Glyma11g36780	126	0.189	0.039	0.897	0.031	S	2	20
	Glyma13g04250	97	0.06	0.229	0.114	0.851	_	2	-
	Glyma13g25870	245	0.005	0.214	0.876	0.035	S	2	33
	Glyma13g27980	124	0.011	0.021	0.872	0.343	S	3	19
	Glyma14g04250	103	0.049	0.092	0.174	0.917	_	2	-
	Glyma14g04260	200	0.15	0.057	0.086	0.815	_	2	-
	Glyma14g04291	184	0.164	0.033	0.177	0.606	_	3	-
	Glyma15g12211	112	0.018	0.155	0.855	0.046	S	2	22
	Glyma15g36180	245	0.015	0.538	0.396	0.034	M	5	11
	Glyma18g00690	142	0.138	0.049	0.857	0.037	S	2	20
	Glyma18g12240	120	0.042	0.109	0.221	0.885	_	2	-
	Glyma19g39400	114	0.008	0.552	0.618	0.024	S	5	23
	Glyma20g08800	97	0.062	0.217	0.119	0.863	_	2	-

6.5 APPENDIX E.

Table A.8 Primer sets to isolate target cystatin gene sequences.

PhytozomeID		5'/3'
Glyma05g28250	F	**GAAGGATCC**CGATGGCGGCGTTG
	R	**GAGAAGAATTC**TCACTGCGTGGAAGGAGCG
Glyma13g04250	F	**AAAGGATCC**AAATGGCAGCACTTGG
	R	**AAGAATTC**CTATGCAGGTGCATCTC
Glyma14g04250	F	**GAAGGATCC**AAATGGCAGCACTGG
	R	**GGAGAATTC**AGACCGTCACCGAAAGAG
Glyma20g08800	F	**GAAGGATCC**GAATGGCAGCACTTG
	R	**GAAAAGAATTC**CTATGCAGGTGCATCTCC