The transcriptome response of leaves of the resurrection plant, *Xerophyta humilis* to desiccation

Arthur Yen-Hsiang Shen

Thesis Presented for the Degree of

DOCTOR OF PHILOSOPHY

In the Department of Molecular and Cell Biology

University of Cape Town

October 2014

Abstract

In angiosperms, desiccation tolerance, a genetic trait that enables tissues to survive loss of more than 95% of cellular water is widely observed in the seeds, but is only found in the vegetative tissues of a small group of species known as the resurrection plants. *Xerophyta humilis* is a small resurrection plant indigenous to Southern Africa. In this study, the hypothesis that vegetative desiccation tolerance is derived from an adaptation of seed desiccation tolerance was tested by characterizing changes in the transcriptome of *X. humilis* leaves during desiccation. The mRNA transcript abundance of a set of 1680 *X. humilis* genes was analyzed at 6 different stages of water loss in the leaves of *X. humilis*. Functional enrichment analysis showed that genes that were down-regulated during desiccation were over-represented with genes involved in photosynthesis, cellular developmental processes, as well as transcription regulator activity. Three distinct clusters of up-regulated genes were identified. The earliest set of up-regulated genes were enriched with genes associated with the turnover of proteins and the simultaneous synthesis of proteins required for protection. Enrichment also included genes associated with lipid body synthesis, as well as the transport of storage proteins to vacuoles. Two groups of late desiccation up-regulated genes were also identified, their expression only increased at later stages of desiccation and remained high in the desiccated leaves. The late embryogenesis abundant proteins and genes associated with chlorophyll biosynthesis and metabolism were over-represented among the late desiccation up-regulated genes, speculated to be stably stored in desiccated tissues ready for the immediate translation in absence of de novo transcription when water becomes available. The expression of a common set of 772 orthologues were directly compared in the *X. humilis* microarray dataset and publicly available *Arabidopsis thaliana* microarray datasets for osmotically stressed shoots, as well as seed development. The transcript abundance of 1-cysteine peroxiredoxin 1, a well-known seed specific antioxidant, and the 3 previously reported seed specific late embryogenesis abundant proteins, were found up-regulated in desiccating *X. humilis* leaves and in developing *A. thaliana* seeds, but were absent in *A. thaliana* shoots during osmotic stress. Furthermore, PICKLE, a CHD3-chromatin-remodeling factor, and SWINGER, a core component of Polycomb Repressive Complex 2 (PRC2), which are known to repress seed maturation genes during the early stages of germination, were found to be down-regulated during desiccation in *X. humilis* leaves, but not during abiotic stress in *A. thaliana*. These results suggest that desiccation tolerance in *X. humilis* leaves, may have evolved by an extension of desiccation tolerance seen in the early

germination programme, where angiosperm seedlings can reactivate the seed desiccation transcriptome by the reversible epigenetic modification of seed maturation genes mediated by PRC2.

Acknowledgements

I would like to thank my Supervisor, Professor Nicola Illing for her guidance and support throughout my PhD study.

I would like to thank the National Research Foundations (NRF) for funding the study.

I would like to thank Dr Helen Collett for her guidance and support when I first started working on the project.

I would like to thank Dr Katherine Denby for her guidance and support in the design of and the analysis of microarray experiment and data.

I would like to thank Mrs Faezah Davids for her help when I have encountered problems with my experiments.

I would like to thank Professor Nicky Mulder, and Mr Gerrit Botha from the Computational Biology Group (CBIO), University of Cape Townfor setting up of the EST sequence analysis pipeline, and their help in Bioinformatics.

I would like to thank Dr Sally-Ann Walford for helping me in developing R scripts for my microarray data analysis.

I would like to thank all current members and past members of Lab 425 (or the Evo-Devo lab) for sharing fun and support.

I would like to thank my family for their support, understanding and believing in me.

Table of contents

Chapter 3 Microarray Gene Expression Data Analysis

Chapter 1
Plant Desiccation Tolerance

1.1 Introduction

Plants are constantly exposed to various abiotic environmental stresses such as lack of water, temperature extremes, high soil salinity, as well as oxidative stress caused by increased sunlight. Plant genomes evolved several protective mechanisms and strategies to cope with and to survive under these unfavourable environmental conditions, as they are immobile and unable to migrate to escape the harsh conditions that arise from these environmental stresses (Gechev and Hille, 2012).

As the key component of life, water comprises more than 90% of the fresh weight of most herbaceous plants. Not only is water important for maintaining cell turgor through intracellular space filling to provide structural support, it is also essential for metabolism, acting as a reactant in a number of critical biochemical reactions, and providing hydrophilic and hydrophobic interactions, which in turn, are important for controlling the intermolecular distances that determine the conformation of macromolecules and membranes, as well as the partitioning of molecules within organelles (Hoekstra *et al.*, 2001; Walters *et al.*, 2002; Wood, 2005). Water availability is one of the major limitations to plant productivity as well as one of the major factors regulating the distribution of plant species (Boyer, 1982; Delmer, 2005; Neumann, 2008).

Most terrestrial plants will face problems of transient decreases in relative water content (RWC) at some stage in their life cycle. Many of them are able to tolerate moderate dehydration, but only a few can survive extreme desiccation. Desiccation is a process of extreme dehydration, in which plants lose their free water almost completely (Alpert 2005; 2006; Wood and Jenks, 2007). Similar to the seeds of angiosperms and spores of bryophytes which survive desiccation (Dickie and Prichard, 2002; Tweddle *et al.*, 2003; Farnsworth, 2004; Alpert, 2005; Berjak *et al.*, 2007), desiccation tolerant plants possess vegetative tissues that are able to tolerate prolonged periods of severe water loss down to less than 5% RWC, and rapidly resume full normal functioning upon rehydration (Gaff, 1971). These include several species of ferns (Bewley and Krochko, 1982), algae (Trainor and Gladych 1995; Abe *et al.*, 2001), lichens (Kranner and Lutzoni, 1999), bryophytes (Oliver, 1996; 2007; Oliver

and Bewley, 1997; Oliver *et al.*, 2005) and a small group of angiosperms. These desiccation tolerant angiosperms are also known as resurrection plants (Gaff, 1971; Farrant, 2007).

1.1.1 Stresses associated with water deficit

Loss of water in plant tissues or cells leads to serious mechanical strains, which in turn results in cellular damage. During dehydration, cell turgor and cell volume are reduced as water is lost from the vacuoles and cytoplasm. As the compaction of organelles and intracellular macromolecules increases due to the shrinkage of the cytoplasm, plasmolysis occurs as plasmalemma rupture, allowing entry of extracellular hydrolases into the cytoplasm, which then causes further cell damage and ultimately cell death (Walters *et al.*, 2002). The reduction in cell volume also increases the concentration of cellular constituents, packing molecules closely together and causing them to interact in ways that might not normally occur. Such molecular aggregations can lead to denaturation of protein as well as distortion of membrane structure.

Membrane integrity plays an important role in survival of the plants. The cell membrane shields the cell from its environment, and is also the site of sensors that interpret environmental conditions (Barkla and Pantoja, 2011). In the hydrated state, cell membranes are kept functional and intact by water molecules, which are intrinsically linked to hydrophilic heads of phospholipids, facilitating their spontaneous alignment to form bilayer structures based on polarity. Hydrophobic acyl chains within bilayers are vital as they allow the anchorage of essential proteins and other constituents within membranes. When water is lost from the system, different membrane systems may become closely compressed that leads to mis-mixing of phospholipids and membrane proteins. Non-bilayer structures formed by oppressed phospholipids of different membrane systems may assemble into different conformation upon rehydration, leading to possible leakage of cellular constituents (Walters *et al.*, 2002).

Loss of water also affects normal metabolic pathways, which can be uncoupled due to mis-folding of essential enzymes. Although general metabolism may be perturbed, not all reactions are affected by dehydration in the same way. Various reactions within photosynthetic and respiratory pathways respond differently to low water content, and such differing responses result in metabolic imbalances (Farrant, 2000; Leprince *et al.*, 2000; Walters *et al.*, 2002). Furthermore, high-energy intermediates accumulate and leak out from

mitochondria and chloroplasts as a consequence of continued respiration and photosynthesis while other metabolic processes are being slowed down or stopped. These intermediates form reactive oxygen species (ROS) and free radicals such as singlet oxygen, hydroxyl radicals, hydrogen peroxide and superoxide anions (Smirnoff, 1998; Apel and Hirt, 2004; Oliver, 2007; Grene *et al.*, 2011; Dinakar *et al.*, 2012), which react with nucleic acids, lipids and proteins, causing permanent damage to chromosomes, membranes and enzymes. (Dizdaroglu, 1994; Leprince *et al.*, 2000; Dean *et al.*, 1993; Walters *et al.*, 2002). The general stresses and damages associated with water deficit, as well as mechanisms which plants have evolved to overcome them, are summarized in Table 1.1.

Table 1.1. Problems caused by desiccation and mechanisms of desiccation tolerance in resurrection plants.

Problem	Mechanism of protection	Selected references
Mechanical damage due to shrinkage of cytoplasm and organelles	Changes in cell wall composition that increase flexibility	(Jones and McQueen-Mason, 2004; Vicre *et al.*, 2004b)
	Folding of cell walls	(Van der Willigen *et al.*, 2004)
	Replacement of water in vacuoles by non-aqueous compounds and fragmentation of vacuoles	(Farrant, 2000; Vicre *et al.*, 2004a)
Physiological damage at low water availability	Up-regulation of proteins that increase membrane permeability	(Smith-Espinoza *et al.*, 2003; Van der Willigen *et al.*, 2004)
Disintegration of membranes and aggregation of macromolecules during, coalescence of lipid bodies and membrane leakage upon rehydration	Accumulation of sugars, especially non reducing disaccharides, that stabilize molecules, depress temperature (*T*m) of membrane phase change from liquid crystal to gel, and form glasses with high melting temperature (*T*g)	(Wingler, 2002; Bernacchia and Furini, 2004; Buitink and Leprince, 2004; Crowe *et al.*, 2005)
	Expression of LEA proteins, which act as molecular chaperones and interact with sugars to form glasses	(Wise and Tunnacliffe, 2004; Goyal *et al.*, 2005; Oliver *et al.*, 2005)
	Partitioning of amphiphiles into membranes	(Hoekstra and Golovina, 2002; Oliver *et al.*, 2002)
	Small stress proteins, which may act as chaperones or repair damage upon rehydration	(Collins and Clegg, 2004; Crowe *et al.*, 2005; Potts *et al.*, 2005)

Table 1.1 (continued)

Problem	Mechanism	Selected references
Disintegration of membranes and aggregation of macromolecules during, coalescence of lipid bodies and membrane leakage upon rehydration (continued)	Changes in lipid composition that stabilize membranes, such as increases in phospholipids, degree of saturation, and free sterols	(Quartacci *et al.*, 2002; Hoekstra, 2005)
	In seeds, oleosins prevent aggregation of individual oil bodies	(Murphy, *et al.*, 2001)
Generation of reactive oxygen species (ROS)	Synthesis of antioxidants during drying, maintenance of pools of reduced antioxidants and ROS-scavenging enzymes	(Shirkey *et al*., 2000; Augusti *et al.*, 2001; Espindola *et al.*, 2003; Kranner and Birtic, 2005)
	Down-regulation of photosynthesis early in drying and programmed chlorophyll loss	(Jensen *et al.*, 1999; Deng *et al*., 2003; Hirai *et al.*, 2004; Illing *et al*., 2005; Tuba *et al.*, 1996)
	Folding of leaves	(Farrant *et al.*, 2003)
Triggering of cell death by oxidized glutathione	Rapid reduction of glutathione upon rehydration	(Kranner and Birtic, 2005)
In plants, disintegration of the photosynthetic apparatus	Modification of proteins in Photosystems	(Peeva and Maslenkova, 2004)
Accumulation of damage from UV and gamma radiation and from Maillard and Fenton reactions while dry	Expression of UV-absorbing pigments	(Potts, 1996)
	Up-regulation of DNA repair pathways	(Wilson *et al.*, 2004)
	DNA protection	(Potts *et al.*, 2005)
In plants, cavitation of xylem	Low hydraulic conductivity	(Sherwin *et al.*, 1998)
Drying too fast for induction of tolerance mechanisms	Signaling for induction of general stress tolerance mechanisms via ABA	(Beckett *et al.*, 2000; Bartels and Salamini, 2001)

Modified from Alpert (2006).

1.2 Plant desiccation tolerance

Desiccation tolerance is a phenomenon widely observed in reproductive organs of some plants such as seeds, pollens, spores, or in dormant buds, as well as in vegetative tissues of a small group of plants that are tolerant to severe water loss such as resurrection plants. Although vegetative desiccation tolerance is frequently found across the plant kingdom, desiccation tolerance is absent in gymnosperms (Hartung *et al.*, 1998; Bernacchia and Furini, 2004; Alpert, 2006). Desiccation tolerance is defined as the ability of an organism to dry to

equilibrium with dry air (50% relative humidity and 20 °C, corresponding water potential of approximately -100 MPa) down to 10% RWC or less, and to resume normal metabolism and growth upon rehydration (Bewley, 1979; Gaff, 1997; Alpert and Oliver, 2002; Proctor and Pence, 2002; Alpert, 2005; 2006; Wood, 2005; 2007; Proctor *et al.*, 2007; Wood and Jenks, 2007). The threshold of 10% water content may correspond to the point at which there is no longer enough water available for forming a monolayer around macromolecules, thus enzymatic reactions cease and in turn, metabolism (Billi and Potts, 2002; Alpert, 2005).

Desiccation tolerance is considered as a critical component in the evolution of green plants that allowed primitive aquatic plants to successfully colonize the land (Oliver *et al.*, 2000). As land plants evolved, although desiccation tolerance was retained in the reproductive structures like spores, pollens and seeds, the trait was lost in vegetative parts of the plants, as more efficient mechanisms that conserve water within the plant were developed to increase the growth rate and the competitive ability of the plant. As vascular plants diversified, vegetative desiccation tolerance is thought to have independently re-evolved several times, giving rise to the resurrection plants observed today (Oliver *et al.*, 2000; Wood and Jenks, 2007; Porembski, 2011; Tuba and Lichtenthaler, 2011; Gaff and Oliver, 2013). Approximately 1300 plant species (1000 pteridophytes and 300 angiosperms) have been reported of possessing vegetative desiccation tolerance (Porembski, 2011; Gechev *et al.*, 2012; Gaff and Oliver, 2013). Resurrection plants have to date been identified in 6 families of monocotyledonous (Anthericaceae, Bromeliaceae, Cyperaceae, Philydraceae, Poaceae, and Velloziceae) and 10 dicotyledonous (Brassicaceae, Cactaceae, Gesneriaceae, Lamiaceae, Linderniaceae, Myrothamnaceae, Ranunculaceae, Scrophulariaceae, Stylidiaceae and Tamaricaceae) angiosperms, which are mostly found in shallow soils or on rocky outcrops in central and southern Africa, Australia and South America, in areas that are subjected to lengthy periods of drought due to sporadic rainfall (Porembski and Barthlott, 2000; Bernacchia and Furini, 2004; Wood and Jenks, 2007; Phillips *et al.*, 2008; Tuba and Lichtenthaler, 2011; Gaff and Oliver, 2013). However, there are exceptions. For example, *Lindernia brevidens*, a close relative to *Craterostigma plantagineum*, was recently reported to exhibit vegetative desiccation tolerance, even though it is endemic to the Montane rainforests of Tanzania and Kenya, where it never experiences seasonal dry periods (Phillips *et al.*, 2008). This resurrection plant is thought to be a neoendemic species that has retained desiccation tolerance through genome stability, despite such tolerance being superfluous to the environmental conditions (Phillips *et al.*, 2008). Most desiccation tolerant angiosperms

tend to be small in size. Metabolic rates, biomass production and competitive abilities are compromised, because of the extra cost of specialist protection and repair mechanisms that are activated in response to desiccation. (Alpert, 2006; Toldi *et al.*, 2009; Porembski, 2011). Thus, although desiccation tolerance may facilitate the survival of plants during prolonged periods of severe water loss, there is a trade-off between the ability of surviving desiccation and growth.

1.2.1 Drought tolerance and desiccation tolerance

Drought by definition is the limitation of water over a prolonged period of time (Alpert, 2005), and may also be defined as any level of water availability that is low enough to reduce plant performance (Alpert and Oliver, 2002). Plants cope with water deficit, or dehydration either by avoiding it (drought avoidance) or tolerating it (drought tolerance). Drought avoidance may be achieved by the formation of seeds before drought conditions prevail, or by specialized morphological adaptations such as (1) development of a specialized leaf surface; (2) reduction of the leaf total surface area; (3) sunken stomata to decrease the rate of water loss via transpiration; (4) the development of water storage organs; or (5) the increase in root length and density to utilize water more efficiently (Ramanjulu and Bartels, 2002; Bernacchia and Furini, 2004). Drought tolerance appears to be a more complex trait, which is the result of the co-ordination of physiological and biochemical alterations at the cellular and molecular level. For example the accumulation of various protective gene products, molecules and compounds, coupled with an efficient free radical scavenging system. These allow plant to survive under various degrees of water loss (Ramanjulu and Bartels; 2002).

Desiccation tolerance is an extreme form of drought tolerance (Alpert and Oliver, 2002; Shao *et al.*, 2008). Drought stress is most often considered to be a moderate loss of water from tissues and cells in plants, which leads to stomatal closure and limitation of gas exchange. Whereas desiccation stress is referred to as a much more extensive loss of water due to prolonged drought, which can potentially lead to gross disruption of metabolism and cell structure and eventually to the cessation of enzyme catalyzing reactions, and finally, death (Wood, 2005; Shao *et al.*, 2008).

Most drought tolerant, but desiccation sensitive plants are usually able to endure a mild degree of water deficit (tissue water contents down to 80% RWC) by employing adaptations that permit metabolism to occur at low water potential until the “permanent wilting point” is

reached, where water potential has declined to such a degree that the irreversible damage sets in, and plant cannot recover upon rehydration (Iljin, 1957; Moore *et al.*, 2008). As drought persists, these plants die from desiccation, unlike desiccation tolerant plants that can survive desiccation and revive upon rehydration (Scott, 2000). Some drought tolerant variants can survive more extensive degree of dehydration, such as desiccation sensitive grasses *Eragrostis curvula*, *Eragrostis teff* and *Eragrostis capensis* which have been reported to survive water deficit at 45%, 50% and 65% RWC respectively (Balsamo *et al.*, 2006; Farrant, 2007).

1.2.2 Desiccation tolerance in orthodox seeds

Desiccation tolerance in seeds is acquired through a programmed phase of embryological development (Berjak *et al.*, 2007). In higher plants that produce desiccation tolerant seed (also known as orthodox seeds), seed development is divided into two major phases: morphogenesis and seed maturation (West and Harada, 1993; Park and Harada, 2008; Gutierrez *et al.*, 2007; Holdsworth *et al.*, 2008; Le *et al.*, 2010) (Fig. 1.1).

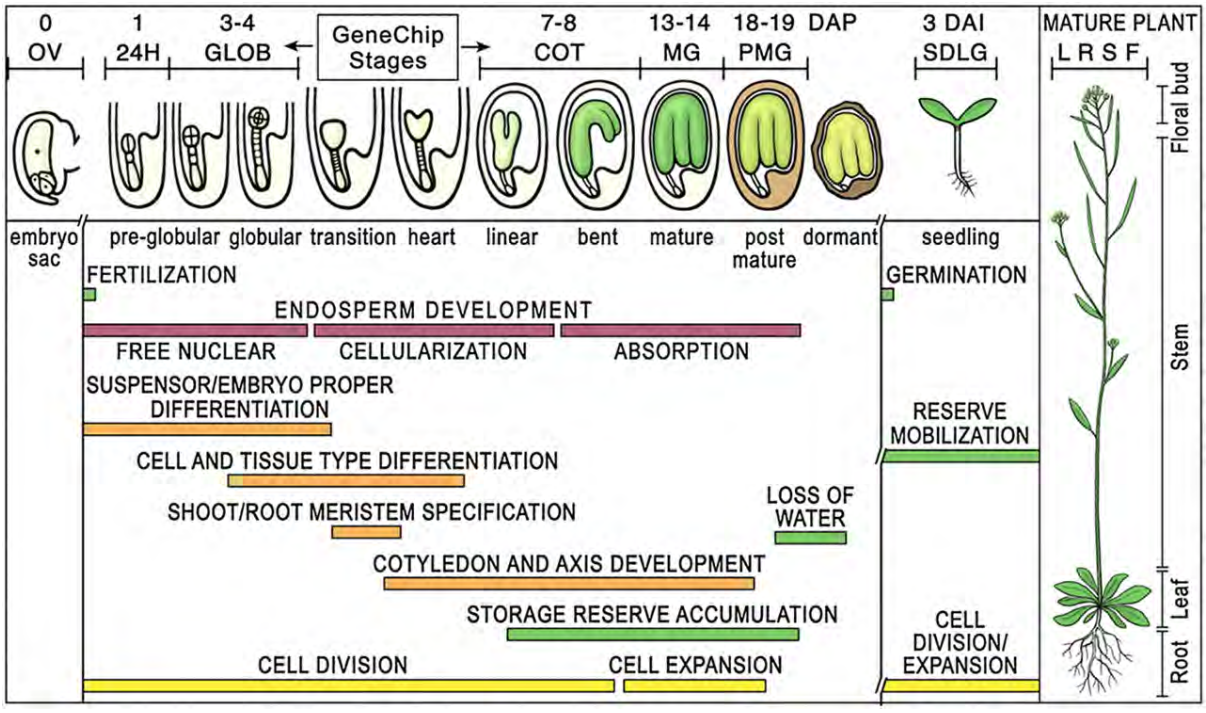


Figure 1.1. Schematic representation of *Arabidopsis thaliana* seed development and stages of the life cycle used for GeneChip analysis. Numbers correspond to days after pollination (DAP) or days after imbibition (DAI). OV, unfertilized ovule; 24H, 24-h postpollination seed; GLOB, globular-stage seed; COT, cotyledon-stage seed; MG, mature-green-stage seed; PMG, postmature-green-stage seed; SDLG, seedling; L, leaf; R, root; S, stem; F, floral buds (taken from Le *et al.*, 2010).

In *Arabidopsis thaliana*, morphogenesis involves embryo and endosperm development. Embryogenesis begins with double fertilization in which one sperm cell fuses with the egg cell and the other fuses with the central cell to form the zygote and endosperm mother cell respectively (Park and Harada, 2008). The zygote then undergoes cell division and develops into the embryo in precisely ordered events, ensuring the correct relative positioning of the various tissues and organs (i.e. meristems, cotyledons and hypocotyl) and the arrangement of cell types within each tissue of the embryo (Santos-Mendoza *et al.*, 2008). During the early maturation phase, cell division and growth of the embryo are halted (Raz *et al.*, 2001). Throughout the rest of the maturation phase, the embryo undergoes a period of cellular expansion and differentiation, dehydration, and filling the embryo sac with various synthesized storage compounds, concomitantly causing the reduction of the endosperm to one cell layer (Santos-Mendoza *et al.*, 2008). By the end of maturation phase, storage compounds have accumulated, water content has decreased, and desiccation tolerance and primary dormancy are established. This facilitates seed dispersal and the survival of seeds for a period in various environmental conditions (Gallardo *et al.*, 2008; Holdsworth *et al.*, 2008).

One of the keys to this success in survival of seeds is the reserves, such as storage proteins, lipids (often triacylglycerols) and carbohydrates (often starch). The critical level of reserves in vacuoles confers mechanical strength to whole cell during water deprivation (Kermode and Finch-Savage, 2002). These reserves are also used as an initial energy source in seedling establishment upon germination, in which desiccation tolerance is lost. Other protective proteins and molecules such as late embryogenesis abundant (LEA) proteins, antioxidants, and sucrose also accumulate and are stored in the mature seed. These proteins and carbohydrates have potential roles in protection against desiccation damage on seeds, both during desiccation and subsequent rehydration, and are to be discussed in details in section 1.3.

1.2.3 Vegetative desiccation tolerance

Unlike desiccation tolerance in orthodox seeds that is pre-programmed during embryo development, vegetative desiccation tolerance is thought to be stochastic, induced when a plant senses severe water deprivation in the environment (Berjak *et al.*, 2007). Mechanisms of vegetative desiccation tolerance vary in different plants. Desiccation tolerant plants can be divided into two classes: fully-desiccation tolerant group and modified-desiccation tolerant group based on the desiccation rate they can tolerate. Fully-desiccation tolerant plants can

tolerate rapid desiccation, i.e. desiccation within a few minutes, withstanding total loss of free protoplasmic water (Oliver *et al.*, 1998). These include the less complex, lower order species like algae, lichens, and mosses. The internal water content of these plants rapidly equilibrates to the water potential of the environment as they possess very few of morphological or physical adaptations for the retention of water. Mechanisms of desiccation tolerance in these plants are considered to be heavily based on the induced cellular repair during rehydration. The cellular protection mechanisms in these plants during desiccation are not induced, but are primitive and constitutive in general (Oliver *et al.*, 2000). However, a desiccation tolerant desert moss *Pterygoneurum lamellatum* has recently been shown to exhibit inducible protective mechanisms upon desiccation (Stark *et al.*, 2013).

The modified-desiccation tolerant plants constitute more complex, higher order vascular plants such as angiosperm resurrection plants. They can only tolerate desiccation at slower rate. These plants rely heavily on inducible morphological and physiological cellular protection systems during desiccation, rather than on cellular repair during rehydration (Toldi *et al.*, 2009). These desiccation tolerant plants can be subdivided into homoiochlorophyllous and poikilochlorophyllous types, based on their ability to maintain or dismantle the photosynthetic apparatus during desiccation.

In orthodox seed, ROS accumulates mainly from ongoing respiratory metabolism (Hendry, 1993; Bailly, 2004), whereas in vegetative tissues, excess excitation energy resulted from photosynthetic activity under water deficit is a critical additional source of ROS production (Smirnoff, 1993). Homoiochlorophyllous species maintain their photosynthetic apparatus and chlorophyll in a readily recoverable form during desiccation. They rely on various mechanisms to minimize light-chlorophyll interaction during the dry state such as leaf folding and/or shading of inner leaves or adaxial surfaces from light, as well as accumulation of sunscreen pigments. Pigments such as anthocyanin accumulate in leaf surfaces that are exposed to light in order to mask chlorophyll and to reflect light back (Sherwin and Farrant, 1998; Farrant, 2000; Farrant *et al.*, 2003; Moore *et al.*, 2007a; 2007b; Tuba and Lichtenthaler, 2011). In contrast, poikilochlorophyllous species dismantle their internal chloroplast structure and break chlorophyll down to shut down photosynthesis completely during desiccation in an effort to minimize photosynthetically associated ROS production. The photosynthetic components are resynthesized and reconstituted upon rehydration (Sherwin and Farrant, 1998; Farrant, 2000; Farrant *et al.*, 2003; Moore *et al.*, 2007; 2007b; Tuba and Lichtenthaler, 2011).

Although homoiochlorophyllous variants have the advantage of resuming photosynthetic activity immediately upon rehydration, they can only survive shorter periods (several days or weeks) of water drought compared to poikilochlorophyllous plants. Poikilocholophyllous desiccation tolerance appears to be restricted to the monocotyledonous desiccation tolerant plants (Gaff, 1989), and such strategy is thought to have evolved in habitats where the plants remain in the desiccated state for 6 to 11 months, where it is evidently more advantageous to dismantle the photosynthetic apparatus and reconstruct it after rehydration, rather than costing energy to maintain photosynthetic apparatus in the readily recoverable form through prolonged periods of desiccation (Tuba *et al.*, 1996; Toldi *et al.*, 2009; Tuba and Lichtenthaler, 2011). Homoiochlorophyllous plants in contrast, are found in habitats with more frequent wet and dry period alternations, where desiccated periods are shorter.

The cellular protection mechanisms induced during desiccation in vegetative tissues of desiccation tolerant plants appear very similar to those observed in orthodox seeds (Farrant, 2000; Vicré *et al.*, 2004; Berjak *et al.*, 2007) (Table 1.1). To prevent the mechanical damages due to cell shrinkage, some desiccation tolerant plants such as *Xerophyta humilis* and *Xerophyta viscosa* synthesize and store compatible solutes or osmolytes in the vacuole i.e. sucrose and proline in replacement of water to maintain turgor pressure and to provide mechanical support. The large vacuole is fragmented into numerous smaller vacuoles that fill the cytoplasm to maintain cell volume, in turn which prevents organelle compaction and plasmolysis that resulted from cell shrinkage (Farrant, 2000; Van der Willigen *et al.*, 2004). *Craterostigma wilmsii* however limits mechanical stresses by active and reversible wall folding (Vicré *et al.*, 2004). Some species have been reported to utilize both mechanisms in different tissues to combat mechanical stresses arise from dehydration. For example, in *Myrothamnus flabellifolia*, wall folding occurs in the epidermis, around seemingly less flexible stomata and gland cells, as well as in the immediately adjacent mesophyll cells. Whereas in the more centrally located mesophyll cells, mechanical support is due to vacuole filling (Moore *et al.*, 2007). In tolerant grass species *Eragrostis nindensis* and *Sporobolus stapfianus*, wall folding occurs in mesophyll cells and vacuole filling in bundle sheath cells, (Van der Willigen *et al.*, 2003; 2004; Farrant, 2007).

Similar to orthodox seeds, sucrose, as well as LEA proteins, antioxidants and other protectants have been reported to be universally accumulated in vegetative tissues of almost

all desiccation tolerant plants (Norwood *et al.*, 2000; 2003; Bartels and Salamini, 2001; Whittaker *et al.*, 2001; 2004). They too have potential roles in protection against desiccation damages in vegetative tissues of resurrection plants, and are to be discussed in details in section 1.3. Given the similarities of desiccation tolerance in orthodox seeds and that in the vegetative tissues of resurrection plants, it has been postulated that vegetative desiccation tolerance is possibly a reiteration of seed desiccation (Illing *et al.*, 2005; Cushman and Oliver, 2011).

1.3 Genes induced during drought stress and desiccation response, and seed development

Genes responsible for the synthesis of common classes of protectants are induced in desiccation sensitive plants during drought stress; in resurrection plants during desiccation; as well as in orthodox seeds during maturation.

1.3.1 Sugars

Under abiotic stress conditions, soluble sugars like sucrose, and raffinose family oligosaccharides such as galactinol and raffinose, have been found to accumulate in *A. thaliana*. Transcripts encoding galactinol synthase (GolS), an enzyme involved in the biosynthesis of raffinose family oligosaccharides have been shown to be induced during abiotic stress response (Taji *et al.*, 2002). Overexpression of GolS was shown to increase endogenous galactinol and raffinose levels and improved abiotic stress tolerance in *A. thaliana*, suggesting the osmoprotective function under abiotic stress conditions (Taji *et al.*, 2002; Yamada *et al.*, 2010). In addition to sugar synthases, transcripts encoding sugar transporters are also found induced under abiotic stress conditions, implicating the transport of sugars to specific tissues or organelles in *A. thaliana* under abiotic stress conditions (Maruyama *et al.*, 2004; Wormit *et al.*, 2006; Yamada *et al.*, 2010). Many studies including transgenic experiments have revealed strong correlations between sugar accumulation and drought tolerance (Ramanjulu *et al.,* 1994; Abd-El Baki *et al.,* 2000; Gilmour *et al.,* 2000; Streeter *et al.,* 2001).

Accumulation of soluble sugars is also observed in seeds during maturation, which are thought to play a role in the acquisition of desiccation tolerance (Kermode and Finch-Savage, 2002; Berjak *et al.*, 2007). In orthodox seeds, sucrose levels gradually accumulate at the end of the maturation phase (Baud *et al.*, 2002; Focks and Benning, 1998). Sucrose has been

suggested to have a signaling function during the transition from embryo morphogenesis to maturation in *A. thaliana*, because the decrease in the hexose to sucrose ratio was found to correlate with transition to the maturation phase (Weber *et al.*, 2005; Gutierrez *et al.*, 2007; Santos-Mendoza *et al.*, 2008).

Sucrose has been shown to accumulate in vegetative tissues of all desiccation tolerant plants that have been assayed, as well as in orthodox seeds during maturation (Farrant, 2007). The resurrection plant *C. plantagineum* has been shown to have high amount of 2-octulose in the hydrated state. During dehydration these sugars are rapidly converted to sucrose (Bianchi *et al*., 1991). This sugar conversion is coupled with up-regulation of sucrose synthase gene and sucrose phosphate synthase gene (Ingram *et al.*, 1997). Increasing sucrose synthesis and sucrose phosphate synthase activity is not only a dehydration response or developmental program found in desiccation tolerant plants such as *C. plantagineum* and in orthodox seeds respectively, but also in plants that are intolerant to desiccation such as spinach (Quick *et al.*, 1989; Zrenner and Stitt, 1991).

Trehalose is a sugar which has been shown to contribute cellular protection as water replacement molecule and membrane stabilizer during dehydration in animal systems (Crowe *et al.*, 1986; 1987; 1998; Kaushik and Bhat, 2003). Similarly, in plant systems, trehalose plays a role as a carbon source and stress protection compound in plants. In addition, trehalose and its precursor, trehalose-6-phosphate, have signaling functions, and are involved in the regulation of plant growth and development (Nunes *et al.*, 2013). Substantial amounts of trehalose were identified in two resurrection plants *M. flabellifolia* and *S. stapfianus* (Bianchi *et al.*, 1993; Drennan *et al.*, 1993; Albini *et al.*, 1994; Phillips *et al.,* 2002), and high levels of trehalose were found in spikemoss *Selaginella lepidophylla* during drying (Adams *et al.*, 1990; Iturriaga *et al.*, 2000). The high levels of trehalose have long been considered crucial in the desiccation tolerant mechanism of *S. lepidophylla*. However, very recently, a large-scale comparative metabolic analysis has reported that the desiccation sensitive *Selaginella moellendorffii* unexpectedly contains higher trehalose levels than *S. lepidophylla* upon drying (Pampurova and Van Dijck P, 2014). Trehalose was reported to be absent in maturing or mature orthodox seeds (Bewley and Black, 1994). However, its presence has recently been recorded in *A. thaliana* seeds (Fait *et al.*, 2006). Nevertheless, sucrose has widely been considered as an alternate to trehalose in plant systems (Berjak *et al.*, 2007).

The current hypothesis is that sugars accumulated during dehydration can act as compatible solutes that fill the vacuoles to maintain cell turgor, and prevent plasmolysis. Sugars can also act as protectors of intracellular macromolecules and contribute to the stabilization of membrane structures through formation of a biological glass that prevents the crystallization of cellular solutes, as well as by maintaining hydrogen bonds within and between the macromolecules (Crowe *et al.*, 1992; Leprince and Buitink, 2007; Oliver, 2007). Furthermore, sugars accumulated during drying can serve as additional carbon source for recovery during rehydration (Farrant, 2007).

To investigate the correlation between sucrose accumulation and desiccation tolerance, Illing *et al.* (2005) analyzed and compared changes in sucrose content in response to drying in vegetative tissues of several desiccation tolerant and desiccation sensitive species, as well as in orthodox seeds of several plant species. Results showed that this sugar does indeed increase to varying extents on dehydration in all resurrection angiosperms and orthodox seeds studied, but only in some desiccation sensitive plants. Analysis on sucrose accumulation in desiccation tolerant and desiccation sensitive *Eragrotis* grasses, *E. nindensis* (tolerant) and *E. curvula* and *E. teff* (sensitive), revealed that this sugar accumulates (unrelated to photosynthesis) in seeds of both tolerant *E. nindensis* and sensitive *E. teff*, but only in vegetative tissues of tolerant *E. nindensis* (Illing *et al.*, 2005).

1.3.2 Late embryogenesis abundant proteins (LEA)

LEA proteins were first identified in mature cotton (*Gossypium hirsutum*) seeds during late maturation stage (Galau *et al.*, 1986; 1993). As their name suggests, LEA proteins are produced in abundance during late embryo development, comprising up to 4% of cellular protein (Roberts *et al.*, 1993; Wise and Tunnacliffe, 2004). LEA proteins are low complexity, generally hydrophilic and intrinsically unstructured under fully hydrated conditions, but have been shown to acquire various degrees of second structures including alpha-helix and beta-sheet in the dry state (Hincha and Thalhammer, 2012). The nomenclature and classification of LEA proteins have been rather confusing in the literature. As more and more LEAs were being identified, different independent researchers simply reported and named their identified LEAs in their own ways. Their low complexity and unstructured nature in hydrated state have made it experimentally difficult to assign structure and determine potential function, for which they can be classified and named (Farrant, 2007). LEAs can be classified into different groups, and are initially grouped on the basis of their similarity to prototypical LEA proteins

from the cotton plant, and are named after particular cDNA clones, i.e. D7, D11, D19, D29, D34, D73, D95 and D113 (Dure, 1993; Dure *et al.*, 1989; Wise, 2003; Bies-Ethève *et al.*, 2008). Similar and additional proteins were discovered in other species thereafter, and these genes were further classified into several groups based on sequence similarities (Bray, 1993; 1994; Cuming, 1999): Group 1 (proteins similar to D19), group 2 (proteins similar to D11), group 3 (proteins similar to D7), group 4 (proteins similar to D113), group 5 (proteins similar to D29) and group 6 (proteins similar to D34) (Bies-Ethève *et al.*, 2008). As new LEA proteins were identified, they were assigned to one of these groups depending on sequence relatedness and the presence of the characteristic motifs (repeated amino acid sequence). However there were few cases that some newly identified LEA proteins could not be assigned to any group (Wise, 2003). Later, the categorization of LEA was further refined based on a novel computational approach, POPP (protein or oligonucleotide probability profile), using similarities of peptide composition, according to InterPro superfamilies and Pfam domains, rather than sequence similarities of proteins (Wise, 2003; Wise and Tunnacliffe, 2004; Berjak *et al.*, 2007) (Table 1.2).

Table 1.2. Nomenclature of LEA groups and their corresponding sequence motif, Pfam, and InetrPro descriptions and identifiers.

Dure, 1993	Bray, 1994	Wise, 2003	Galau *et al.*, 1989	Harada *et al.*, 1989	Illing *et al.*, 2005	Pfam	InterPro	Sequence motif
D19	Group 1	Class I	-	-	LEA-1	LEA-5, small hydrophilic plant seed protein; PF00477	IPR000389	GGQTRREQLGEEGYSQMGRK
D11	Group 2	Class II	-	-	LEA-2	Dehydrin; PF00257	IPR000167	Y motif (DEYGNP) S motif (S_n) K motif (EKKGIMDKIKEKLPG)
D7	Group 3	Class III	-	-	LEA-3	LEA_4; PF02987	IPR004238	TAQAAKEKAXE
D113	Group 4	Class II, III	-	-	LEA-4	LEA_1; PF03760	IPR005513	-
D29	Group 5	Class III	-	-	-	LEA_4; PF02987	IPR004238	-
D34	Group 6	Class IV	-	-	LEA-6	Seed maturation protein; PF04927	IPR007011	-
D73	-	Class V	Lea5	-	LEA-7	LEA_3; PF03242	IPR004926	-
D95	-	Class VI	Lea14	-	LEA-8	LEA_2; PF03168	IPR004864	-
-	-	Class III	-	Lea76	-	LEA_4; PF02987	IPR004238	-
-	-	-	-	-	LEA-10	AWPM-19-like; PF05512	IPR008390	-

Modified from Tunnacliffe and Wise (2007) and Fisher (2008).

LEA genes have been shown to be induced in orthodox seeds during the maturation phase; in vegetative tissues of desiccation tolerant plants during desiccation; as well as in many desiccation sensitive species including *A. thaliana* and rice in response to abiotic stress conditions. In *A. thaliana*, microarray analysis of 22476 genes revealed several LEA mRNA transcripts present in the dry seeds from groups 1, 2, 3, 4, 6, 7 and 8 (Illing *et al.*, 2005 nomenclature) (Nakabayashi *et al.*, 2005). LEA mRNAs from groups 1, 2, 4 and 6 were also shown to be preferentially expressed during seed maturation, and with decreased level during germination in a microarray analysis of *Brassica oleracea* seeds (Soeda *et al.*, 2005). Fourteen LEAs were identified as being up-regulated in *A. thaliana* seedlings that were exposed to either drought, cold or high salinity stress, which comprised from groups 2, 3, 7 and 8 (Seki *et al.* 2001; 2002). Transcriptome studies in four resurrection plants, *C. plantagineum, X. humilis, X. viscosa and S. stapfianus* have identified LEA transcripts from all groups (Table 1.3) as being up-regulated in the vegetative tissues during desiccation (Piatkowski *et al.*, 1990; Blomstedt *et al.*, 1998; Ndima *et al.*, 2001; Mundree and Farrant, 2002; Collett *et al.*, 2004; Farrant, 2007) (Table. 1.3). This may not simply imply that the transcription rate of these LEA genes was increased during dehydration in these plants. The terms “up-regulated” and “down-regulated” are frequently used in the similar differential expression studies, and they usually refer to the changes in the steady-state levels of mRNA transcripts, which can be the consequence in the change of rate of transcription or turn-over of mRNA transcripts.

Table 1.3. Summary of desiccation induced LEA genes reported in vegetative tissues of resurrection plants.

Species	LEA Superfamily (Illing *et al.*, 2005)	Genebank Accession number	Description	Reference
C. plantagineum	LEA-3	P23283	Desiccation-related protein [*C. plantagineum*]	Piatkowski *et al.*, 1990
	LEA-8	P22241	Desiccation-related protein [*C. plantagineum*]	
	LEA-2	P22238	Desiccation-related protein [*C. plantagineum*]	
	LEA-2	S43775	Desiccation-related protein [*C. plantagineum*]	
X. humilis	LEA-2	CK906385	none	Collett *et al.*, 2004; Illing *et al.*, 2005
	LEA-2	CK988413	44 kDa dehydrin-like protein [*C. sericea*]	
	LEA-2	CK906432	Embryogenic abundant protein (radish)	
	LEA-2	CK906386	Dehydrin-like protein [*M. sativa*]	
	LEA-3	CK906406	LEA-like protein [*L. longiflorum*]	
	LEA-3	CK906427	LEA-like [*A. thaliana*]	
	LEA-3	CK906404	LEA protein 76 (rape)	
	LEA-3	CK906402	LEA 1 protein [*T. aestivum*]	
	LEA-3	CK906398	LEA protein [*B. inermis*]	
	LEA-4	CK906399	Putative seed maturation protein [*O. sativa*]	
	LEA-6	CK906408	Seed maturation protein PM26 [*G. max*]	
	LEA-7	CK906401	LEA homolog (tomato)	
	LEA-8	CK906400	LEA protein Lea 14-A (upland cotton)	

Table 1.3. (continued)

Species	LEA Superfamily (Illing *et al.*, 2005)	Genebank Accession number	Description	Reference
X. humilis (continued)	LEA-10	CK906403	LEA protein with hydrophobic domain [*G. max*]	
	LEA-10	CK906405	Hydrophobic LEA-like protein [*O. sativa*]	Collett *et al.*, 2004; Illing *et al.*, 2005
	LEA-10	CK906407	Putative plasmamembrane associated protein [*O. sativa*]	
X. viscosa	LEA-2	AAP22171	*X. viscosa*	Mundree and Farrant, 2002; Ndima *et al.*, 2001
	LEA-2	NA		
S. stapfianus	LEA-2	EMBL:Y10778	Dehydrin	Blomstedt *et al.*, 1998
	LEA-3	Y10779	LEA-like protein (Wheat)	

Modified from Farrant (2007).

Group 1 LEA proteins are characterized by a 20-amino acid conserved domain that occurs between one and four times (Cuming, 1999). They are highly hydrophilic that possess a high potential for hydration, and are predicted to have structural flexibility indicated by random coil configuration (McCubbin *et al.*, 1985; Baker *et al.*, 1988). Group 1 LEA proteins have therefore been hypothesized to play roles in providing a hydration shell around intracellular structures such as membranes and macromolecules, in orthodox seeds and vegetative tissues of resurrection plants during desiccation (Berjak *et al.*, 2007). Unlike other LEAs, Group 1 LEA genes are only found to be significantly expressed during seed development, but not during abiotic stresses in vegetative tissues of *A. thaliana*, suggesting that these LEA proteins may thus be uniquely associated with vegetative desiccation tolerance (Illing *et al.*, 2005). Group 2 LEA proteins are also referred to as dehydrins, and have been shown to have detergent and chaperone-like properties that stabilize membranes and protein cellular components during drying (Close, 1996; Bartels *et al.*, 2007). They are characterized by 1 to 11 copies of a lysine-rich K domain at carboxyl terminus, as well as a S domain which consists of a tract of serine residues in some dehydrins, and a consensus Y domain near the

amino terminus in most dehydrins (Close, 1996; 1997). The K segment has the tendency of forming amphipathic α helices, suggesting their possible stabilizing role by interacting with hydrophobic domains of other proteins that could be exposed with increasing dehydration, which in turn minimizes the incidents of inappropriate inert-molecular hydrophobic associations during desiccation (Close, 1997; Cuming, 1999). Additionally, group 2 LEAs possess highly polar and unstructured repeating π segments, which may interact with, and stabilize various intracellular constituents during desiccation (Close, 1997; Cuming, 1999; Berjak *et al.*, 2007). Group 2 LEA proteins have been shown to be up-regulated in response of abiotic stresses, as well as seed maturation and desiccation in resurrection plants. Group 3 LEA proteins are characterized by their 11-mer repeats, which are proposed to have potential of forming α helices. The dimerization of two helices via hydrophobic interactions may result in a structure with highly charged surfaces, which can sequester ions during dehydration to prevent irreversible damages of intracellular structures due to increased ionic stress during dehydration (Dure, 1993). The carboxyl-terminal domains of group 4 LEA proteins are suggested to show functional similarity to the π segment of group 2 LEAs due to their high potential for hydration. In addition, they are predicted to form amphipathic α helices similar to groups 2 and 3 LEAs (Cuming, 1999). Despite the differences from the other LEA groups in their heat insolubility and relatively higher portion of hydrophobic residues, some group 5 LEA proteins have seen found to be composed largely of 11-mer repeats similar to group 3 LEAs (Cuming, 1999; Dure, 1993). Based on these observations, Wise and Tunnacliffe (2004) in their attempt to refine categorization of LEAs, proposed to incorporate group 4 and group 5 LEA members into group 2 and group 3 respectively.

Plants activate different groups of LEAs simultaneously during water deficit. It is possible that different groups of LEAs are specifically targeted to different organelles or cellular structures, where they serve to protect or stabilize proteins, nucleic acids and membranes under dehydration locally. Alternatively, it may also suggest that the formation of an interacting network of different LEAs may be necessary for the protection and stabilization of macromolecules (Illing *et al.*, 2005). Hundertmark and Hincha (2008) studied subcellular distribution and localization of all 51 LEA proteins identified in *A. thaliana*. LEA4 proteins (LEA-3 group in Illing *et al.*, 2005) are computationally predicted to be localized in all cellular compartments; the LEA3 proteins (LEA-7 group in Illing *et al.*, 2005) are exclusively targeted to chloroplasts and mitochondria. The AtM proteins (LEA-9 group in Illing *et al.*, 2005) are predicted to enter the secretory pathway; and one member in the SMP (seed

maturation protein) group (LEA-6 group in Illing *et al.*, 2005), At5g53270, is more likely to be targeted to chloroplasts. Subcellular localization of several different groups of LEA proteins has also been studied on experiment base. Cold regulated COR15a (At2g42530), a LEA4 protein (LEA-3 group in Illing *et al.*, 2005) has been shown to localize in the chloroplast stroma (Lin and Thomashow, 1992). The SMP group protein RAB28 (At3g22490) (LEA-6 group in Illing *et al.*, 2005) has been shown to localize in the nucleus (Borrell *et al.*, 2002). These results suggest that LEA proteins can be present in all subcellular compartments. Olvera-Carrillo *et al.* (2010) analyzed the transcript and protein accumulation pattern of the 3 members of group 4 LEA proteins: LEA4-1, LEA4-2 and LEA4-5 in *A. thaliana* during the last stages of seed development where desiccation tolerance was required, as well as in vegetative tissues upon water deficit. They have found that LEA4-5 was differentially regulated as compared to LEA4-1 and LEA4-2, under normal developmental stages and upon osmotic and salt stress treatments. Together with phenotypic results obtained from different LEA4 mutant plants, as well as transgenic plants over-expressing different LEA4 proteins, they have reported that the functional redundancy among different groups or within a particular group of LEAs is likely to be minimal. However, whether different groups of LEAs have different functions in different compartments and what these functions are remains to be determined.

The exact roles and functions of LEA proteins still remain unclear. Many potential functions of LEA proteins have been predicted based on their rich hydrophilic amino acid content and their thermostability. LEA proteins are proposed to act as water replacement molecules or hydration buffers; as ion sequesters; as chaperonins; in prevention of protein and membrane aggregation; and in facilitating glass formation with sugars (Bray, 1997; Vicré *et al.*, 2004a; Berjak, 2006; Mtwisha *et al.*, 2006; Farrant, 2007). Two dehydrins in *A. thaliana*, ERD10 and ERD14, have been shown to act as chaperones *in vitro*, which were able to prevent the heat-induced aggregation and/or inactivation of various substrates, such as lysozyme, alcohol dehydrogenase, firefly luciferase, and citrate synthase (Kovacs *et al.*, 2008). MtPM25, a group 5 LEA found in seeds of *Medicago truncatula*, not only has been shown to protect membranes and to prevent aggregation of proteins against heating, freezing or drying, it was also able to dissolve aggregates resulted from stress conditions in a non-specific manner (Boucher *et al.*, 2010). The involvement of LEAs in increasing tolerance to various abiotic stresses has also been demonstrated in various transgenic studies. Constitutive expression of the barley HVA1 gene, encoding group 3 LEA protein, has been reported to enhance

tolerance of water stress and salt stress in transgenic rice (Xu *et al.*, 1996; Rohila *et al.*, 2002). Introduction of the same gene into wheat resulted in better growth and water usage under water stress conditions (Sivamani *et al.*, 2000).

Due to their close association in seed development and in conferring tolerance during various abiotic stresses, they have been utilized in comparative studies to evaluate the hypothesis that vegetative desiccation is possibly derived from an activation of the seed desiccation program. Illing *et al.* (2005) compared the expression of 35 *A. thaliana* LEA genes during seed development, during various abiotic stresses in vegetative tissues, as well as the expression of 16 LEA genes identified in leaves of the resurrection plant *X. humilis* (Collett *et al.*, 2004) during desiccation-rehydration cycle. The study identified 13 *A. thaliana* LEA genes which are exclusively expressed in seeds, which mainly encode for LEA-1, -6 and -9 proteins. In addition with the presence and induction of a seed-specific LEA-6 orthologue observed in *X. humilis* leaves during desiccation, suggested that desiccation tolerance in vegetative tissues might be a seed-derived trait. Fisher (2008) analyzed ancestral expression of LEA genes from desiccation sensitive Arabidopsis, Physcomitrella, and desiccation tolerant *Tortula ruralis*, *C. plantagineum* and *X. humilis* by Bayesian reconstruction to evaluate the origin of vegetative desiccation tolerance in plants. Results showed that none of the ancestral LEA nodes were estimated to have expression pattern that was exclusively stress related. All the significant reconstructions included seed and/pollen expression either exclusively or in combination with stress expression. The result further supports the hypothesis that desiccation tolerance in vegetative tissues being a seed-derived trait, rather than the derivation from co-option of genes exclusively related to stress tolerance.

1.3.3 Small heat shock proteins (sHSP)

Small heat shock proteins genes have been shown to be induced in *A. thaliana* seedlings when subjected to drought, cold and high salinity stresses (Seki *et al.*, 2001; 2002), in the maturing orthodox seeds of many plant species (Wehmeyer *et al.*, 1996; Kermode and Finch-Savage, 2002; Berjak *et al.*, 2007), as well as in the desiccated leaves of resurrection plant *X. humilis* during desiccation (Collett *et al.*, 2004). Under hydrated conditions, sHSPs are usually undetectable in vegetative tissues, but can be induced by environmental stress stimuli in addition to high temperature (Waters, 2013). In contrast to undetectable levels during hydrated conditions observed in many plants, sHSPs are however constitutively expressed in the leaves of resurrection plant *C. plantagineum*, and accumulate to higher levels during

desiccation (Alamillo *et al.*, 1995). Because the constitutive expression of sHSPs has only been previously observed in orthodox seeds, such observation has once again strengthened the hypothesis of vegetative desiccation tolerance observed in desiccation tolerant plants being a seed-derived trait (Oliver, 2007).

sHSPs have been categorized into six classes based on DNA sequence similarity, immunological cross-reactivity, and intracellular localization (Waters *et al.*, 1996; Scharf *et al.*, 2001). In plants, classes CI, CII, and CIII sHSPs are localized in the cytosol or in the nucleus, and classes P, ER and M sHSPs are localized in the plastids, endoplasmic reticulumn and mitochondria respectively (Scharf *et al.*, 2001; Vierling, 1991; Helm *et al.*, 1993; 1995; Lenne and Douce, 1994; Lenne *et al.*, 1995; LaFayette *et al.*, 1996). The complex expression patterns of sHSPs and their unusual abundance and diversity observed in plants may reflect their importance in conferring tolerance to the environmental stresses (Sun *et al.*, 2002; Waters, 2013).

The mechanisms of cellular protection by sHSPs during water deficit stress are still largely unknown. Nevertheless, they have been proposed to act as molecular chaperones that bind to, and stabilize another protein and by controlled binding and release, facilitate its correct folding, oligomeric assembly, transport to a particular subcellular compartment, or disposal by degradation (Sun *et al.*, 2002). Furthermore, they possess ability to recognize and bind unfolded proteins, to minimize the incidents of possible irreversible protein aggregations during desiccation state (Sun *et al.*, 2002; Buitink *et al.*, 2002; Berjak *et al.*, 2007; Waters, 2013). Despite the observation of sHSPs being abundantly expressed in recalcitrant *Castanea sativa* seeds that are intolerant to desiccation (Collada *et al.*, 1997), the general protective roles of some sHSPs during water deficit have been demonstrated by several transgenic studies with use of sHSP genes (Batels and Sunkar; 2005). For instance, in transgenic *A. thaliana* plants overexpressing AtHSPl7.6A, a gene encodes for a class CI sHSP, tolerance to drought and high salinity is enhanced (Sun *et al.*, 2001).

1.3.4 Antioxidants

In plant vegetative tissues and orthodox seeds, reactive oxygen species (ROS) and free radicals such as singlet oxygen, hydroxyl radicals, hydrogen peroxide and superoxide anions are produced during water loss, which cause various cellular damages in plants (Smirnoff, 1998; Apel and Hirt, 2004; Oliver, 2007; Dizdaroglu, 1994; Leprince *et al.*, 2000; Dean *et al.*,

1993; Walters *et al.*, 2002). However, they may also act as signals, not only induce ROS scavengers, but also in activation of downstream cascades via Ca^{2+} (Price *et al.*, 1994; Loiacono and De Tullio, 2012). Sufficient defence mechanisms to detoxify excesses of ROS while maintaining their levels at minimal as required for signaling during drying, appear to be essential in survival during water deficit conditions in vegetative plants, as well as in matured seeds (Bartels *et al.*, 2007). The importance of antioxidant systems has been demonstrated in the resurrection shrub *M. flabellifolia*, whose ability to recover from desiccation is directly correlated to the state of its antioxidant defence system (Kranner *et al.*, 2002). The plant is able to recover after four months of desiccation, but not after eight months, once the antioxidants have been depleted.

ROS detoxification mechanisms in plants consist of non-enzymatic and enzymatic mechanisms. Key non-enzymatic antioxidants include ascorbate (vitamin C), glutathione, flavanoids, caratenoids and polyphenols. Major enzymatic mechanisms include superoxide dismutase (SOD), peroxidases, and catalase (CAT) (Mittler, 2002; Apel and Hirt, 2004; Bartels and Sunkar, 2005; Ahmad *et al.*, 2010). SOD converts superoxide to hydrogen peroxide, which is then detoxified by ascorbate peroxidase (APX) and CAT (Asada and Takahashi, 1987). Both SOD and CAT enzymes exist as multiple isozymes in the chloroplast and cytosol (Asada, 1994). In addition to these major free radical scavenging enzymes that neutralize primary ROS produced during dehydration, several detoxification enzymes are also involved in quenching of secondary products that are resulted from interaction of primary ROS and biomolecules. Aldehyde dehydrogenases (ALDH) are crucial enzymes in detoxification of highly reactive aldehydes resulted from free radical-mediated lipid peroxidation, to less toxic carboxylic acid forms. On the other hand, aldose/aldehyde reductases are responsible in reducing a wide range of aldehydes and ketones to alcohols. Aldehyde dehydrogenase genes have been identified in *A. thaliana* and resurrection plant *C. plantagineum*, induced by diverse abiotic stressors that induce oxidative stress (Sunkar *et al.*, 2003; Kirch *et al.*, 2001). An aldose reductase gene has been identified to be up-regulated in resurrection plant *X. viscosa* during water deficit (Mundree *et al.*, 2000). Peroxiredoxins are involved in the breakdown of cellular-toxic peroxides to the corresponding alcohols (Dietz *et al.*, 2002), they have been shown to protect DNA, membranes and certain enzymes against damage by ROS (Haslekås *al.*, 1998). Peroxiredoxin genes have been identified in *A. thaliana* seedling during drought and cold stress (Seki *et al.*, 2001), and in *X. viscosa* during desiccation (Mowla *et al.*, 2002). Increased resistance to environmental stresses is often

correlated with an efficient antioxidative system (Smirnoff, 1998; Kranner *et al.*, 2002). Overproductions of SOD, APX or CAT have been shown to improve oxidative stress tolerance in transgenic plants (Allen, 1995; Roxas *et al.,* 1997; Sunkar *et al.*, 2003; Kotchoni *et al.*, 2006).

In a comparative study, Illing *et al.* (2005) compared the expression of 68 *A. thaliana* antioxidant encoding genes during seed development, and during abiotic stresses in vegetative tissues. Results revealed that approximately 72% of the *A. thaliana* antioxidant genes were expressed at high levels in control stress-free plants. These include several APXs, SODs, GPs, and were referred to as housekeeping antioxidants as they are constitutively expressed at high levels under normal conditions, and responsive to most abiotic stresses. Enzyme activity assays of these housekeeping antioxidants during desiccation in desiccation tolerant grass *E. nindensis*, desiccation sensitive *E. curvula* and *E. teff*, revealed that although APX, SOD and GP may be housekeeping antioxidants constitutively expressed to protect desiccation tolerant and sensitive plants from oxidative damages arise from most abiotic stresses, their activities at lower RWC levels are only maintained in desiccation tolerant *E. nindensis* (Illing *et al.*, 2005). Furthermore, four genes were identified as seed specific antioxidants in this comparative study. Among them, 1-cys peroxiredoxin (1-cys-PrxR) has been shown to be abundantly expressed in vegetative tissues of desiccation tolerant *T. ruralis* (Oliver, 1996), *X. viscose* (Mowla *et al.*, 2002), and *X. humilis* (Collett *et al.*, 2004). The observation of seed-specific antioxidant gene from desiccation sensitive plants expressed at high levels in vegetative tissues of desiccation tolerant plants during desiccation further supports the seed-derived evolutionary trait of vegetative desiccation tolerance.

1.4 Regulatory networks during drought stress and desiccation response, and seed development

There is a clear overlap in the classes of protection proteins induced in seed maturation and the abiotic stress response, as well as the cellular signals that activate these programmes. Abscisic acid (ABA) is recognized as the major plant hormone involved both in integrating environmental changes in water availability with adaptive responses as well as during seed maturation in plants (Koornneef *et al.*, 1998; Bartels *et al.*, 2007; Yamaguchi *et al.*, 2007; Grene *et al.*, 2011; Nakashima and Yamaguchi-Shinozaki, 2013). ABA levels rapidly increase following water deficit stress as well as during seed maturation, resulting in the activation of genes involved in conferring protection and tolerance in plants against damages

caused by water-deficit. In addition, exogenous ABA has also been shown to mimic the induction of genes observed during stress in many plants and seeds (Bartels and Sunkar, 2005). Although ABA signaling is important in both the seed maturation and abiotic stress responses in *A. thaliana*, these responses are activated by different sets of transcription factors in association with ABA, such as ABA intensive 5 (ABI5) and ABA responsive element binding protein 1 (AREB1) respectively (Nakashima and Yamaguchi-Shinozaki, 2013).

1.4.1 Regulatory networks during seed development

Investigations on seed development and germination in *A. thaliana* at the genetic and physiological levels has identified many of the key regulators that are important for seed maturation, dormancy induction, maintenance and the completion of germination, including light, temperature, transcription factors and plant hormones (Santos-Mendoza *et al.*, 2008) (Fig. 1.2).

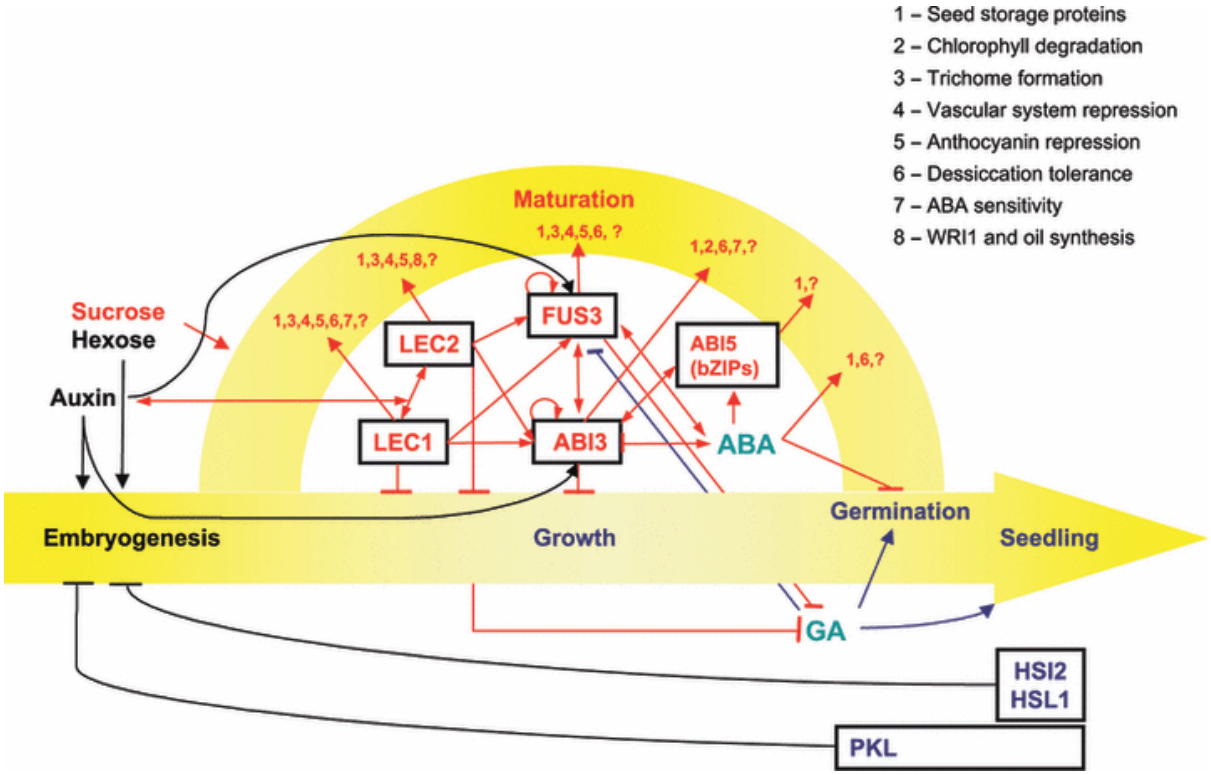


Figure 1.2. Proposed model of genetic and molecular interactions in the regulatory network involved in the control of seed development and maturation in *A. thaliana.* Arrows and T bars indicate positive and negative effects, respectively. The factors that induce and/or maintain seed maturation are shown in red. The factors that promote cell growth and differentiation are shown in blue. The numbers indicate the various targets of the regulators given in the key on the right. (Taken from Santos-Mendoza *et al.*, 2008).

In *A. thaliana*, seed maturation and dormancy induction are controlled by at least 5 major transcriptional regulators which were originally identified in mutant screens: leafy cotyledon 1 (LEC1) and LEC2, FUSCA3 (FUS3), ABA intensive 3 (ABI3) and ABI5, (Kagaya *et al.*, 2005; Gutierrez *et al.*, 2007). LEC1 gene encodes a HAP3 subunit of the CCAAT-binding transcription factor (CBF) required for normal development during early and late phases of embryogenesis (Lotan *et al*., 1998), acting as a master regulator of the LEC2, FUS3, ABI3 and ABI5 transcription factor cascade (Table 1.4). LEC2, FUS3 and ABI3 all encode transcription factors containing the conserved B3-binding domain. They act downstream of LEC1 to activate genes important for the maturation phase of seed development. Enhancers recognized by these transcription factors contain RY motif in their promoters (Giraudat *et al*., 1992; Luerssen *et al*., 1998; Stone *et al*., 2001; Braybrook *et al*., 2006) (Table 1.4). The phenotype of mutants of all four genes includes decreased dormancy at maturation and reduced expression of seed storage proteins (Raz *et al*., 2001; Gutierrez *et al*., 2007).

Table 1.4. Major transcription factors and the interacting *cis*-acting regulatory elements involved in abiotic stress response and seed development identified in *A. thaliana*.

Transcription factor family	*cis*-acting regulatory element	
	Name	Core sequence
AP2/ERF (ABI4)	CE1	CACC(G)
B3 (LEC2, FUS3, ABI3)	RY	CATGCA
bHLH	E-box	CANNTG
bZIP (AREB or ABF, ABI5)	ABRE	(C/T)ACGTG(G/T)C
bZIP (HY5)	LTRE	CCGAC
CBF	CRT	TGGCCGAC
DREB	DRE	TACCGACAT
DST	DBS	TGCTANNATTG
ERF	GCC box	AGCCGCC
HD-ZIP (I, II)		CAATNATTG
HD-ZIP (III, IV)		GTAAT(G/C)ATTAC or TAAATG(C/T)A
HSF	HSE	AGAAnnTTCT
MYB	MYBRS	TGGTTAG
MYC	MYCRS	CACATG
NAC	NACRS	TCNNNNNNNACACGCATGT
NF-YB (LEC1)	CAAT-box	CCAAT
WRKY	G-box	TTGACC/T
ZF-HD	rps1 site 1-like sequence	CACTAAATTGTCAC

Modified from Lata *et al.*, 2011 and Liu *et al.*, 2013

Seed storage protein (SSP) genes such as 2S albumins and 12S globulins, have been identified to be expressed at early and mid-maturation phases during seed development (Hughes & Galau, 1989; Parcy *et al*., 1994). ABA-responsive element (ABRE) and RY element have been identified in the promoter region of these SSP genes (Ellerstrom *et al*., 1996; Ezcurra *et al*., 1999; Ezcurra *et al*., 2000). ABI5, as well as other basic leucine zipper (bZIP) domain containing transcription factors such as AtbZIP10 and AtbZIP25 are involved in the regulation of SSP genes that directly bind to the ABRE. ABI5 proteins have been shown to binds to ABRE that are present in the promoters of several LEA genes, controlling theirs expression in *A. thaliana* seeds (Lopez-Molina and Chua, 2000; Finkelstein and Lynch, 2000; Carles *et al.,* 2002). Furthermore, global gene expression analysis during seed maturation revealed an overrepresentation of ABREs in genes with high expression levels in mature seeds (Nakabayashi *et al*., 2005; Cadman *et al.*, 2006), which supports and emphasizes the involvement of ABA in the acquisition of dormancy and desiccation tolerance in mature seeds.

The transcription factors encoded by the other 4 major regulators of seed maturation, ABI3, FUS3, LEC1 and LEC2 do not interact with the ABREs, they regulate the SSP genes by recognizing and binding to the additional RY elements or CAAT-box present in the promoter regions of these genes (Ezcurra *et al*., 1999; Reidt *et al*., 2000; Monke *et al*., 2004). In combination with closely associated ABREs, this module acts as an enhancer of seed-specific transcription (Dickinson *et al*., 1988; Suzuki *et al*., 1997).

Lastly, additional transcription factors belonging to different protein families which are up-regulated or repressed during seed maturation in *A. thaliana*, have been identified in several microarray studies (Girke *et al.*, 2000; Ruuska *et al.*, 2002; Nakabayashi *et al*., 2005; Cadman *et al.*, 2006; Le *et al.*, 2010). Le *et al.*, (2010) identified 48 transcription factor genes that were specifically involved in *A. thaliana* seed development, as well as transcription factors that were specifically induced in each of the seed development stages. These include 4 transcription factors up-regulated only during the mature green (MG) stage, and 19 transcription factors specifically induced during the post mature green (PMG) stage (Fig. 1.1). However, the exact role and function of most of these regulatory genes in controlling seed development still remain unknown, because mutations in 24 of these seed specific transcription factors analyzed did not result in seed-lethal phenotype or detectably alter seed development (Le *et al.*, 2010).

1.4.2 Regulatory networks during drought stress in desiccation sensitive plants

Stress responses triggered by dehydration and other environmental stress stimuli have been shown to be elicited via a network of signals that include both ABA-dependent and ABA-independent pathways in *A. thaliana* (Yamaguchi-Shinozaki and Shinozaki, 2005; 2006; Lata *et al.*, 2011) (Fig. 1.3).

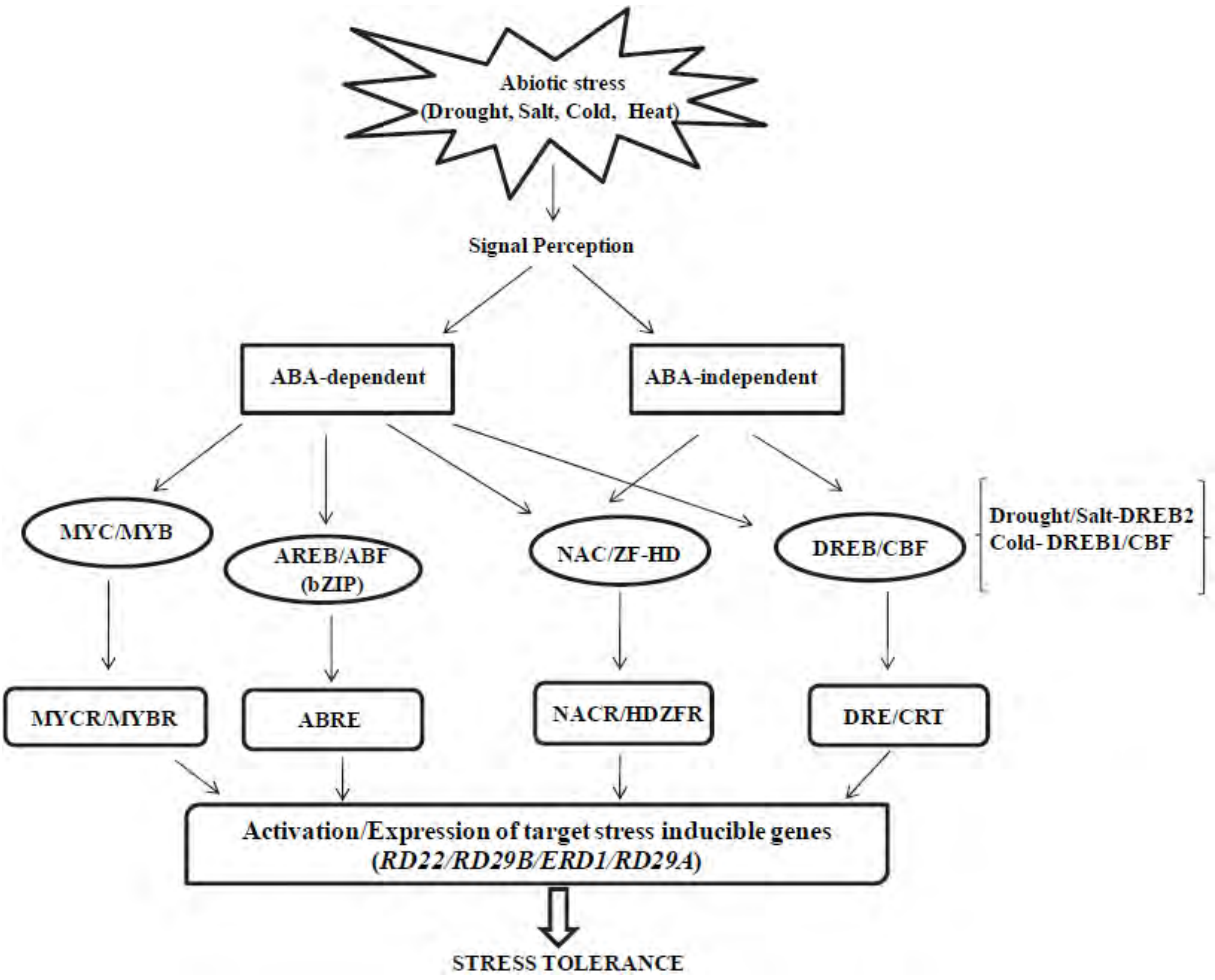


Figure 1.3. A schematic representation of transcriptional regulatory networks of *cis*-acting elements and transcription factors involved in abiotic-stress-responses in *A. thaliana*. Transcription factors are shown in ellipses; *cis*-acting elements are shown in boxes; and target stress inducible genes are shown in long rectangular box at the bottom (taken from Lata *et al.*, 2011).

Through the promoter analysis on the stress responsive genes, CREs involved in ABA-dependent and ABA-independent pathways have been identified. Among the CREs identified, ABREs have been found in the promoter regions of many genes, but not all, in ABA-dependent pathways (Guiltinan *et al.,* 1990; Mundy *et al.,* 1990; Ingram and Bartels, 1996; Busk and Pages, 1998; Ciarmiello *et al.*, 2011). The core CACGTG motif of ABREs is also

known as G-box motif, which functions in the regulation of plant genes stimulated by a variety of environmental signals (Marcotte *et al.*, 1989; Shen *et al.*, 1993; Bartels and Sunkar, 2005). The dehydration responsive element (DRE) (Table 1.4) is an essential CBE involved in the ABA-independent response to dehydration in *A. thaliana* (Yamaguchi-Shinozaki and Shinozaki, 1992; 1994; 2006). DRE motifs are also found in the promoter regions of many cold inducible genes (Thomashow, 1999; Shinozaki and Yamaguchi-Shinozaki, 2000). Similar CRE, C-repeat (CRT) and low temperature responsive element (LTRE) (Table 1.4) , have been identified, which both possess the CCGAC motif of the core DRE sequence, and which regulate cold inducible promoters (Baker *et al.*, 1994; Jiang *et al.*, 1996; Stockinger *et al.*, 1997; Thomashow, 1999; Catalá *et al.*, 2011).

A number of stress-inducible transcription factors have been identified which bind to these CREs. They include members of several transcription factor families such as the drought responsive element binding proteins (DREB), ethylene responsive element binding factors (ERF), zinc finger protein, WRKY family protein, MYB protein, basic helix-loop-helix (bHLH) protein, basic leucine zipper (bZIP) transcription factor, NAC family protein, and zinc finger homeodomain (ZF-HD) transcription factor subfamilies (Seki *et al.*, 2002; Sreenivasulu *et al.*, 2007). In addition, the heat shock transcription factors (HSFs) are also among the transcription factors found induced upon abiotic stress. They are the major regulator during the heat stress response, regulating the expression of heat shock proteins (HSPs), which are critical in the protection against heat damage and many other important biological processes (Guo *et al.*, 2008). These transcription factors regulate various stress-inducible genes either cooperatively or separately.

The promoter of a drought, high-salinity and cold inducible gene RD29A/COR78/LTI78 was first shown to contain both ABRE as well as DRE/CRT elements (Stockinger, 1997). Different families of transcription factors have been identified which bind to these CREs in response to stress in *A. thaliana* (Chen *et al.*, 2002; Seki *et al.*, 2004). DREB1/CBF and DREB2 are AP2/EREBP family transcription factors which have been shown to bind to the ABA-independent DRE/CRT element of up-regulated genes identified during drought and cold stress (Yamaguchi-Shinozaki and Shinozaki, 2005; Shinozaki and Yamaguchi-Shinozaki, 2007). DREB1/CBF genes are rapidly induced specifically by cold stress (Seki *et al.*, 2001; 2002; Fowler and Thomashow, 2002; Maruyama *et al.*, 2004; Vogel *et al.*, 2005). The DREB2 transcription factor is induced by drought is likely to activate genes involved in the

drought stress tolerance (Shinozaki and Yamaguchi-Shinozaki, 2007). Several drought inducible genes however appeared to be non-responsive to either cold or ABA treatment, such as ERD1, a Clp protease regulatory subunit ClpD encoding gene. This suggested the possible existence of other CREs involved in the ABA-independent pathway in drought stress response in addition to ABRE and DRE/CRT (Simpson *et al.*, 2003). Promoter analysis of ERD1 revealed two different novel CREs involved in drought stress response. DNA-binding proteins interacting with these novel CREs were identified as NAC and ZF-HD transcription factors (Table 1.4) (Simpson *et al.*, 2003; Tran *et al.*, 2004).

Several transcription factors have been shown to be important in the ABA-dependent abiotic stress response pathway such as ABA-responsive element binding (AREB) proteins and ABRE binding factors (ABFs). These bZIP transcription factors recognize ABRE, respond at the transcriptional and post-transcriptional level to dehydration and salt stress (Choi *et al.*, 2000; Uno *et al.*, 2000; Bartels and Sunkar, 2005). ABREs are not the only *cis*-acting element involved in the ABA-dependent abiotic stress pathway. For example, the RD22 promoter lacks the typical ABRE and DRE/CRT elements, but contains the MYB recognition sequence (MYBRS) and MYC recognition sequence (MYCRS) (Table 1.4) (Iwasaki *et al.*, 1995). The AtMYB2 and AtMYC2 transcription factors, have been shown to bind to these *cis*-acting elements in the RD22 promoter and co-operatively activate RD22 (Abe *et al.*, 1997; 2003). Overexpression of both AtMYB2 and AtMYC2 genes has resulted in an ABA-hypersensitive phenotype in transgenic plants, as well as improved osmotic stress tolerance (Abe *et al.*, 2003). Target genes of AtMYB2 and AtMYC2 include alcohol dehydrogenase and other ABA or jasmonic acid (JA) inducible genes (Abe *et al.*, 2003).

RD26, which encodes a NAC transcription factor, has also been shown to respond to ABA signaling (Fujuta *et al.*, 2004). RD26 is induced by drought, high-salinity, ABA treatment and JA treatment. Typical ABA inducible genes such as LEAs are not to be targeted by RD26, whereas many JA-inducible genes are (Fujuta *et al.*, 2004). These results suggest that RD26 may play a role in mediating cross-talk between ABA signaling and JA signaling during drought and wounding stress responses, as well as in activation of wounding-related genes (Shinozaki and Yamaguchi-Shinozaki, 2007).

Recently, the link between the major ABA signaling pathway and other signaling factors in abiotic stress response and seed development was established and reviewed (Nakashima and

Yamaguchi-Shinozaki, 2013) (Fig. 1.4). SNF1-related protein kinases 2 (SnRK2) are the key regulators of ABA signaling including the AREB/ABF regulon. When endogenous ABA level increases during abiotic stress response, or during seed development, the subclass III SnRK2s (SRK2D/E/I, Fig. 1.4) phosphorylate AREB/ABFs. The phosphorylated AREB/ABFs then bind to the ABREs in the promoter regions of the target genes to activate their expression. The SnRK2s have also been shown to phosphorylate transcription factors that are involved in non-ABRE regulons (Nakashima and Yamaguchi-Shinozaki, 2013).

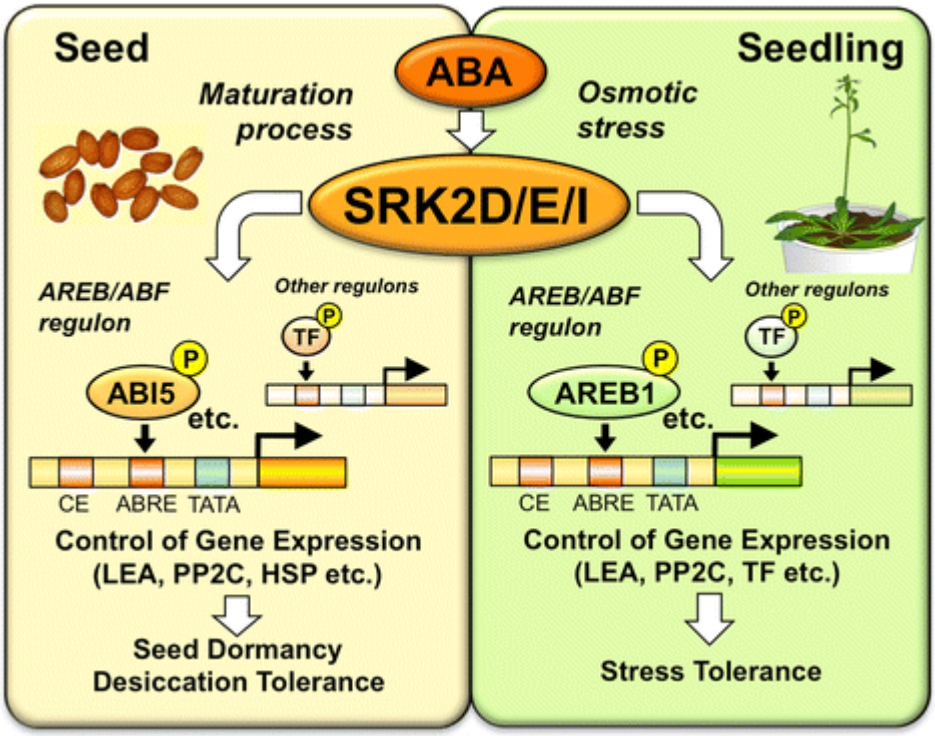


Figure 1.4. Model of the ABA-SnRK2-AREB/ABF pathway that controls ABRE-mediated transcription during seed maturation and under osmotic stress in seedlings. SRK2D/E/I: subclass III SNF1-related protein kinases 2 (SnRK2) (taken from Nakashima and Yamaguchi-Shinozaki, 2013).

1.4.3 Regulatory networks during desiccation in desiccation tolerant plants

In contrast to desiccation sensitive plants such as *A. thaliana*, relatively little is known about which transcription factors regulate gene expression during desiccation in vegetative tissues of desiccation tolerant plants. Similar to desiccation sensitive plants and orthodox seeds, ABA is thought to play a coordinating role in the activation of tolerance genes in response to desiccation (Toldi *et al.*, 2009). However, several genes expressed during the early phase of desiccation in *C. plantagineum* were found not to respond to ABA. This indicates the

possibility of other signaling pathways involved in desiccation tolerance (Frank *et al.*, 2000; Toldi *et al.*, 2009).

Several transcription factors that are activated by water deficiency have been identified in different homoiochlorophyllous, dicotyledonous resurrection plants. RNA-seq analysis of *Haberlea rhodopensis* leaves taken from well hydrated, desiccating (42% or 4% RWC) and rehydrated (4 days after watering) plants identified transcription factors belonging to the NAC, NF-YA, MADS box, HSF, GRAS, and WRKY families as being highly induced in response to water deficiency (Gechev *et al.*, 2013). Wang *et al.* (2009a) characterized the promoter of the galactinol synthase gene (BhGolS1) that was dehydration and ABA-inducible in another homoiochlorophyllous resurrection plant *Boea hygrometrica*. They identified a WRKY transcription factor (BhWRKY1) that binds to a W-box promoter element of the BhGolS1 during the early phase of dehydration, which induces the synthesis of raffinose and raffinose RFOs in an ABA-dependent manner. Several transcription factors have also been identified in *C. plantagineum* upon water loss. These include a heat-shock transcription factor (Bockel *et al.*, 1998), 7 homeodomain leucine zipper (HDZip) family proteins (Frank *et al.*, 1998; Deng *et al.*, 2002) and 3 MYB genes (Iturriaga *et al.*, 1996). Deng *et al.* (2002) have reported the expression of 5 HDZip transcripts in response to ABA treatment in the undifferentiated callus of *C. plantagineum*. Two of which were induced by exogenous ABA while the other 3 were not, suggesting that these HDZip transcription factors act in different pathways of the dehydration response, i.e. ABA-mediated and ABA independent (Bartels and Hussain, 2011). Overexpression of *C. plantagineum* MYB transcription factor gene CpMYB10 has been shown to enhance tolerance to drought and hypersensitivity to ABA in transgenic *A. thaliana* (Villalobos *et al.*, 2004). Regulatory CDT-1 gene identified in *C. plantagineum* during desiccation and exogenous ABA treatment has been shown to have structural features resemble a group of SINE-retrotransposons, suggesting that it may function by expressing a small regulatory RNA molecule or small polypeptide rather than a protein (Furini *et al.*, 1997; Bernacchia and Furini, 2004; Smith-Espinoza *et al.*, 2005). Overexpression of CDT-1 gene in transgenic callus results in conferred desiccation tolerance in absence of exogenous ABA, as well as induction of ABA and dehydration induced gene expression (Furini *et al.*, 1997).

The recognition promoter elements of the regulatory genes identified in the resurrection plants seemed to be conserved, and no desiccation-specific elements have been identified so

far (Ditzer and Bartels, 2006; Leprince and Buitink, 2010; Bartels and Hussain, 2011). This suggests that new combinations of existing regulatory elements are sufficient, and are used for expressing genes in desiccation tolerance (Bartels and Hussain, 2011).

Although there is an overlap in genes involved in desiccation tolerance in orthodox seeds and vegetative tissue of resurrection plants (Illing *et al.*, 2005), the regulatory networks involved in vegetative desiccation tolerance may not be as simple as a reiteration of seed regulators taking place in the vegetative tissues of resurrection plants (Cushman and Oliver, 2011). Prieto-Dapena *et al.* (2008) constitutively expressed a seed specific transcription factor (HaHSFA9) from sunflower in transgenic tobacco plants. The overexpression of seed specific HaHSFA9 induced the ectopic expression of seed specific sHSPs and improved tolerance to severe dehydration in transgenic tobacco. However, the signature seed traits such as the expression of LEA proteins, elevation of sucrose or proline levels were not observed upon overexpression of HaHSFA9. This suggests that additional regulators of genes encoding different desiccation protectants exist, which contribute further to the desiccation tolerance of orthodox seeds and resurrection plants (Cushman and Oliver, 2011).

In my PhD, I set out to study the transitional changes of transcription factors, as well as to identify groups of co-expressed genes during desiccation in a poikilochlorophyllous, monocotyledonous, resurrection plant model to further understand the complex regulatory mechanism of vegetative desiccation tolerance.

1.5 Gene expression profiling in *X. humilis* during desiccation

The current study involved transcriptional profiling of genes regulated in response to desiccation signals in *X. humilis* (Bak.) Dur. and Schinz, a monocotyledonous resurrection plant indigenous to Southern Africa. Physiological studies done on Xerophyta species (Sherwin and Farrant, 1998; Farrant *et al.*, 1999; Farrant, 2000) revealed that during desiccation, many defensive mechanisms were activated to protect the plant from various stress damages associated with severe loss of water. These mechanisms include (1) synthesis and accumulation of sugars (i.e. sucrose), proteins (i.e. Late embryogenesis abundant (LEA) proteins, small heat shock proteins (HSP) and various compatible solutes (i.e. proline), to maintain cell integrity and to stabilize intracellular biomolecules. (2) dismantling of the photosynthetic system, folding of leaves and deposition of sunscreen pigments (i.e. anthocyanin), to limit formation of reactive oxygen species or free radicals arising from

photosynthetic activity in the absence of water and damage from UV. (3) production and protection of antioxidant and free radical scavenging enzymes to neutralize any free radicals present during the desiccated state.

For the purpose of a genome-scale approach for investigating global gene expression patterns in *X. humilis*, 4 normalized libraries were constructed, representing 10900 *X. humilis* cDNA clones from root and leaf tissues that are expressed during the desiccation-rehydration cycle (Collett *et al.*, 2004). In a preliminary study, expression analysis on a small subset of 424 annotated cDNA in fully hydrated and desiccated leaves by microarray and reverse northern blots, identified a total of 55 desiccation-inducible genes and 79 desiccation down-regulated genes (Collett *et al.*, 2004). Sixteen LEA genes from groups 1, 2, 3, 4, 6, 7, 8 and 10, as well as 10 antioxidants including an orthologue of seed specific 1-cys peroxiredoxin, were among the list of up-regulated genes. Fourteen of these differentially expressed genes were confirmed by northern blot analysis. In addition, the functional classification of the up- and down-regulated genes identified is in keeping with the physiological behaviour of poikilochlorophyllous *X. humilis* during desiccation.

In the current study, we extended our analysis to a set of 3105 *X. humilis* cDNAs that were printed on a 'boutique' microarray slide. The aims of this study are (1) to investigate the changes in mRNA transcript abundance in leaf tissue at six different stages of water loss (100%, 80%, 60%, 40%, 20% and 5% RWC) during a desiccation treatment; (2) to identify different temporal classes of genes, as well as functional classes that are activated or repressed during desiccation; (3) to compare the results to the publicly available results obtained from other microarray studies that measured the transitional changes in gene expression in the vegetative tissues of desiccation sensitive plants such as *A. thaliana* during abiotic stresses, as well as the transitional expression changes in seed during seed development, in order to test the hypotheses that vegetative desiccation tolerance observed in *X. humilis* may have derived from the activation of seed specific traits.

Chapter 2

Clustering and annotation of *X. humilis* clone cDNAs

2.1 Introduction

There are many challenges in performing a microarray experiment on non-model organisms. The first being the printing of a species-specific microarray slide. In this study, EST clones were randomly selected from four normalized cDNA libraries derived from leaves and roots at various stages of desiccation (LD & RD) and rehydration (RL and RR) (Collett *et al.*, 2004), and printed on glass slides prior to any knowledge on their identity. An important part of the study was thus the full sequence analysis and annotation of all the EST clones printed on the microarray slides. Although a subset had been manually annotated (Collett *et al.*, 2004), it was not possible to take a manual approach to annotate the full EST dataset, as the process will be too time consuming and is less accurate, as no annotation score is built into retrieving GO terms directly from BLAST results.

There are a number of steps that had to be inserted in a pipeline to process and annotate the sequence data for each EST clone printed on the microarray slide. The majority of the EST clones were sequenced in both directions by high-throughput single-pass sequencing. EST sequences generated from high throughput sequencing usually have poor base read quality at the initial 50 to 100 bases, as well as the bases towards the end of the EST reads (Aaronson *et al.*, 1996). Furthermore, unnecessary additional sequences such as vector sequences have to be removed from the sequence reads. It is possible that genetic material from the host organism may be incorporated into the EST clones, and these sequences need to be excluded for the generation of a high quality EST dataset. Furthermore, even though the number of redundant clones in the four *X. humilis* EST libraries was minimized by normalization, redundant clones had been identified in the EST libraries (Collett *et al.*, 2004). It was possible that further redundancy existed by the same cDNA being represented in each of the four *X. humilis* EST libraries. It was thus highly likely that a unique *X. humilis* gene may be represented by two or more EST clones printed on the microarray slide. The redundancy can be reduced by clustering the sequence data to identify ESTs that are derived from the same gene into a unique contig. Thus, in summary, the pipeline to analyze EST sequence data involves several steps including pre-processing of raw sequence data to remove low quality reads and unwanted vector and adapter sequences, clustering and assembly of sequence data into contigs, generation of a consensus sequence and peptide sequence prediction, and

annotation of the EST contig consensus sequences (Nagaraj *et al.*, 2007; González and Vizcaíno, 2011).

The very first step of EST analysis involves base calling and the removal of unwanted and low quality sequences to yield high-quality EST sequence data. Base calling is a process where the nucleotide sequences in the raw sequence are processed and converted from chromatograms to sequences in a FASTA format. PHRED is probably the most widely used program to derive base calls and base sequence quality (Ewing and Green, 1998; Ewing *et al.*, 1998; Hansen *et al.*, 2005). PHRED is designed to read nucleotide sequence chromatogram files, evaluating on the strength of the signal, the shape of the peak, and the local environment of the peak to call bases, and to assign quality scores to each base call (Schmid and Blaxter, 2009). These quality scores range from 4 to about 60, and are logarithmically linked to the probability that the base is called wrong. Higher values correspond to a higher quality of base call. For example, a quality score of 10 denotes an error probability of 1 in 10, which corresponds to 90% accuracy of base call. A quality score of 40 denotes an error probability of 1 in 10000, which corresponds to 99.99% accuracy. PHRED scores are then used to extract either entire sequences, or segments of specified quality. A score of 20 or more, representing accuracy of 99% or higher, is commonly regarded as a good quality indication on a base call (González and Vizcaíno, 2011). Trace2dbest is one of the software tools developed for base calling that incorporates PHRED (Parkinson *et al.*, 2004). In addition to base calling, trace2dbest also identifies and removes vector sequences, as well as host contamination sequences and adaptor sequences via a cross_match algorithm (Schmid and Blaxter, 2009). The Poly(A) tail is usually trimmed to retain a few adenines (usually 6–10) to get high-quality ESTs for clustering and assembly process (Nagaraj *et al.*, 2007). In summary, only high quality EST sequence fragments are derived from this pre-processing improving the efficacy of subsequent analyses.

2.1.1 EST clustering and assembly

The second phase of EST analysis involves clustering and assembly which groups the high quality EST sequences into contigs based on sequence similarity, and generates a consensus sequence for each contig, which represents a unique cDNA. In general, an ideal EST clustering program needs to be stringent enough to separate paralogues, while tolerating a certain level of sequencing error to avoid separating EST gene variants such as alternative spliced transcripts, or polymorphisms, originated from the same gene (Wang *et al.*, 2004a).

Clustering is useful as a first step in sequence assembly pipelines for the grouping of reads sharing high sequence similarity, because sequence assembly algorithms often perform better when running on sets of closely related sequences than on the whole set (Imelfort, 2009). D2_cluster is an agglomerative single linkage clustering method, which clusters the sequences by identifying and counting matching n-length words (usually n=6). This loose approach of clustering is particularly useful in detecting related cluster members resulting from alternative splicing (Burke *et al.*, 1999; Christoffels *et al.*, 2001). Sequence assembly algorithms are responsible for the more stringent grouping of sequences into different contigs based on the multiple overlapping alignment analysis of the reads, and to generate consensus sequences for these contigs. Many assembly programs have been developed, including PHRAP (Ewing and Green, 1998) and CAP3 (Huang and Madan, 1999) which are the most extensively used. While PHRAP makes use of PHRED quality scores to determine the correct consensus sequence at positions where the assembled sequences have discrepant bases, CAP3 simply removes any sequence overlaps which do not match, before the generation of the contig consensus sequences. Huang and Madan (1999) have reported, that based on the results from comparison of CAP3 and PHRAP performances on same data, PHRAP was shown to produce longer contigs than CAP3, whereas CAP3 produced fewer errors in the final consensus sequences. However, CAP3 has been reported to produce more errors in EST clustering, where ESTs from the same gene do not form a cluster, and ESTs from distinct genes are wrongly clustered together (Wang *et al.*, 2004a; Nagaraj *et al.*, 2007). STACK (Sequence tag alignment and consensus knowledgebase) is an algorithm developed by The South African National Bioinformatics Institute that combines d2_cluster and PHRAP (Miller *et al.*, 1999; Christoffels *et al.*, 2001). It first performs a loose, unsupervised clustering to group the pre-processed ESTs using d2_cluster algorithm, then uses the PHRAP algorithm for assembly and consensus sequence generation on the clusters identified by d2_cluster.

2.1.2 EST contig peptide prediction

The next challenge, after determining the consensus sequence for each EST contig, is annotation. Although establishing possible gene identities and predicting gene function can be assigned via database similarity searches based on nucleotide sequences, it has been argued that amino acid sequences may be better than nucleotide sequences for identifying protein domains and motifs, as well as predicting the localization and functions of putative gene products. Thus the robust translation of EST contig consensus sequences into polypeptides is important to complete before annotation (Wasmuth and Blaxter, 2004; 2009).

Prot4EST is one of the software pipelines developed for EST translation that incorporates a suite of algorithms involved with open reading frame (ORF) prediction and conceptual translations (Wasmuth and Blaxter, 2004). In prot4EST, EST contig consensus sequences are processed for peptide predictions in a 6 tiered, rule-based system. During the first tier of analysis, nucleotide sequences of ribosomal RNA (rRNA) genes are identified via a BLASTN search against the rRNA sequence database. These rRNA sequences are excluded from the translation process. In the second tier analysis, a BLASTX search is performed against proteins encoded by mitochondrial genomes. The remaining sequences with no significant hits are then searched against the SwissProt database (Boeckmann *et al.*, 2003) during tier 3 analysis via BLASTX, and the representative ORF is selected based on the scores of the significant hits. Sequences with no significant similarity to mitochondrial proteins, nor to the proteins from SwissProt are moved to the remaining tiers of the process. Tier 4 and tier 5 involve the detection and extraction of coding regions, correction of frame shift errors, as well as conceptual translations on the sequences using algorithms ESTScan (Iseli *et al.*, 1999) and DECODER (Fukunishi and Hayashizaki, 2001) respectively. Two threshold criteria are applied to each putative polypeptide before it is accepted. The ORF must be at least 30 codons in length, and cover at least 10% of the input sequence. Sequences that fail these criteria are passed onto tier 6. In a last attempt to provide a putative polypeptide translation, prot4EST determines and identifies the longest string of amino acids resulted from all six-frame translations of the sequence. If a methionine is found in this longest string of amino acids, it is noted as a potential initiation site (Wasmuth and Blaxter, 2004).

2.1.3 EST contig annotation

The first generation of methods used to annotate EST contigs, were based on determining sequence similarities via BLAST searches of the queried nucleotide or protein sequences against publically available databases at Genbank, and extracting the related functional information linked to these records. This simple method may be convenient, but the identities obtained from BLAST searches were often arbitrary and poorly defined, as different databases, as well as different groups working on different organisms may have different ways of naming or interpreting their genes. In addressing this problem, the Gene Ontology (GO) project was initiated that aims to provide a controlled vocabulary that describes genes, and any associated information that is applicable to genes from all organisms (Harris *et al.*, 2004; Maere *et al.*, 2005). GO consists of three hierarchically structured vocabularies (GO terms) that describe gene products in terms of their associated biological processes, molecular

functions, and cellular components in an organism independent manner. Each GO term has a name, a unique identifier number, and is explicitly related to its parent and child terms in each of the three hierarchical structures. Information on GO terms and their relationships are revised and updated regularly by the Gene Ontology Consortium (Maere *et al.*, 2005; Dimmer *et al.*, 2007). GO terms can be assigned to a gene product by several different methods including for example experimental evidence such as direct essays, or by computational means which uses sequence similarity to other known genes to transfer associated GO terms. Different codes are used to describe the method of evidence used to link GO terms to gene products (Harris *et al.*, 2004; http://www.geneontology.org/GO.evidence.shtml).

The next generation of methods to annotate sequences not only captures the names of the subject genes identified by BLAST that share high sequence similarity with the query genes, but also captures any associated GO terms that describe the biological, molecular and cellular functions of the subject genes after BLAST search. Several tools have been developed for the automated annotation of datasets and include GoFigure (Khan *et al.*, 2003), GOblet (Groth *et al.*, 2004) and GOtcha (Martin *et al.*, 2004). These tools perform a BLAST search on the submitted nucleotide or peptide sequences to find matching homologs, and then retrieve GO terms associated with the BLAST hit sequences, and apply them to the input sequences. These tools have their own scoring system in assessing the retrieval of correct GO terms, but none of them takes the degree of similarity between the BLAST hits and the query sequence into account, nor do they consider the evidence codes when retrieving GO terms. Blast2GO, a species independent annotation tool (Conesa *et al.*, 2005; Conesa and Götz, 2008; Götz *et al.*, 2008) directly addresses these points, by including rules for annotation in retrieving GO terms associated with BLAST hits. In Blast2GO, the first step is to identify significant similarity between a dataset of nucleotide or peptide sequences via BLAST searching against either the NCBI or custom databases. After capturing a list of the best significant BLAST hits above a threshold value defined by the user, for each uploaded sequence, GO terms are retrieved from constantly updated annotation file stored at the Blast2GO server obtained from the GO Consortium, and are mapped to these BLAST results. An annotation rule is used to only retrieve reliable GO terms. The annotation rule involves calculating an annotation score that takes into consideration the % similarity between the query sequence and each of the BLAST hits; the evidence code weight that describes how each GO term is linked to each BLAST hit as well as the relationship between each GO term and its parent and child terms.

GO terms with a calculated annotation score greater than a defined threshold are considered as reliable and linked to each sequence entry. In addition to the annotation of GO terms via the BLAST, mapping, and annotation route, additional GO terms can also be retrieved and included via an InterPro scan. InterPro terms, or the domain/motif information of each sequence entry are retrieved during the InterPro scan. Any additional GO terms associated with these InterPro terms can be transferred from the InterPro database to the query sequences and merged with existing annotation. The GO records can also be augmented with ANNEX, a set of manually reviewed relations between the three GO categories, which adds additional GO annotation information based on the secure relationships (Myhre *et al.*, 2006). For example, cellular and biological information such as nucleus and transcription would be added for a contig identified as a histone.

Blast2GO also offers visualization tools on the combined annotation for all the sequences, or a group of sequences, on a GO DAG known as a combined graph, which is very useful in studying the collective biological meaning of a set of sequences. Annotation results in Blast2GO can be summarized with a combined graph through GOSlim mapping. GOSlim is a cut-down version of the GO ontologies containing a subset of the key terms in the whole GO. By reducing total number of terms, GOSlim summarizes a set of GO annotations from a detailed complex representation of the functional content to a simpler one. Thus GOSlim on the whole is very useful for giving a summary of the results of GO annotation (Conesa and Götz, 2008). Different GOSlim mappings adapted to specific species are available in Blast2GO, for example, "GOSlim_plant" developed by The Arabidopsis Information Resource (TAIR), and "GOSlim_yeast" developed by Saccharomyces Genome Database (SGD).

In this chapter, a dataset of 7312 sequence reads generated from 3123 cDNA clones, with 3105 being analyzed by the current microarray study and 18 being sequenced previously, was pre-processed, and clustered to identify 1775 contigs representing unique *X. humilis* genes. Blast2GO was used to annotate this dataset, and to link GO and Interpro terms to the dataset.

2.2 Material and Methods

2.2.1 Sequencing of X. humilis cDNAs

X. humilis cDNA library clones (Collett *et al.*, 2004) were inoculated and allowed to grow overnight at 37 °C with gentle agitation in Luria broth supplemented with 100 µg/ml ampicillin in 96-well plates. These cultures were supplemented with glycerol to a final concentration of 25% before sending them to the High-Throughput Genomics Unit (HTGU), Department of Genomic Sciences, University of Washington, Seattle, WA, USA, for sequencing with M13F and M13R primers. Sequence information published previously (Collett *et al.*, 2004) was included in the sequence dataset.

2.2.2 Preprocessing and clustering of X. humilis cDNA sequences

Galaxy, a web-based genome analysis tool (Blankenberg *et al.*, 2010; http://omweso.cbio.uct.ac.za/galaxy/) was used to build a pipeline for the preprocessing, clustering and peptide prediction of the EST sequence dataset (Fig. 2.1). The pipeline and the analyses were established and carried out by Mr Gerrit Botha from the Computational Biology Group (CBIO), University of Cape Town (August 2009 to June 2011). The bases from the resulted raw sequence chromatogram files (.abI) were called, and check for quality scores, and preprocessed to remove vector and adaptor sequences, and any contaminating *E. coli* sequences, before being converted into FASTA formatted sequence trace files by trace2dbest (Parkinson *et al.*, 2004) (Step 1, Fig. 2.1). Sequences derived from sequencing with the M13F primer, were reverse complemented (Step 2, Fig. 2.1), before submitting all sequences to stackPACK for clustering (Miller *et al.*, 1999; Christoffels *et al.*, 2001). In stackPACK, the sequences were initially clustered by d2_cluster (Burke *et al.*, 1999) (Step 3, Fig. 2.1). These initial clusters were then further assembled and partitioned into contigs by PHRAP (Ewing and Green, 1998) (Step 4, Fig. 2.1). Prot4EST was used to predict a peptide sequence for the consensus nucleotide sequence derived from each PHRAP contig, using *Oryza sativa* codon usage information (Step 5, Fig. 2.1).

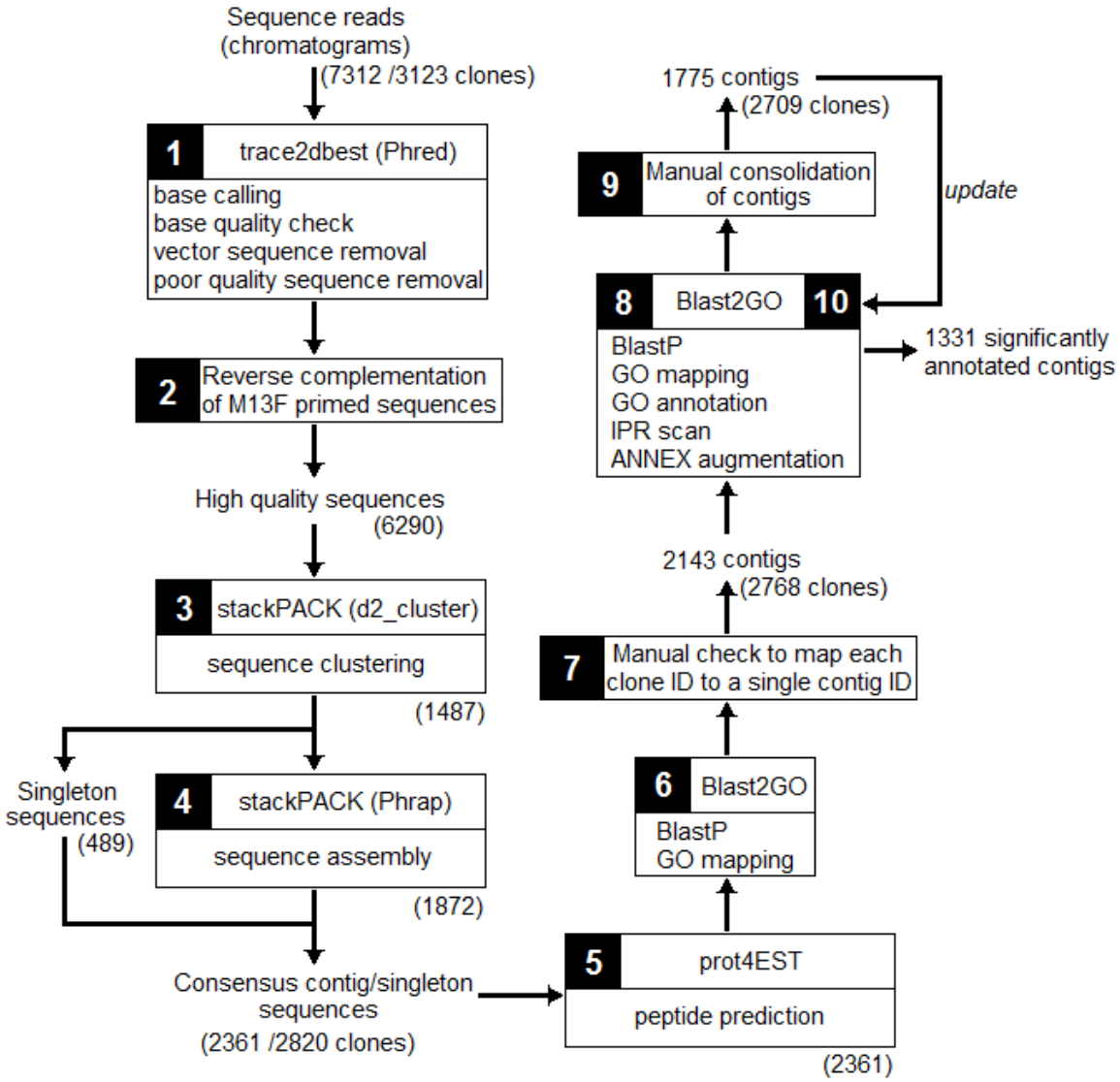


Figure 2.1. An overview of the different steps and algorithms involved in the *X. humilis* EST pipeline. Figures in parenthesis refer to the number of clones, reads, clusters or contigs processed at each step. The chromatogram .abI sequence files from the sequencing of the *X. humilis* cDNA clones printed on the microarray slides, together with sequence files for 18 cDNA clones from a previous were submitted to trace2dbest for base calling, vector sequence and adaptor sequence trimming, and base quality assessment (Step 1), as well as sequence orientation correction (Step 2). Sequences passed by the PHRED algorithm were and grouped by the d2_cluster algorithm (Step 3), then subsequently reclustered by (Step 4). Peptide sequences of the resulting PHRAP contigs and singleton sequences (referred to as contigs too) were predicted by the prot4EST algorithm (Step 5). The peptide sequences were used to perform a first round of BLASTP and mapping in Blast2GO (Step 6), which was used to manual curate the ambiguous clustering of cDNA clones (Step 7). Annotation of curated data was carried out in Blast2GO (Step 8). The data of clustering of cDNA clones into contigs was manually checked and consolidated (Step 9). The annotation results of the contigs were updated accordingly (Step 10). These results are described in detail in section 2.3.1-2.3.

2.2.3 Annotation of X. humilis contigs

The *X. humilis* contigs were annotated in Blast2GO (http://www.blast2go.com) (Steps 8, Fig. 2.1). The default configuration settings were used for each BLAST run, with BLAST DB was set as "nr" (April, 2012), the BLAST Program was set as BLASTP, the BLAST Mode set as QBlast-NCBI, and the BLAST Descriptor Annotator option was set to be on. BLAST was repeated three times, with different settings selected for the low complexity filter and the BLAST E Value. The first run was performed with a BLAST E Value= 1.0E-3, and the Low Complexity Filter= ON. Contigs with no BLAST results identified after the first BLAST run (red coloured in Blast2GO) were selected, and subjected to the second BLAST run with BLAST E Value= 1.0E-3, and Low Complexity Filter= OFF. Contigs which failed to have any BLAST results after the second BLAST run were selected, and subjected to the third BLAST run with BLAST E Value= 0.1, and Low Complexity Filter= OFF. GO terms linked to each BLAST result were then mapped to the respective *X. humilis* contigs in Blast2GO. Annotation was done in several steps, starting with the most stringent parameters, and then relaxing the parameters (summarized in Table 2.1) to capture additional information for sequences that failed to be annotated in the first pass.

Table 2.1. Summary of settings used for the four annotation steps in Blast2GO.

Step no	E-value-Hit-Filter	Annotation CutOff	GO Weight	Hsp-Hit Coverage CutOff	Evidence Code weighting
1	1.00E-03	50	5	0	IEA:0.7
2	1.00E-03	45	5	0	IEA:0.7
3	1.00E-03	45	5	0	IEA:0.9
4	0.1	35	5	0	IEA:0.9

An InterPro Scan was run on all the contig sequences, and additional GO terms linked to InterPro terms were accessed and merged with the annotated records for each contig. Lastly, ANNEX augmentation was run to add additional GO terms.

GOSlim annotation on the contigs was also performed in Blast2GO, using "goslim_plant", the predefined GOSlim mapping data specifically adapted for plant originated sequences. The GOSlim version of annotation was used for data statistics visualizations on the annotated *X. humilis* contigs carried out in Blast2GO.

The mapping results of GO terms to the respective *X. humilis* contigs in Blast2GO were used for the manually curation of the ambiguously clustered cDNA clones (Step 6; 7, Fig. 2.1), which is described in detail in section 2.3.2.

2.2.4 Database submission of X. humilis ESTs

All EST sequencing records, with their mapped contig IDs and associated annotation information were submitted to dbEST at Genbank under the following accession numbers:
Leaf Dehydration (LD) library ESTs: JK688342 - JK691274;
Leaf Rehydration (LR) library ESTs: JK691275 - JK691375;
Root Dehydration (RD) library ESTs: JK691376 - JK693744;
Root Rehydration (RR) library ESTs: JK693745 - JK693821.

The submission of *X. humilis* EST data was carried out by Mr Gerrit Botha from the Computational Biology Group (CBIO), University of Cape Town (January 2012).

2.2.5 Identification of LEA, antioxidant and transcription factor contigs in X. humilis

X. humilis contigs annotated as LEAs, antioxidants and transcription factors were identified in Blast2GO by selections of relevant contig descriptions, GO terms or InterPro terms described in Table 2.2, Table 2.3 and Table 2.4 respectively.

Table 2.2. Criteria used in the identification of *X. humilis* LEA contigs in Blast2GO.

Selection by contig description	
LEA	
late embryogenesis	
seed maturation	
Selection by InterPro term	**Term description**
IPR000389 (class 1 LEAs)	Stress induced protein
IPR000167 (class 2 LEAs)	Dehydrin
IPR004238 (class 3 LEAs)	Late embryogenesis abundant protein, LEA-3
IPR005513 (class 4 LEAs)	Late embryogenesis abundant protein, LEA-25/LEA-D113
IPR007011 (class 6 LEAs)	Seed maturation protein
IPR004926 (class 7 LEAs)	Late embryogenesis abundant protein, LEA-5
IPR004864 (class 8 LEAs)	Late embryogenesis abundant protein, LEA-14
IPR008390 (class 10 LEAs)	AWPM-19-like
IPR018930 (class 11 LEAs)	Late embryogenesis abundant protein, LEA-18

Table 2.3. Criteria used in the identification of *X. humilis* antioxidant contigs in Blast2GO.

Selection by GO terms	Term description
GO:0000302	response to reactive oxygen species
GO:0004096	catalase activity
GO:0004362	glutathione-disulfide reductase activity
GO:0004601	peroxidase activity
GO:0004602	glutathione peroxidase activity
GO:0004784	superoxide dismutase activity
GO:0004791	thioredoxin-disulfide reductase activity
GO:0006749	glutathione metabolic process
GO:0006802	catalase reaction
GO:0006804	peroxidase reaction
GO:0006979	response to oxidative stress
GO:0016209	antioxidant activity
GO:0032542	sulfiredoxin activity
GO:0045174	glutathione dehydrogenase (ascorbate) activity
GO:0050605	superoxide reductase activity

Table 2.4. Criteria used in the identification of *X. humilis* transcription factor contigs in Blast2GO.

Selection by GO term	Term description
GO:0000785	chromatin
GO:0003676	nucleic acid binding, Interacting selectively and non-covalently with any nucleic acid
GO:0003677	DNA binding; Any molecular function by which a gene product interacts selectively with DNA
GO:0003700	transcription factor activity
GO:0003712	transcription cofactor activity
GO:0003713	transcription coactivator activity
GO:0004402	histone acetyltransferase activity
GO:0005667	transcription factor complex
GO:0005669	transcription factor TFIID complex
GO:0006350	The cellular synthesis of either RNA on a template of DNA or DNA on a template of RNA
GO:0006355	regulation of cellular transcription, DNA-dependent
GO:0008017	microtubule/chromatin interaction
GO:0008134	transcription factor binding
GO:0016481	negative regulation of transcription
GO:0016563	transcription activator activity
GO:0016564	transcription repressor activity
GO:0016568	chromatin modification
GO:0030528	transcription regulator activity
GO:0043565	sequence-specific DNA binding
GO:0045449	regulation of cellular transcription
GO:0045893	positive regulation of transcription, DNA-dependent

2.3 Results and Discussion

2.3.1 First round of clustering and annotation

The 7312 chromatogram .abI sequence files from the sequencing of the 3105 cDNA clones printed on the microarray slides, together with sequence files for 18 cDNA clones from a previous study (Collett *et al.*, 2004), were submitted to trace2dbest for base calling, vector sequence and adaptor sequence trimming, and base quality assessment (Step 1, Fig. 2.1). A total of 6290 sequences were passed by the PHRED algorithm (Step 2, Fig. 2.1). These 6290 sequences were grouped by the d2_cluster algorithm into 489 singleton groups and 1487 d2 cluster groups (with two or more member sequences in each group) (Step 3, Fig. 2.1). These 1487 d2 clusters were subsequently reclustered by PHRAP into 1872 PHRAP contigs (Step 4, Fig. 2.1). The 1872 PHRAP contigs and 489 singleton sequences (referred to as contigs too henceforward), (i.e. 2361 contigs in total) represented a total of 2820 *X. humilis* cDNA clones. Based on the consensus nucleotide sequence derived by PHRAP, the 2361 peptide sequences representing these cDNA clones were predicted by the prot4EST algorithm (Step 5, Fig. 2.1). These peptide sequences were used to perform a first round of mapping in Blast2GO (Step 6, Fig. 2.1).

2.3.2 Second round of clustering and annotation

The quality of the 2361 PHRAP contigs was inspected manually, and the clustering of EST sequence data from 264 cDNA clones was found to be ambiguous, with sequences derived from the same cDNA clone being clustered into different contigs. Logically, sequences derived from the same cDNA template should be grouped into the same contig group. One reason of the failure of PHRAP to group forward and reverse sequences from the same clone into the same contig, could be when a cDNA insert is too big for the sequence reads to overlap. However, in these cases, the peptide sequences derived from the forward and reverse sequencing reactions, would be predicted to generate similar BLAST results, and the GO terms associated with these results should be in agreement. Thus, the information on the gene identity, and the associated GO terms, identified by Blast2GO (Step 6, Fig. 2.1), were used to manually curate the ambiguous clustering data of the 264 cDNA clones (Step 7, Fig. 2.1).

The following criteria and set of rules was defined to assign these cDNA clones to different groups, and to deal with them consistently:

1. If the contigs had unrelated gene description or GO terms, the sequences derived from this clone were considered as unreliable and all data associated with this clone were excluded from the dataset.

2. If only one of the contigs had annotation information, all sequences of this clone were then manually assigned to that contig group with annotation information.

3. If the contigs had highly related gene descriptions, and overlapping GO terms, the sequences of the clone were then manually assigned to a newly created contig group with the extension_M, together with associated annotation information.

4. If neither of the contigs were annotated, the contig with the longest peptide sequence length was selected to represent the sequences of the clone.

5. If neither contig had any associated GO terms, but had similar gene descriptions, then the sequences of the clone were then manually assigned to a newly created contig group with the extension_M, together with associated gene description.

The assignment of the 264 clones to these groups, and the action taken are summarized in Table 2.5. A total of 2143 contigs remained after this manual curation step, which represented 2768 cDNA clones. The predicted peptide sequences of these 2143 contigs were then annotated in Blast2GO (Step 8, Fig. 2.1).

Table 2.5. A breakdown of the 264 EST cDNAs whose sequence reads did not overlap.

Rule no	Description	No of EST cDNAs
1	Annotation associated with contigs were in disagreement	52
2	Only one contig had annotation information	70
3	Annotation terms in agreement	127
4	None of the contigs had annotation information, contig of longest peptide selected	11
5	Annotation terms did not overlap, but gene descriptions showed association	4

2.3.3 Third iteration of clustering and annotation

Although the majority of the 1487 d2 cluster groups mapped to one PHRAP contig, 225 d2 clusters were split into two or more contigs by PHRAP. For example d2 cluster XHP00030 was split into 9 PHRAP contigs, XHP00030_1 to 3, and XHP00030_6 to 11. A manual inspection of these PHRAP contigs showed that in many cases the consensus sequence of these PHRAP contigs shared a high similarity (Fig. 2.2; Fig. 2.3). It could be potentially inaccurate to regard each PHRAP contig as representing a unique gene, and could lead to inflation of particular classes of genes in the microarray data analysis. A further round of curation was thus performed to manually check the validity of keeping these PHRAP contigs separate or not (Step 9, Fig. 2.1).

```
XHP00030_1      1  GCACGAGG-- ---------A ACAGATCTTT CAGTTAAAAG CTAGTAGTTT
XHP00030_2      1  GCACGAGG-- ---------- ---------- ---------- ----------
XHP00030_3      1  g-CCGAGG-- ---------- ---------- ---------- ----------
XHP00030_6      1  cCACGAGG-- ---------- ---------- ---------- ----------
XHP00030_7      1  GCACGAGGCa acAGCAGTAA ACAGATCTTT CAGTTAAAAG CTAGTAGTTT
XHP00030_8      1  GCACGAGG-- -----AATAA ACAGATCTTT CAGATAAAAG CTAGAAGTTT
XHP00030_9      1  GCACGAGGC- --AGCAATAA ACAGATCTTT CAGATAAAAG CTAGTAGTTT
XHP00030_10     1  GCACGAGG-- -----AGTAA ACAGATCTTT CAGTTAAAAG CTAGTAGTTT
XHP00030_11     1  GCACGAGGC- --AGCAGTAA ACAGATCTTT CAGTTAAAAG CTAGTAGTTT

XHP00030_1     40  TAAGTTCAGT TTTCAAGACA TAGAGATGGA GGGTTTTGGG AGCCAGCAGC
XHP00030_2      9  ---------- ---------- ---------- ---------- ----------
XHP00030_3      8  ---------- ---------- ---------- ---------- ----------
XHP00030_6      9  ---------- ---------- ---------- ---------- ----------
XHP00030_7     51  TAAGTTCAGT TTTCAAGACA TAGAGATGGA GGGTTTTGGG AGCCAGCAGC
XHP00030_8     44  TAAGTTTAGT TTTCAAGACA TCGAGATGGA GGGTTTTGGG AACCAGCAGC
XHP00030_9     48  TAAGTTCAGT TTTCAAGACA TAGAGATGGA GGGTTTTGGG AGCCAGCAGC
XHP00030_10    44  TAAGTTCAGT TTTCAAGACA TAGAGATGGA GGGTTTTGGG AGCCAGCAGC
XHP00030_11    48  TAAGTTCAGT TTTCAAGACA TAGAGATGGA GGGTTTTGGG AGCCAGCAGC

XHP00030_1     90  ACGACCAGCA CCGACACCAG CAGGGCACCG ACCAGTTCGG CTCCCATGTC
XHP00030_2      9  ---------- ---------- ---------- ---------- ----------
XHP00030_3      8  ---------- ---------- ---------- ---------- ----------
XHP00030_6      9  ---------- ---------- ---------- ---------- ----------
XHP00030_7    101  ACGACCAGCA CCGACACCAG CAGGGCACCG ACCAGTTCGG CTCCCATGTC
XHP00030_8     94  ACGACCAGCA CCGACACCAG CAGGGCACCG ACCAGTTCGG CTCCCATGTC
XHP00030_9     98  ACGACCAGCA CCGACACCAG CAGGGCACCG ACCAGTTCGG CTCCCATGTC
XHP00030_10    94  ACGACCAGCA CCGACACCAG CAGGGCACCG ACCAGTTCGG CTCCCATGTC
XHP00030_11    98  ACGACCAGCA CCGACACCAG CAGGGCACCG ACCAGTTCGG CTCCCATGTC

XHP00030_1    140  CAGCCCGGGC ACGGTGGTCA GCAGGGTGTA CTCGGCGGGC AGCAGCAACA
XHP00030_2      9  ---------- ---------- ---------- ---------- ----------
XHP00030_3      8  ---------- ---------- ---------- ---------- ----------
XHP00030_6      9  ---------- ---------- ---------- ---------- ----------
XHP00030_7    151  CAGCCCGGGC ACGGTGGTCA GCAGGGTGTA CTCGGCGGGC AGCAGCAACA
XHP00030_8    144  CAGCCCGGGC ACGGTGGTCA GCAGGGTGTA CTCGGCGGGC AGCAGCAACA
XHP00030_9    148  CAGCCCGGGC ACGGTGGTCA GCAGGGTGTA CTCGGCGGGC AGCAGCAACA
XHP00030_10   144  CAGCCCGGGC ACGGTGGTCA GCAGGGTGTA CTCGGCGGGC AGCAGCAACA
XHP00030_11   148  CAGCCCGGGC ACGGTGGTCA GCAGGGTGTA CTCGGCGGGC AGCAGCAACA

XHP00030_1    190  CCAGCAGCAC AAGGACCAGA GTCAGGGTAT TGGTTCCGGC ATTAGCAGCA
XHP00030_2      9  ---------- ---------- ---------- ---------- ----------
XHP00030_3      8  ---------- ---------- ---------- ---------- -------GCA
XHP00030_6      9  ---------- ---------- ----GGGTAT TGGTTCCGGC ATTAGCAGCA
XHP00030_7    201  CCAGCAGCAC AAGGACCAGA GTCAGGGTAT TGGTTCCGGC ATTAGCAGCA
XHP00030_8    194  CCAGCAGCAC AAGGACCAGG GTCAGGGTAT TGGTTCCGGC ATTAGCAGCA
XHP00030_9    198  CCAGCAGCAC AAGGACCAGA GTCAGGGTAT TGGTTCCGGC ATTAGCAGCA
XHP00030_10   194  CCAGCAGCAC AAGGACCAGA GTCAGGGTAT TGGTTCCGGC ATTAGCAGCA
XHP00030_11   198  CCAGCAGCAC AAGGACCAGA GTCAGGGTAT TGGTTCCGGC ATTAGCAGCA

XHP00030_1    240  AGCTTCACCG CTCCAACAGT TCCAGCTCCA GCTC------ TGAAAGTGAT
XHP00030_2      9  ---------- ---------- ---------- ---------- ----------
XHP00030_3     11  AGCTTCACCG CTCCAACAGT TCCAGCTCCA GCTC------ TGAAAGTGAT
XHP00030_6     35  AGCTTCACCG CTCCAACAGT TCCAGCTCCA GCTC------ TGAAAGTGAT
XHP00030_7    251  AGCTTCACCG CTCCAACAGT TCCAGCTCCA GCTC------ TGAAAGTGAT
XHP00030_8    244  AGCTTCACCG CTCCAACAGT TCCAGCTCCA GCTC------ TGAGAGTGAT
XHP00030_9    248  AGCTTCACCG CTCCAACAGT TCCAGCTCCA GCTC------ TGAAAGTGAT
XHP00030_10   244  AGCTTCACCG CTCCAACAGT TCCAGCTCCA GCTCcagctc TGAAAGTGAT
XHP00030_11   248  AGCTTCACCG CTCCAACAGT TCCAGCTCCA GCTC------ TGAAAGTGAT
```

```
XHP00030_1    284  GGAGAAGGAG GGAGGAGGAA GAAGGGTATT AAGGACAAGA TCAAGGAGAA
XHP00030_2      9  ---------- ---------- ---------- ---------- ----------
XHP00030_3     55  GGAGAAGGAG GGAGGAGGAA GAAGGGTATT AAGGACAAGA TCAAGGAGAA
XHP00030_6     79  GGAGAAGGAG GGAGGAGGAA GAAGGGTATT AAGGACAAGA TCAAGGAGAA
XHP00030_7    295  GGAGAAGGAG GGAGGAGGAA GAAGGGTATT AAGGACAAGA TCAAGGAGAA
XHP00030_8    288  GGAGAAGGAG GGAGGAGGAA GAAGGGTATT AAGGACAAGA TCAAGGAGAA
XHP00030_9    292  GGAGAAGGAG GGAGGAGGAA GAAGGGTATT AAGGACAAGA TCAAGGAGAA
XHP00030_10   294  GGAGAAGGAG GGAGGAGGAA GAAGGGTATT AAGGACAAGA TCAAGGAGAA
XHP00030_11   292  GGAGAAGGAG GGAGGAGGAA GAAGGGTATT AAGGACAAGA TCAAGGAGAA

XHP00030_1    334  ACTGCCAGGG CAACACAACC AAGGACAGAC CGGTCAGCAT GGCATGACTG
XHP00030_2      9  -------GGG CAACACAACC AAGGACAGAC CGGTCAGCAT GGCATGACTG
XHP00030_3    105  ACTGCCAGGG CAACACAACC AAGGACAGAC CGGTCAGCAT GGCATGACTG
XHP00030_6    129  ACTGCCAGGG CAACACAACC AAGGACAGAC CGGTCAGCAT GGCATGACTG
XHP00030_7    345  ACTGCCAGGG CAACACAACC AAGGACAGAC CGGTCAGCAT GGCATGACTG
XHP00030_8    338  ACTGCCAGGG CAGCACAACC AAGGACAGAC CGGCATGACT GGCATGACTG
XHP00030_9    342  ACTGCCAGGG CAACACAACC AAGGACAGAC CGGTCAGCAT GGCATGACTG
XHP00030_10   344  ACTGCCAGGG CAACACAACC AAGGACAGAC CGGTCAGCAT GGCATGACTG
XHP00030_11    42  ACTGCCAGGG CAACACAACC AAGGACAGAC CGGTCAGCAT GGCATGACTG

XHP00030_1    384  GCGGCCATCA GCAGGGCATG ACCGGCATGA CTGGCGGCCA TCAGCAGGGC
XHP00030_2     52  GCGGCCATCA GCAGGGCATG ACTGGCATGA CTGGCGGCCA TCAGCAGGGC
XHP00030_3    155  GCGGCCATCA GCAGGGCATG ACCGGCATGA CTGGCGGCCA TCAGCAGGGC
XHP00030_6    179  GCGGCCATCA GCAGGGCATG ACCGGCATGA CTGGCGGCCA TCAGCAGGGC
XHP00030_7    395  GCGGCCATCA GCAGGGCATG ACTGGCATGA CTGGCGGCCA TCAGCAGGGC
XHP00030_8    388  GCGGCCATCA GCAGGGCATG ACTGGCATGA CTGGCGGCCA TCAGCAGGGC
XHP00030_9    392  GCGGCCATCA GCAGGGCATG ACTGGCATGA CTGGCGGCCA TCAGCAGGGC
XHP00030_10   394  GCGGCCATCA GCAGGGCATG ACCGGCATGA CTGGCGGCCA TCAGCAGGGC
XHP00030_11   392  GCGGCCATCA GCAGGGCATG ACCGGCATGA CTGGCGGCCA TCAGCAGGGC

XHP00030_1    434  TACGGAGCCA CCGACCAGCA TGGAGAGCAA GAGGGAATGA TGGATAAGAT
XHP00030_2    102  TACGGAGCCA CCGGCCAGCA TGGAGAGCAA GAGGGAATGA TGGATAAGAT
XHP00030_3    205  TACGGAGCCA CCGGCCAGCA TGGAGAGCAA GAGGGAATGA TGGATAAGAT
XHP00030_6    229  TACGGAGCCA CCGGCCAGCA TGGAGAGCAA GAGGGAATGA TGGATAAGAT
XHP00030_7    445  TACGGAGCCA CCGGCCAGCA TGGAGAGCAA GAGGGAATGA TGGATAAGAT
XHP00030_8    438  TACGGAGCCA CCGGCCAGCA TGGAGAGCAA GAGGGAATGA TGGATAAGAT
XHP00030_9    442  TACGGAGCCA CCGGCCAGCA TGGAGAGCAA GAGGGAATGA TGGATAAGAT
XHP00030_10   444  TACGGAGCCA CCGACCAGCA TGGAGAGCAA GAGGGAATGA TGGATAAGAT
XHP00030_11   442  TACGGAGCCA CCGGCCAGCA TGGAGAGCAA GAGGGAATGA TGGATAAGAT

XHP00030_1    484  CAAGGACAAG CTTTCCGGCA ATCAGTAAAC CTAAATACCT CCAGAATTGC
XHP00030_2    152  CAAGGACAAG CTTTCCGGCA ACCAGTAAAC CTAAGTACCT CCAGGATTGC
XHP00030_3    255  CAAGGACAAG CTTTCCGGCA ATCAGTAAAC CTAAATACCT CCAGAATTGC
XHP00030_6    279  CAAGGACAAG CTTTCCGGCA ACCAGTAAAC CTAAGTACCT CCAGGATTGC
XHP00030_7    495  CAAGGACAAG CTTTCCGGCA ATCAGTAAAC CTAAATACCT CCAGAATTGC
XHP00030_8    488  CAAGGACAAG CTTTCCGGCA ATCAGTAAAC CTAAATACCT CCAGAATTGC
XHP00030_9    492  CAAGGACAAG CTTTCCGGCA ACCAGTAAAC CTAAGTACCT CCAGAATTGC
XHP00030_10   494  CAAGGACAAG CTTTCCGGCA ATCAGTAAAC CTAAATACCT CCAGAATTGC
XHP00030_11   492  CAAGGACAAG CTTTCCGGCA ACCAGTAAAC CTAAGTACCT CCAGGATTGC

XHP00030_1    534  ATGATGAGAC GCATAAATAT ATATATTTAT GTGTATATAA GAATATGTTT
XHP00030_2    202  ATGATGAGAC GCATAAATAT ATATATTTAT GTGTATATAA GAATATGTTT
XHP00030_3    305  ATGATGAGAC GCATAAATAT ATATATTTAT GTGTATATAA GAATATGTTT
XHP00030_6    329  ATGATGAGAC GCATAAATAT ATATATTTAT GTGTATATAA GAATATGTTT
XHP00030_7    545  ATGATGAGAC GCATAAATAT ATATATTTAT GTGTATATAA GAATATGTTT
XHP00030_8    538  ATGATGAGAC GCATAAATAT ATATATTTAT GTGTATATAA GA--ATGTTT
XHP00030_9    542  ATGATGAGAC GCATAAATAT ATATATTTAT GTGTATATAA GAATATGTTT
XHP00030_10   544  ATGATGAGAC GCATAAATAT ATATATTTAT GTGTATATAA GAATATGTTT
XHP00030_11   542  ATGATGAGAC GCATAAATAT ATATATTTAT GTGTATATAA GAATATGTTT
```

```
XHP00030_1    584  GCTGTGTTTC GGTGCTGCTA AGCACCTGTG TTGATCGGTG TGTAAGTTAA
XHP00030_2    252  GCTGTGTTTC GGTGCTGCTA AGCACCTGTG TTGATCGGTG TGTAA-----
XHP00030_3    355  GCTGTGTTTC GGTGCTGCTA AGCACCTGTG TTGATCGGTG TGTAA-----
XHP00030_6    379  GCTGTGTTTC GGTGCTGCTA AGCACCTGTG TTGATCGGTG TGTAA-----
XHP00030_7    595  GCTGTGTTTC GGTGCTGCTA AGCACCTGTG TTGATCGGTG TGTAA-----
XHP00030_8    586  GCTGTGTTTC GGTGCTGCTA AGCACCTGTG TTGATCGGTG TG-----TAA
XHP00030_9    592  GCTGTGTTTC GGTGCTGCTA AGCACCTGTG TTGATCGGTG TGTAA-----
XHP00030_10   594  GCTGTGTTTC GGTGCTGCTA AGCACCTGTG TTGATCGGTG TGTAAGTTAA
XHP00030_11   592  GCTGTGTTTC GGTGCTGCTA AGCACCTGTG TTGATCGGTG TGTAA-----

XHP00030_1    634  GTGTGTGTGC GTCGTGAATA ATTCATGTGT AGCAGTGAAT ACACATGAAC
XHP00030_2    297  GTGTGTGTGC GTCGTGAATA ATTCATGTGT AGCAGTGAAT ACACATGAAC
XHP00030_3    400  GTGTGTGTGC GTCGTGAATA ATTCATGTGT AGCAGTGAAT ACACATGAAC
XHP00030_6    424  GTGTGTGTGC GTCGTGAATA ATTCATGTGT AGCAGTGAAT ACACATGAAC
XHP00030_7    640  GTGTGTGTGC GTCGTGAATA ATTCATGTGT AGCAGTGAAT ACACATGAAC
XHP00030_8    631  GTGTGTGTGC GTCGTGAATA ATCCATGTGT AGCAGTGAAT ACACATGAAC
XHP00030_9    637  GTGTGTGTGC GTCGTGAATA ATTCATGTGT AGCAGTGAAT ACACATGAAC
XHP00030_10   644  GTGTGTGTGC GTCGTGAATA ATTCATGTGT AGCAGTGAAT ACACATGAAC
XHP00030_11   637  GTGTGTGTGC GTCGTGAATA ATTCATGTGT AGCAGTGAAT ACACATGAAC

XHP00030_1    684  GCTATGGTTC ATCTTTTATC GTACTTGAAT GCGAAGTC-T TT-GAGTTTT
XHP00030_2    347  GCTATGGTTC ATCTTTTATC GTAAAAAAAa ---------- ----------
XHP00030_3    450  GCTATGGTTC ATCTTTTATC GTACTTGAAT GCGAAGTCTT TT-GAGTTTT
XHP00030_6    474  GCTATGGTTC ATCTTTTATC GTACTTGAAT GCGAAGTCTT TT-GAGTTTT
XHP00030_7    690  GCTATGGTTC ATCTTTTATC GTACTTGAAT GCGAAGTCTT TT-GAGTTTT
XHP00030_8    681  GCTATGGTTC ATTTTTTATC GTACTTGAAT GCGAAGTC-T TTtAAGTTTT
XHP00030_9    687  GCTATGGTTC ATCTTTTATC GTACTTGAAT GCGAAGTCTT TT-GAGTTTT
XHP00030_10   694  GCTATGGTTC ATCTTTTATC GTACTTGAAT GCGAAGTC-T TT-GAGTTTT
XHP00030_11   687  GCTATGGTTC ATCTTTTATC GTACTTGAAT GCGAAGTCTT TT-GAGTTTT

XHP00030_1    732  TTCTTAAAAA A--------- ---------- ---------- ----------
XHP00030_2    377  ---------- ---------- ---------- ---------- ----------
XHP00030_3    499  TTCTTCATGA CCAGTGTGTA CCTCGTTGAG GTGTTCCAAA AAA-------
XHP00030_6    523  TTCTTAAACA AAA------- ---------- -------AAA AA--------
XHP00030_7    739  TTCTTCAAAA AAA------- ---------- ---------- ----------
XHP00030_8    730  TTCTTCAAAA AA-------- ---------- ---------- ----------
XHP00030_9    736  TTCTTCATGA AAA------- ---------- -------AAA AAA-------
XHP00030_10   742  TTCTTCATGA CCAGTGTGTA CCTCGTTGAG GTGTTCCAGA TTTACTGTTT
XHP00030_11   736  TTCTTCATGA CCAGTGTGTT CCTCGTTGAG ATGTTCCAGA TTTACTGTTT

XHP00030_1    743  ---------- ---------- ---------- ---------- ----------
XHP00030_2    377  ---------- ---------- ---------- ---------- ----------
XHP00030_3    542  ---------- ---------- ---------- ---------- ----------
XHP00030_6    541  ---------- ---------- ---------- ---------- ----------
XHP00030_7    752  ---------- ---------- ---------- ---------- ----------
XHP00030_8    742  ---------- ---------- ---------- ---------- ----------
XHP00030_9    755  ---------- ---------- ---------- ---------- ----------
XHP00030_10   792  TCTTGGCTTG GTTTTTGTTT ---------- ---------- ----------
XHP00030_11   786  TCTTGGCTTG GTTTTTGTTT aaagcctgcg aagtgttgac actcactgat


XHP00030_1    743  ---------- ---------- ---------- ---------
XHP00030_2    377  ---------- ---------- ---------- ---------
XHP00030_3    542  ---------- ---------- ---------- ---------
XHP00030_6    541  ---------- ---------- ---------- ---------
XHP00030_7    752  ---------- ---------- ---------- ---------
XHP00030_8    742  ---------- ---------- ---------- ---------
XHP00030_9    755  ---------- ---------- ---------- ---------
XHP00030_10   812  ---------- ---------- ---------- ---------
XHP00030_11   836  agttttgggg acttgcatat cactactgtt aaaaaatat
```

Figure 2.2. Nucleotide sequence similarity among the PHRAP contigs assembled in d2 cluster XHP00030. The consensus nucleotide sequences of the contigs from d2 cluster XHP0003 derived after PHRAP assembly were aligned using DIALIGN-TX with default settings (http://dialign-tx.gobics.de; Subramanian *et al.*, 2008).

```
XHP00030_1     1  MEGFGSQQHD QHRHQQGTDQ FGSHVQPGHG GQQGVLGGQQ QHQQHKDQSQ
XHP00030_2     1  xtr------- ---------- ---------- ---------- ----------
XHP00030_3     1  a--------- ---------- ---------- ---------- ----------
XHP00030_6     1  prg------- ---------- ---------- ---------- ----------
XHP00030_7     1  MEGFGSQQHD QHRHQQGTDQ FGSHVQPGHG GQQGVLGGQQ QHQQHKDQSQ
XHP00030_8     1  MEGFGNQQHD QHRHQQGTDQ FGSHVQPGHG GQQGVLGGQQ QHQQHKDQGQ
XHP00030_9     1  MEGFGSQQHD QHRHQQGTDQ FGSHVQPGHG GQQGVLGGQQ QHQQHKDQSQ
XHP00030_10    1  MEGFGSQQHD QHRHQQGTDQ FGSHVQPGHG GQQGVLGGQQ QHQQHKDQSQ
XHP00030_11    1  ---------- ---------- ---------- ---------- ----------

XHP00030_1    51  GIGSGISSKL HRSNSSSSSS --ESDGEGGR RKKGIKDKIK EKLPGQHNQG
XHP00030_2     4  ---------- ---------- ---------- ---------- ----GQHNQG
XHP00030_3     2  ------EGKL HRSNSSSS-- SSESDGEGGR RKKGIKDKIK EKLPGQHNQG
XHP00030_6     4  GIGSGISSKL HRSNSSSSSS --ESDGEGGR RKKGIKDKIK EKLPGQHNQG
XHP00030_7    51  GIGSGISSKL HRSNSSSSSS --ESDGEGGR RKKGIKDKIK EKLPGQHNQG
XHP00030_8    51  GIGSGISSKL HRSNSSSSSS --ESDGEGGR RKKGIKDKIK EKLPGQHNQG
XHP00030_9    51  GIGSGISSKL HRSNSSSSSS --ESDGEGGR RKKGIKDKIK EKLPGQHNQG
XHP00030_10   51  GIGSGISSKL HRSNSSSSSS SSESDGEGGR RKKGIKDKIK EKLPGQHNQG
XHP00030_11    1  ---------- HRSNSSSS-- SSESDGEGGR RKKGIKDKIK EKLPGQHNQG

XHP00030_1    99  QTGQHGMTGG HQQGMTGMTG GHQQGYGATD QHGEQEGMMD KIKDKLSGNQ
XHP00030_2    10  QTGQHGMTGG HQQGMTGMTG GHQQGYGATG QHGEQEGMMD KIKDKLSGNQ
XHP00030_3    44  QTGQHGMTGG HQQGMTGMTG GHQQGYGATG QHGEQEGMMD KIKDKLSGNQ
XHP00030_6    52  QTGQHGMTGG HQQGMTGMTG GHQQGYGATG QHGEQEGMMD KIKDKLSGNQ
XHP00030_7    99  QTGQHGMTGG HQQGMTGMTG GHQQGYGATG QHGEQEGMMD KIKDKLSGNQ
XHP00030_8    99  QTGMTGMTGG HQQGMTGMTG GHQQGYGATG QHGEQEGMMD KIKDKLSGNQ
XHP00030_9    99  QTGQHGMTGG HQQGMTGMTG GHQQGYGATG QHGEQEGMMD KIKDKLSGNQ
XHP00030_10  101  QTGQHGMTGG HQQGMTGMTG GHQQGYGATD QHGEQEGMMD KIKDKLSGNQ
XHP00030_11   39  QTGQHGMTGG HQQGMTGMTG GHQQGYGATG QHGEQEGMMD KIKDKLSGNQ
```

Figure 2.3. Peptide sequence similarity among PHRAP contigs assembled in d2 cluster XHP00030. The peptide sequences of the contigs from d2 cluster XHP00030 predicted by prot4EST were aligned using DIALIGN-TX with default settings (http://dialign-tx.gobics.de; Subramanian *et al.*, 2008).

Use was made of the microarray data (see Chapter 3) to evaluate the validity of keeping multiple PHRAP contigs which were derived from the same d2 cluster separate. It was reasoned that if the division of a d2 clusters into multiple PHRAP contigs was indeed redundant the expression values for all the contigs corresponding to a particular d2 cluster would be identical or highly similar. An example of d2 cluster accommodating redundant PHRAP contigs was XHP00030. In addition to these 9 PHRAP contigs that shared a high similarity in nucleotide and peptide sequences (Fig. 2.2; Fig. 2.3), they also shared high similarity in their microarray expression patterns identified in *X. humilis* leaves during desiccation (Fig. 2.4). This suggested that the 9 PHRAP contigs should not be regarded as 9 different genes, and all the cDNA clones from the 9 contigs should be merged into a single representative contig regarded as 1 unique gene.

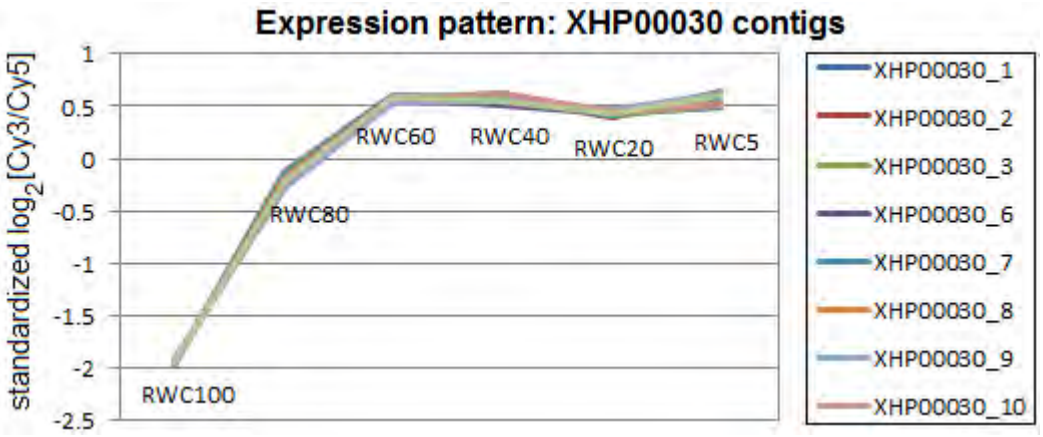


Figure 2.4. Expression pattern similarity among PHRAP contigs assembled in d2 cluster XHP00030. The normalized microarray $\log_2$ ratio values of clones (See Chapter 3) assembled in each of the contigs were extracted and averaged to represent expression pattern of each contig in *X. humilis* during desiccation. Expression patterns of all contigs were standardized using GEPAS version 4.0 (http://gepas.bioinfo.cipf.es; Tárraga *et al.*, 2008) and plotted in Excel.

With 225 d2 clusters identified having multiple PHRAP contigs, it was impossible to assess the sequence and expression similarity of PHRAP contigs within each of these d2 clusters. A more systematic approach was adapted. If the PHRAP contigs within a d2 cluster were derived from the same gene, then one would expect their expressions to be highly correlated, showing very small variation. In other words, a low data variance would be expected within the expression values obtained from the same RWC sample across all PHRAP contigs analyzed. For each of the 225 d2 clusters accommodating multiple PHRAP contigs, microarray expression data of all member contigs were extracted. A variance value was first obtained from the expression values obtained from each of the 6 RWC samples. The 6 resulted variance values were then averaged to give a mean variance, which served as an arbitrary indication of how well the member PHRAP contigs in a d2 cluster correlated on their microarray expression patterns. These mean variances ranged across the 225 d2 cluster groups from 1.34E-04 to 1.92. XHP00030 that has been shown to have redundant PHRAP contigs was determined a mean variance of 1.86E-03. The similarity assessment on PHRAP contig sequence and microarray expression pattern was decided to carry out on all 36 d2 clusters whose mean variances were above 0.5, to determine whether their division into multiple PHRAP contigs was redundant (Table 2.6, indicated in red). An additional set of 30 d2 clusters whose mean variances lower than 0.5 was also tested (Table 2.6).

Table 2.6. List 66 d2 clusters selected and tested for PHRAP contig consolidation assessment.

d2_cluster	No of PHRAP contig	No of clone	Mean variance
XHP00630	2	2	1.92
XHP00638	2	2	1.85
XHP00432	2	2	1.79
XHP00496	2	2	1.74
XHP00280	2	4	1.73
XHP00687	2	2	1.63
XHP00217	2	5	1.62
XHP00645	2	2	1.59
XHP00356	2	3	1.58
XHP00651	2	2	1.51
XHP00640	2	2	1.31
XHP00641	2	2	1.25
XHP00604	3	3	1.24
XHP00692	2	2	1.21
XHP00017	3	4	1.19
XHP00032	2	3	1.07
XHP00668	2	4	1.04
XHP00698	2	2	0.95
XHP00561	2	2	0.92
XHP00636	2	3	0.81
XHP00446	2	2	0.81
XHP00027	5	7	0.77
XHP00506	2	2	0.71
XHP00212	2	2	0.70
XHP00554	2	2	0.70
XHP00164	4	8	0.67
XHP00311	2	2	0.66
XHP00491	5	8	0.66
XHP00420	6	18	0.65
XHP00115	5	12	0.65
XHP00051	5	5	0.63
XHP00108	6	7	0.59
XHP00637	2	2	0.58
XHP00357	2	2	0.58
XHP00635	2	3	0.56
XHP00026	7	11	0.56
XHP00694	2	2	0.41
XHP00330	2	2	0.26
XHP00306	2	2	0.16
XHP00665	2	2	8.61E-02
XHP00221	2	3	6.43E-02
XHP00487	2	5	6.01E-02

Table 2.6. (continued)

d2_cluster	No of PHRAP contig	No of clone	Mean variance
XHP00062	6	17	5.58E-02
XHP00089	4	23	2.92E-02
XHP00272	2	4	1.97E-02
XHP00028	2	4	1.95E-02
XHP00416	3	6	1.58E-02
XHP00131	3	5	1.42E-02
XHP00014	3	11	1.33E-02
XHP00109	4	6	1.28E-02
XHP00664	2	2	1.22E-02
XHP00106	4	14	1.13E-02
XHP00066	3	12	8.93E-03
XHP00182	6	26	8.13E-03
XHP00531	2	3	7.13E-03
XHP00016	2	5	6.90E-03
XHP00039	2	8	5.63E-03
XHP00105	3	6	4.31E-03
XHP00052	4	14	4.04E-03
XHP00406	2	3	2.78E-03
XHP00030	9	62	1.86E-03
XHP00054	3	12	1.37E-03
XHP00055	2	6	1.34E-03
XHP00200	4	10	1.26E-03
XHP00024	2	25	7.66E-04
XHP00098	2	8	1.34E-04

D2 clusters with mean variance greater than 0.5 were indicated in red.

Many d2 clusters had contigs with highly similar nucleotide or peptide sequence content, as well as highly correlated expression data, but with individual outlier contigs present. An example was XHP00115 (mean variance of 0.65). Despite the high sequence similarity shared among the five XHP00115 contigs (Fig. 2.5; Fig. 2.6), the expression pattern of XHP00115_1 did not agree with the other 4 contigs (Fig. 2.7).

```
XHP00115_1     1 ---------- --CTCATCAA CCCCGCCATT GTTGTTTCTC TTGCAGCTTC
XHP00115_2     1 ---------- ---------- ---------- ---------- ----------
XHP00115_3     1 ---------- ---------- ---------- GTTGTTTCTC TTGCAGCTTC
XHP00115_4     1 ------CTAC GTCTCATCAA TCCCGCCATT GTTGTTTCTC TTGCAGCTTC
XHP00115_5     1 gttcgtCTAC GTCTCATCAA TCCCGCCATT ATTGTTTCTC TTGCAGCTTC

XHP00115_1    39 TCTGAACCAA AAAgTCTCTT TACCTCTGTC TTCTTCTTCT GCTGGGTATT
XHP00115_2     1 ---------- ---------- ---------- ---------- ----------
XHP00115_3    21 TCTGAAGCAA ---CTCTCTT TACCCCTATC TTCTTCTTCT GCTGGGTATT
XHP00115_4    45 ACTGAACCAA AAACTCTCTT TACCTCTATC TTCTTCTTCT GCTGGGTATT
XHP00115_5    51 ACTGAACCAA AAACTCTCTT TACCTCTATC TTCTTCTTCT GCTGGGTATT
```

```
XHP00115_1   89 TCAGCAGAGA TACAACAGTT CCACTTCTGA GATCAAAAAC TCTGTTATCT
XHP00115_2    1 ---------- ---------- ---------- ---------- ----------
XHP00115_3   68 TCAGCAGAGA TACAACAGTT CCAATTCTGA GATATAAGCC TCTGTTATCT
XHP00115_4   95 TCAGCAGAGA TACAACAGTT CCAATTCTGA GATATAAGCC TCTGTTATCT
XHP00115_5  101 TCAGCAGAGA TACAACAGTT CCACTTCTGA GATCAAAACC TCTGTTATCT

XHP00115_1  139 TCTTCGTTCA CGTCAATGGC GTCCAACGGC AACCGTCAGT TCCCACCACA
XHP00115_2    1 ---------- ---------- --------GC AACCGTCAGT TCCCACCACA
XHP00115_3  118 TCTTCATTCA CGTCAATGGC GTCCAACGGC AACCGTCAGT TCCCACCACA
XHP00115_4  145 TCTTCATTCA CGTCAATGGC GTCCAACGGC AACCGTCAGT TCACACCACA
XHP00115_5  151 TCTTCATTCA CGTCAATGGC GTCCAACGGC AACCGTCAGT TCCCACCACA

XHP00115_1  189 AGCCCAGGAC ACGCAGCCCG GTAAGGAGCA CGTCATGGAC CCGTCACCTG
XHP00115_2   23 AGCCCAGGAC ACGCAGCCCG GTAAGGAGCA CGCCATGAAC CCGTCACCTG
XHP00115_3  168 AGCCCAGGAC ACGCAGCCTG GTAAGGAGCA CGTCATGGAC CCGTCACCTG
XHP00115_4  195 AGCCCAGGAC ACGCAGCCTG GTAAGGAGCA CGTCATGGAC CCGTCACCTG
XHP00115_5  201 AGCCCAGGAC ACGCAGCCCG GTAAGGAGCA CGTCATGGAC CCGTCACCTG

XHP00115_1  239 AATTCGTCAG GCCGCACTAC AAACCCGCCA ACAAACTCCA AGGGAAAGTG
XHP00115_2   73 AATTCGTCAG GCCGCACTAC AAACCCGCCA ACAAACTCCA AGGGAAAGTG
XHP00115_3  218 AATTCGTCAG GCCGCACTAC AAACCCGCCA ACAAACTCCA AGGGAAAGTG
XHP00115_4  245 AATTCGTCAG GCCGCACTAC AAACCCGCCA ACAAACTCCA AGGGAAAGTG
XHP00115_5  251 AATTCGTCAG GCCGCACTAC AAACCCGCCA ACAAACTCCA AGGGAAAGTG

XHP00115_1  289 GCGTTGGTGA CTGGCGGCGA CTCCGGCATC GGACGGGCGG TGGCCCACCA
XHP00115_2  123 GCGTTGGTGA CTGGCGGCGA CTCCGGCATC GGACGGGCGG TGGCCCACCA
XHP00115_3  268 GCGTTGGTGA CTGGCGGCGA CTCCGGTATT GGACGGGCGG TGGCGCACCA
XHP00115_4  295 GCGTTGGTGA CTGGCGGCGA CTCCGGTATC GGACGGGCGG TGGCGCACCA
XHP00115_5  301 GCGTTGGTGA CTGGCGGCGA CTCCGGTATC GGACGGGCGG TGGCGCACCA

XHP00115_1  339 CTTCGTTCTG GAAGGCGCCA CCGTGGCCTT CACGTACGTC AAGGGTAAGG
XHP00115_2  173 CTTCGTTCTG GAAGGCGCCA CCGTGGCCTT CACGTACGTC AAGGGTAAGG
XHP00115_3  318 CTTCGTTCTG GAAGGCGCCA CCGTGGCCTT CACGTACGTC AAGGGTAAGG
XHP00115_4  345 CTTCGTTCTG GAAGGCGCCA CCGTGGCCTT CACGTACGTC AAGGGTAAGG
XHP00115_5  351 CTTCGTTCTG GAAGGCGCCA CCGTGGCCTT CACGTACGTC AAGGGTAAGG

XHP00115_1  389 AAGACAAGGA TGCGCATGAG ACCCTCAAGA TACTGAAGGA GGCCAAAGTT
XHP00115_2  223 AAGACAAGGA TGCGCATGAG ACCCTCAAGA TATTGAAGGA GGCCAAAGTT
XHP00115_3  368 AAGACAAGGA TGCGCATGAG ACTCTCAAGA TATTGAAGGA GGCCAAGGTT
XHP00115_4  395 AAGACAAGGA TGCGCATGAG ACCCTCAAGA TATTGAAGGA GGCCAAGGTT
XHP00115_5  401 AAGACAAGGA TGCGCATGAG ACCCTCAAGA TATTGAAGGA GGCCAAAGTT

XHP00115_1  439 TCTGACGCGA AGGACCCCAT AGCCATCCCC GCGGATCTCG GCTTCGAGGA
XHP00115_2  273 TCTAACGCGA AGGACCCCAT AGCCATCCCC GCGGATCTCG GCTTTGAGGA
XHP00115_3  418 TCTGATGCGA AGGAGCCCAT AGCCATCCCC GCGGATCTCG GCTTCGAGGA
XHP00115_4  445 TCTGATGCGA AGGAGCCCAT AGCCATCCCC GCGGATCTCG GCTTCGAGGA
XHP00115_5  451 TCTGACGCGA AGGACCCCAT AGCCATCCCC GCGGATCTCG GCTTCGAGGA

XHP00115_1  489 GAACTGCGCC AAGGTCGTAG AAGAGGTGGC CAAAGCTTAC GGCAAGATCG
XHP00115_2  323 GAACTGCGCC AAGGTCGTAG AAGAGGTGGC CAAAGCTTAC GGCAAGATCG
XHP00115_3  468 GAACTGCGCC AAGGTCGTAG AAGAGGTGGC CAAAGCTTAC GGCAAGATCG
XHP00115_4  495 GAACTGCGCC AAGGTCGTAG AAGAGGTGGC CAAAGCTTAC GGCAAGATCG
XHP00115_5  501 GAACTGCGCC AAGGTCGTAG AAGAGGTGGC CAAAGCTTAC GGCAAGATCG


XHP00115_1  539 ACATACTCGT CAACAACGCC GCGGAGCAGT GGGTTCAGGG CTCTATTGAA
XHP00115_2  373 ACATACTCGT CAACAACGCC GCGGAGCAGT GGGTTCAGGG CTCTATTGAA
XHP00115_3  518 ACATACTCGT CAACAACGCC GCGGAGCAGT GGGTTCAGGG CTCTATTGAA
XHP00115_4  545 ACATACTCGT CAACAACGCC GCGGAGCAGT GGGTTCAGGG CTCTATTGAA
XHP00115_5  551 ACATACTCGT CAACAACGCC GCGGAGCAGT GGGTTCAGGG CTCTATTGAA

XHP00115_1  589 GATATCAGCG CCGAGCAGCT CCAACGTGTT TTCCAAACAA ACATCTTCTC
XHP00115_2  423 GATATCAGCG CCGAGCAGCT CCAACGTGTT TTCCAAACAA ACATCTTCTC
XHP00115_3  568 GATATCAGCG CCGAGCAGCT CCAACGTGTT TTCCAAACAA ACATCTTCTC
XHP00115_4  595 GATATCAGCG CCGAGCAGCT CCAACGTGTT TTCCAAACAA ACATCTTCTC
XHP00115_5  601 GATATCAGCG CCGAGCAGCT CCAACGTGTT TTCCAAACAA ACATCTTCTC
```

```
XHP00115_1  639 CCACTTCTAC ATGACAAAGT TTGCGCTGAA GCACATGTCA GCCGGAGGCA
XHP00115_2  473 CCACTTCTAC ATGACAAAGT TTGCGCTGAA GCACATGTCA GCCGGAGGCA
XHP00115_3  618 CCACTTCTAC ATGACAAAGT TTGCGCTGAA GCACATGTCA GCCGGAGGCA
XHP00115_4  645 CCACTTCTAC ATGACAAAGT TTGCGCTGAA GCACATGTCA GCCGGAGGCA
XHP00115_5  651 CCACTTCTAC ATGACAAAGT TTGCGCTGAA GCACATGTCA GCCGGAGGCA

XHP00115_1  689 GCATAATCTG TACGACGTCG GTGAACGCAT ACAAGGGCAA CAATTCACTG
XHP00115_2  523 GCATAATCTG TACGACGTCG GTGAACGCAT ACAAGGGCAA CAATTCACTG
XHP00115_3  668 GCATAATCTG TACGACGTCG GTGAACGCGT ACAAGGGCAA CAATTCACTG
XHP00115_4  695 GCATAATCTG TACGACGTCG GTGAACGCGT ACAAGGGCAA CAATTCACTG
XHP00115_5  701 GCATAATCTG TACGACGTCG GTGAACGCGT ACAAGGGCAA CAATTCACTG

XHP00115_1  739 CTTGATTACA CGTCGACGAA GGGGGCGATA GTGGGTTTCA TCAGGGGACT
XHP00115_2  573 CTTGATTACA CGTCGACGAA GGGGGCGATA GTGGGTTTCA TCAGGGGACT
XHP00115_3  718 CTTGATTACA CGTCGACGAA GGGGGCGATA GTGGGTTTCA TCAGGGGACT
XHP00115_4  745 CTTGATTACA CGTCGACGAA GGGGGCGATA GTGGGTTTCA TCAGGGGACT
XHP00115_5  751 CTTGATTACA CGTCGACGAA GGGGGCGATA GTGGGTTTCA TCAGGGGACT

XHP00115_1  789 CGCTCTGCAG CTGGTGGAGA GAGGAATCAG GGTGAACGGG GTGGCGCCTG
XHP00115_2  623 CGCTCTGCAG CTGGTGGAGA GAGGAATCAG GGTGAACGGG GTGGCGCCTG
XHP00115_3  768 CGCTCTGCAG CTGGTGGAGA GAGGAATCAG GGTGAACGGG GTGGCGCCTG
XHP00115_4  795 CGCTCTGCAG CTGGTGGAGA GAGGAATCAG GGTGAACGGG GTGGCGCCTG
XHP00115_5  801 CGCTCTGCAG CTGGTGGAGA GAGGAATCAG GGTGAACGGG GTGGCGCCTG

XHP00115_1  839 GTCCAATCTG GATGCCGTTG ATCCCGTCGT CGTTCCCGCC GGAGAAGGTA
XHP00115_2  673 GTCCAATCTG GACGCCGTTG ATCCCGTCGT CGTTCCCGCC GGAGAAGGTG
XHP00115_3  818 GTCCAATCTG GACTCCGTTG ATCCCGTCGT CGTTCCCGCC GGAGAAGGTA
XHP00115_4  845 GTCCAATCTG GACGCCGTTG ATCCCGTCGT CGTTCCCGCC GGAGAAGGTA
XHP00115_5  851 GTCCAATCTG GACTCCGTTG ATCCCGTCGT CGTTCCCGCC GGAGAAGGTA

XHP00115_1  889 GAGAGCTTCG GGCTGGAGGT GCCGATGAAG CGGGCCGGAC AGCCGTCCGA
XHP00115_2  723 GAGAGCTTCG GGCTGGAGGT GCCGATGAAG CGGGCCGGAC AGCCGTCCGA
XHP00115_3  868 GAGAGCTTCG GGCTGGAGGT GCCGATGAAG CGGGCCGGAC AACCGTCCGA
XHP00115_4  895 GAGAGCTTCG GGCTGGAGGT GCCGATGAAG CGAGCCGGAC AGCCGTCCGA
XHP00115_5  901 GAGAGCTTTG GGCTGGAGGT GCCGATGAAG CGGGCCGGAC AGCCGTCCGA

XHP00115_1  939 GGTGGCCACG TCGTTCGTCT TCCTGGCGTC TGACGATTCT TCATACTTCA
XHP00115_2  773 GGTGGCCACG TCGTTCGTCT TCCTGGCGTC TGACGATTCT TCGTACTTCA
XHP00115_3  918 GGTGGCCACG TCGTTCGTCT TTCTGGCGTC TGACGATTCT TCGTACTTCA
XHP00115_4  945 GGTGGCCACG TCGTTCGTCT TCCTGGCGTC TGACGATTCT TCATACTTCA
XHP00115_5  951 GGTGGCCACG TCGTTCGTCT TCCTGGCGTC TGACGATTCT TCGTACTTCA

XHP00115_1  989 GCGGGCAAGT CCTCCACCCT AACGGCGGTA TGGTCGTCAA CGGTTAAGTC
XHP00115_2  823 GCGGGCAAGT CCTCCACCCT AACGGCGGTA TGGTCGTAAA CGGTTAAGTC
XHP00115_3  968 GCGGGCAAGT CCTCCACCCT AACGGCGGTA TGGTCGTAAA CGGTTAAGTC
XHP00115_4  995 GCGGGCAAGT CCTCCACCCT AACGGCGGTA TGGTCGTCAA CGGTTAAGTC
XHP00115_5 1001 GCGGGCAAGT CCTCCACCCT AACGGCGGTA TGGTCGTAAA CGGTTAAGTC

XHP00115_1 1039 CCCTGCCATT AATGGCGCAA AATATATATA TTATGGAGTT AAGATGTTGG
XHP00115_2  873 CCCTGCCATT AATGGCGCAA AATATATATA TTATGGAGAT AAGACGATAG
XHP00115_3 1018 CCCTTCCATT AATGGCGCAA AATATATATA TTATGGAGAT AAGACGATAG
XHP00115_4 1045 CC-TGCCATT AATGGCGCAA AATATACATA TTATGGAGTT AAGATGTTAG
XHP00115_5 1051 CCCTGCCATT AATGGCGCAA AATATATATA TTATGGAGAT AAGACGATAG

XHP00115_1 1089 ATATACGCCT TTCTTATCAT GTT-CGTGTT GTAGTTACGT GCGTGTAATA
XHP00115_2  923 ATATACGCCT ATCTTATCAT GTTTCATGTT GTAGTTAAGT GCGTGTAATA
XHP00115_3 1068 ATATACGCCT ATCTTATCAT GTTTCGTGTT GTAGTTAAGT GCGTGTAATA
XHP00115_4 1094 ATATACGCCT ATCTTATCTT GTT-CGTGTT GTAGTTACGT GCGTGTAATA
XHP00115_5 1101 ATATACGCCT ACCTTATCAT GTTTCGTGTT GTAGTTAAGT GCGTGTAATA

XHP00115_1 1138 TGTAACTTTA TGTGCTTCGA TAATAAATAA ATAAATAAAT ATTTTCGTTT
XHP00115_2  973 TGTAATTTTA TGAGCTGCGA TAATAAATAA ATAAATAAAT CTTTTTCTTA
XHP00115_3 1118 TGTAATTTTA TGAGCTGCGA TAATAAATAA ATAAATAAAT CTTTTTCTTA
XHP00115_4 1143 TGTAACTTTA TGTGCTTCGA TAATAAATAA ATAAATAAAT CTTTTCGTTT
XHP00115_5 1151 TGTAATTTTA TGAGCTGCGA TAATAAATAA ATAAATAAAT CTTTTTCTTA
```

```
XHP00115_1  1188 ---------- -------
XHP00115_2  1023 AAA------- -------
XHP00115_3  1168 AAAAAAAAAT AAAAAAa
XHP00115_4  1193 AAAAAAA--- -------
XHP00115_5  1201 AATTTAAAAA AAAAAA-
```

Figure 2.5. Nucleotide sequence similarity among the PHRAP contigs assembled in d2 cluster XHP00115. The consensus nucleotide sequences of the contigs from d2 cluster XHP00115 derived after PHRAP assembly were aligned using DIALIGN-TX with default settings (http://dialign-tx.gobics.de; Subramanian *et al.*, 2008).

```
XHP00115_1    1 ----LINPAI VVSLAASLNQ KVSLPLSSSS AGYFSRDTTV PLLRSKTLLS
XHP00115_2    1 ---------- ---------- ---------- ---------- ----------
XHP00115_3    1 ---------- VVSLAASLKQ -LSLPLSSSS AGYFSRDTTV PILRYKPLLS
XHP00115_4    1 --LRLINPAI VVSLAASLNQ KLSLPLSSSS AGYFSRDTTV PILRYKPLLS
XHP00115_5    1 vrLRLINPAI IVSLAASLNQ KLSLPLSSSS AGYFSRDTTV PLLRSKPLLS

XHP00115_1   47 SSFTSMASNG NRQFPPQAQD TQPGKEHVMD PSPEFVRPHY KPANKLQGKV
XHP00115_2    1 ---------- NRQFPPQAQD TQPGKEHAMN PSPEFVRPHY KPANKLQGKV
XHP00115_3   40 SSFTSMASNG NRQFPPQAQD TQPGKEHVMD PSPEFVRPHY KPANKLQGKV
XHP00115_4   49 SSFTSMASNG NRQFTPQAQD TQPGKEHVMD PSPEFVRPHY KPANKLQGKV
XHP00115_5   51 SSFTSMASNG NRQFPPQAQD TQPGKEHVMD PSPEFVRPHY KPANKLQGKV

XHP00115_1   97 ALVTGGDSGI GRAVAHHFVL EGATVAFTYV KGKEDKDAHE TLKILKEAKV
XHP00115_2   41 ALVTGGDSGI GRAVAHHFVL EGATVAFTYV KGKEDKDAHE TLKILKEAKV
XHP00115_3   90 ALVTGGDSGI GRAVAHHFVL EGATVAFTYV KGKEDKDAHE TLKILKEAKV
XHP00115_4   99 ALVTGGDSGI GRAVAHHFVL EGATVAFTYV KGKEDKDAHE TLKILKEAKV
XHP00115_5  101 ALVTGGDSGI GRAVAHHFVL EGATVAFTYV KGKEDKDAHE TLKILKEAKV

XHP00115_1  147 SDAKDPIAIP ADLGFEENCA KVVEEVAKAY GKIDILVNNA AEQWVQGSIE
XHP00115_2   91 SNAKDPIAIP ADLGFEENCA KVVEEVAKAY GKIDILVNNA AEQWVQGSIE
XHP00115_3  140 SDAKEPIAIP ADLGFEENCA KVVEEVAKAY GKIDILVNNA AEQWVQGSIE
XHP00115_4  149 SDAKEPIAIP ADLGFEENCA KVVEEVAKAY GKIDILVNNA AEQWVQGSIE
XHP00115_5  151 SDAKDPIAIP ADLGFEENCA KVVEEVAKAY GKIDILVNNA AEQWVQGSIE

XHP00115_1  197 DISAEQLQRV FQTNIFSHFY MTKFALKHMS AGGSIICTTS VNAYKGNNSL
XHP00115_2  141 DISAEQLQRV FQTNIFSHFY MTKFALKHMS AGGSIICTTS VNAYKGNNSL
XHP00115_3  190 DISAEQLQRV FQTNIFSHFY MTKFALKHMS AGGSIICTTS VNAYKGNNSL
XHP00115_4  199 DISAEQLQRV FQTNIFSHFY MTKFALKHMS AGGSIICTTS VNAYKGNNSL
XHP00115_5  201 DISAEQLQRV FQTNIFSHFY MTKFALKHMS AGGSIICTTS VNAYKGNNSL

XHP00115_1  247 LDYTSTKGAI VGFIRGLALQ LVERGIRVNG VAPGPIWMPL IPSSFPPEKV
XHP00115_2  191 LDYTSTKGAI VGFIRGLALQ LVERGIRVNG VAPGPIWTPL IPSSFPPEKV
XHP00115_3  240 LDYTSTKGAI VGFIRGLALQ LVERGIRVNG VAPGPIWTPL IPSSFPPEKV
XHP00115_4  249 LDYTSTKGAI VGFIRGLALQ LVERGIRVNG VAPGPIWTPL IPSSFPPEKV
XHP00115_5  251 LDYTSTKGAI VGFIRGLALQ LVERGIRVNG VAPGPIWTPL IPSSFPPEKV

XHP00115_1  297 ESFGLEVPMK RAGQPSEVAT SFVFLASDDS SYFSGQVLHP NGGMVVNG
XHP00115_2  241 ESFGLEVPMK RAGQPSEVAT SFVFLASDDS SYFSGQVLHP NGGMVVNG
XHP00115_3  290 ESFGLEVPMK RAGQPSEVAT SFVFLASDDS SYFSGQVLHP NGGMVVNG
XHP00115_4  299 ESFGLEVPMK RAGQPSEVAT SFVFLASDDS SYFSGQVLHP NGGMVVNG
XHP00115_5  301 ESFGLEVPMK RAGQPSEVAT SFVFLASDDS SYFSGQVLHP NGGMVVNG
```

Figure 2.6. Peptide sequence similarity among all PHRAP contigs assembled in d2 cluster XHP00115. The peptide sequences of the contigs from d2 cluster XHP00115 predicted by prot4EST were aligned using DIALIGN-TX with default settings (http://dialign-tx.gobics.de; Subramanian *et al.*, 2008).

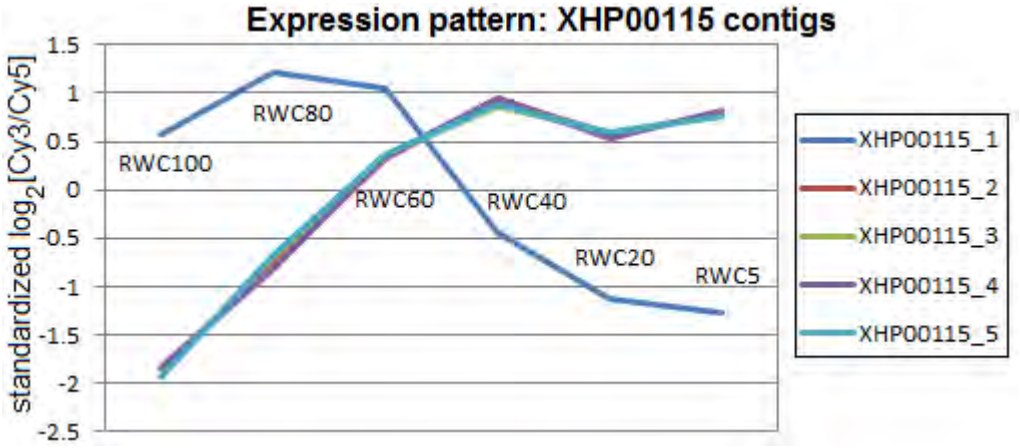


Figure 2.7. Expression pattern similarity among PHRAP contigs assembled in d2 cluster XHP00115. The normalized microarray log2 ratio values of clones (See Chapter 3) assembled in each of the contigs were extracted and averaged to represent expression pattern of each contig in *X. humilis* during desiccation. Expression patterns of all contigs were standardized using GEPAS version 4.0 (http://gepas.bioinfo.cipf.es; Tárraga *et al.*, 2008) and plotted in Excel.

It was difficult to resolve these contradictions as labeled cDNA would probably not have been able to distinguish between probes corresponding to XHP00115_1 through to XHP00115_5 in the microarray hybridizations on the basis of sequence dissimilarity. It is possible that printing errors might have occurred in some of the clones linked to XHP00115_1 which would explain why it gave a different microarray hybridization signal. Signals from XHP00115_1 were thus deemed unreliable, and all clones and sequences associated with this contig were flagged, and excluded from the microarray analysis.

There were a few cases where the nucleotide sequence alignment, predicted peptide sequence alignments of contigs showed high similarity, but were not in complete agreement. Furthermore, their microarray expression values did not correlate. An example was the XHP00446 contigs, given in Fig. 2.8, Fig. 2.9 and Fig. 2.10A. Unlike in the case of XHP00115 contigs, identification of outlier within in the 2 XHP00446 contigs was not as straight forward. XHP00446_1 and XHP00446_2 each had one clone member only, Xh_RD_08G05 and Xh_LD50A07 respectively. cDNA spots of both clones were printed 4 times on microarray slide. The expression patterns of the 4 technical replicate spots of Xh_RD_08G05 or Xh_LD50A07 were shown to be highly intact (Fig. 2.10B). This suggested that the difference in expression patterns observed in the two XHP00446 contigs was in fact substantial. Because there was no addition information in determining the

predominant expression pattern of XHP00446, it was thus decided to keep the two contigs of separate.

```
XHP00446_1    1 GCACGAGGCT ttgtcacaag actctcctta gttcttccat agagagctcc
XHP00446_2    1 GCACGAGGCT c--------- ---------- ---------- ----------

XHP00446_1   51 aatggccacc ctcaacgcca atgctctgGC CTCCACCGTC CCCCGCTTCG
XHP00446_2   12 ---------- ---------- --------GC TTCCACCATC CCCCGCTTCG

XHP00446_1  101 CGGTGCGCCA GAAGGGGTCT GTATGCGGCT CATCACCTGT TCTTGGTCTA
XHP00446_2   34 CGGTGCGCCC GAAGGGGTCT GTATGCGGCT CATCGCCTGT TCTTGGTCTA

XHP00446_1  151 CCCCAAATGA TTGTTAAcGG TGGTGGCAAG GTGAGGTGCT CAGCGTCCGA
XHP00446_2   84 CCCCAAATGA TTGTTAAgGG TGGTGGCAAG GTGAGGTGCT CAGCGTCCGA

XHP00446_1  201 GAAGAAAACA ACAAGTGTCA CAGCCGTTGC AGCTTCATCT CTGCTAGCCA
XHP00446_2  134 GAAGAAAACA ACAAGTGTCA CAGCCATTGC AGCTTCATCT CTGCTAGCCA

XHP00446_1  251 CCGCAAGTGC GGTCATGTCG AGCCCTGCCT TGGCCCTTGT CGATGAAAGG
XHP00446_2  184 CCGCAAGTGC AGTCATGTCG AGCCCTGCCT TGGCCCTTGT CGATGAGAGG

XHP00446_1  301 CTGAGCACCG AGGGTACCGG GCTCCCGTTC GGGCTAAGCA ACAACTTGTT
XHP00446_2  234 CTGAGCACCG AGGGTACCGG GCTCCCATTT GGCCTAAGCA ACAACTTGTT

XHP00446_1  351 AGGTTGGATC TTGTTCGGCG TGTTCGGTCT AATCTGGGCA CTTTTCTTTG
XHP00446_2  284 AGGCTGGATC TTGTTCGGCG TGTTCGGTCT AATCTGGGCA CTTTTCTTTG

XHP00446_1  401 TGTACACCGG CACTCTTGAG GAGGATGAGG AGTCTGGATT ATCTCTCTAA
XHP00446_2  334 TGTACACCGG CACTCTTGAG GAGGATGAGG AGTCTGGATT ATCTCTCTAA

XHP00446_1  451 GTaAAAAAAA AAGAAATTAA GGGTTATTTA CTTGTATGTA TTTTAATTTC
XHP00446_2  384 GT-AAAGAAA AAGAAATTAA GGGTTATTTA CTTGTATGTA TTTTAATTTC

XHP00446_1  501 ATATGAACTG AACAAAAAAT TAAGTACTT- -CCTTTTTTC Agttgtataa
XHP00446_2  433 ATATGAACTG AACAAAAAAT TAAGTACTTt cCATTTTTTC Aaaaaa----

XHP00446_1  549 tcagttattc tgaagttgtg gctatttgat ctctt
XHP00446_2  479 ---------- ---------- ---------- -----
```

Figure 2.8. Nucleotide sequence similarity between the PHRAP contigs assembled in d2 cluster XHP00446. The consensus nucleotide sequences of the contigs from d2 cluster XHP00446 derived after PHRAP assembly were aligned using DIALIGN-TX with default settings (http://dialign-tx.gobics.de; Subramanian *et al.*, 2008). Sequence differences are highlighted in red.

```
XHP00446_1    1 matlnanaLA STVPRFAVRQ KGSVCGSSPV LGLPQMIVNG GGKVRCSASE
XHP00446_2    1 tr------LA STIPRFAVRP KGSVCGSSPV LGLPQMIVKG GGKVRCSASE

XHP00446_1   51 KKTTSVTAVA ASSLLATASA VMSSPALALV DERLSTEGTG LPFGLSNNLL
XHP00446_2   45 KKTTSVTAIA ASSLLATASA VMSSPALALV DERLSTEGTG LPFGLSNNLL

XHP00446_1  101 GWILFGVFGL IWALFFVYTG TLEEDEESGL SL
XHP00446_2   95 GWILFGVFGL IWALFFVYTG TLEEDEESGL SL
```

Figure 2.9. Peptide sequence similarity between all PHRAP contigs assembled in d2 cluster XHP00446. The peptide sequences of the contigs from d2 cluster XHP00115 predicted by prot4EST were aligned using DIALIGN-TX with default settings (http://dialign-tx.gobics.de; Subramanian *et al.*, 2008). Sequence differences are highlighted in red.

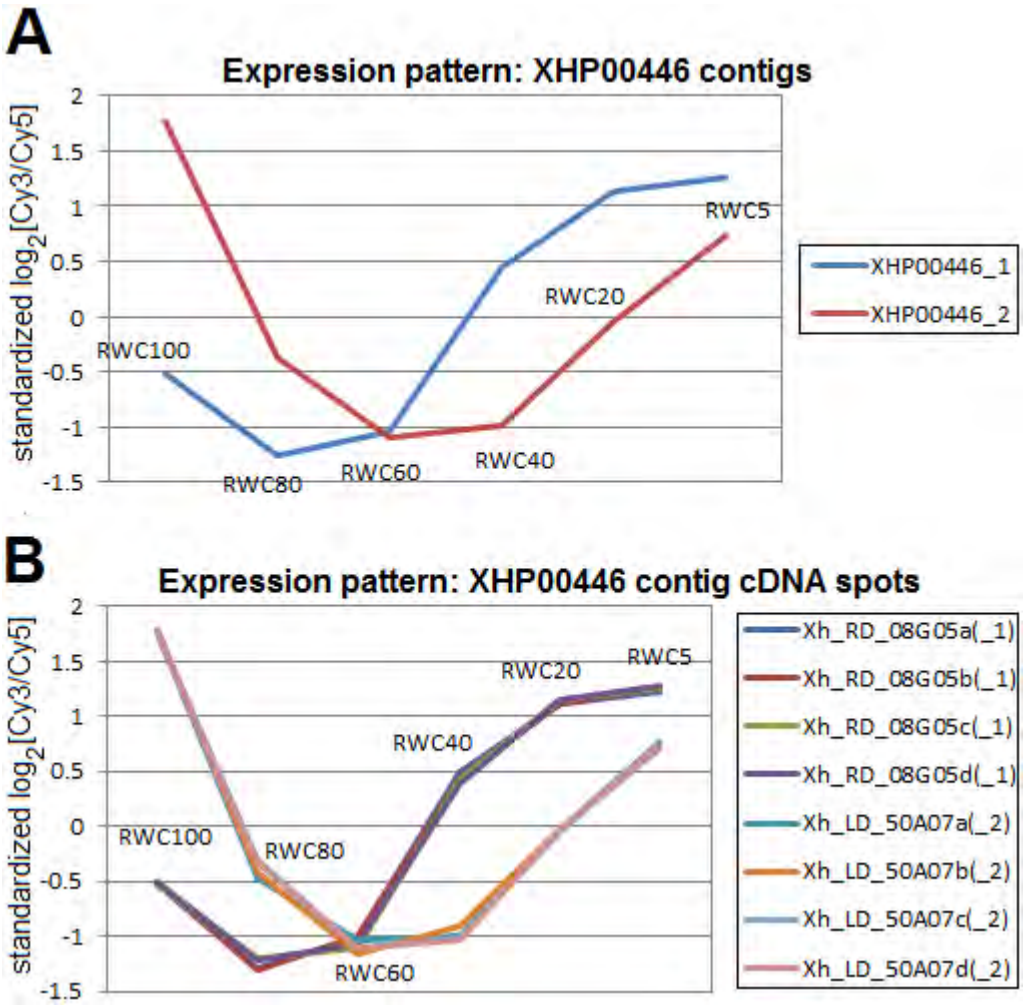


Figure 2.10. Expression pattern similarity between PHRAP contigs assembled in d2 cluster XHP00446. The normalized microarray log2 ratio values of clones (See Chapter 3) assembled in each of the contigs (A), or of arrayed cDNA spots of a clone (B), were extracted and averaged to represent expression pattern of each contig in *X. humilis* during desiccation. Expression patterns of all contigs or spots were standardized using GEPAS version 4.0 (http://gepas.bioinfo.cipf.es; Tárraga *et al.*, 2008) and plotted in Excel.

In general, of the 66 d2 cluster groups assessed, the 30 with mean variances lower than 0.5, such as XHP00030 (Fig. 2.2, Fig. 2.3, Fig. 2.4), showed good similarity in nucleotide and peptide sequence alignments, and in expression data, thus the contigs within each d2 cluster group were merged. In contrast, greater variations between contigs were observed in the 36 d2 clusters with mean variance greater than 0.5. The actions taken in the consolidation of the PHRAP contigs in the 66 assessed d2 clusters were summarized in Table 2.7.

Table 2.7. Summary of PHRAP contig consolidation assessment.

Mean variance	Comments on PHRAP contig assessment	No of d2 cluster
1.34E-04 to 0.50	Sequences aligned, expression patterns correlated. Contigs merged.	30
0.51 to 1.92	Sequences aligned, expression patterns correlated. Contigs merged.	5
	Predominant pattern determined, outliers identified and removed. Contigs merged.	12
	No additional information in determining predominant pattern. Contigs stayed separate.	17
	Data variation too large, unable to identify any patterns. d2 cluster and associated contigs removed.	2

Manual inspection of the aligned PHRAP contigs showed that majority of the 66 d2 clusters had been divided on the basis of minor differences in nucleotide sequences due to polymorphisms and differences in alternative splicing. Because the PHRAP contigs from the 30 d2 clusters assessed with mean variance less than 0.5 all showed high similarity in nucleotide sequence, peptide sequence, and expression pattern, it was decided that for the remainder 159 d2 clusters (accommodated with multiple PHRAP contigs) with lower mean variance (ranged from 2.04E-04 to 0.49), to merge the multiple contigs. Because the d2_cluster algorithm does not possess functions in deriving a consensus sequence, it was decided to select the consensus sequence from one PHRAP contig to be representative of each of the d2 clusters. PHRAP contig with the most informative Blast2GO annotation results was chosen to represent the d2 cluster. All sequences of clones within the d2 cluster group were manually assigned to the representative PHRAP contig, which was regarded as a unique gene of *X. humilis*.

As for the 17 d2 clusters in which their PHRAP contigs were decided to remain separate (Table 2.7), the clustering of the clone sequences stayed unchanged. Each PHRAP contig was represented by its original consensus sequence derived from PHRAP algorithm, and regarded as a unique gene.

A data set of 1775 contigs, representing the unique groups of 2709 clones, was generated after the manual consolidation of PHRAP contigs (Step 9, Fig. 2.1). The previously Blast2GO annotated data set of 2413 contigs (Step 8, Fig. 2.1) was updated accordingly (Step 10, Fig. 2.1).

2.3.4 Final set of annotated genes printed on the X. humilis microarray slides

The final set of contigs represented 1775 unique genes, of which, 1680 were present on the *X. humilis* microarray slide. 1268 of the 1680 contigs were successfully annotated in Blast2GO. The orthologues sharing high sequence similarity to the *X. humilis* peptide sequences came from a wide variety of plant species, with *Oryza sativa*, *Vitis vinifera* and *A. thaliana* recording the majority of hits (Fig. 2.11A). The majority of GO terms associated with these orthologues were inferred from electronic annotation (evidence code IEA), with information being transferred between database records on the basis of sequence similarity (Fig. 2.11B). GOSlim was used to consolidate and reduce the number of GO terms for the 1268 *X. humilis* contigs so that the distribution of GO terms could be simplified and visualized. GO terms associated with a diversity of biological processes (Fig. 2.12), molecular functions (Fig. 2.13) and cellular components (Fig. 2.14) were visualized in pie charts on the GO level where the highest number of GO terms was found annotated to the *X. humilis* contigs.

At GO level 4, a variety of biological processes were represented in the annotated set of 1268 *X. humilis* contigs printed on the microarray slide, with the majority of contigs annotated to macromolecule metabolic process (10.6%), cellular aromatic compound metabolic process (7.7%), cellular nitrogen compound metabolic process (7.7%), heterocycle metabolic process (7.7%), organic cyclic compound metabolic process (7.7%) and transport (6.6%) (Fig. 2.12A). These biological processes were also found to have similar representations in the annotated set of 57495 *A. thaliana* genes (Fig. 2.12B). Two terms represented in the *A. thaliana* dataset, regulation of localization and response to karrikin, were found absent in the *X. humilis*, albeit that they both had low representations (<0.1%) in *A. thaliana*.

The majority of the *X. humilis* contigs at GO level 3 were annotated to molecular functions of heterocyclic compound binding (22.5%), organic cyclic compound binding (22.5%), protein binding (13.1%), hydrolase activity (12.1%) and transferase activity (10.6%) (Fig. 2.13A). A similar distribution of these level 3 molecular function terms was also observed in the *A. thaliana* dataset (Fig. 2.13B). No *X. humilis* contig was found annotated to oxygen binding, which had only a 0.6% representation in *A. thaliana*.

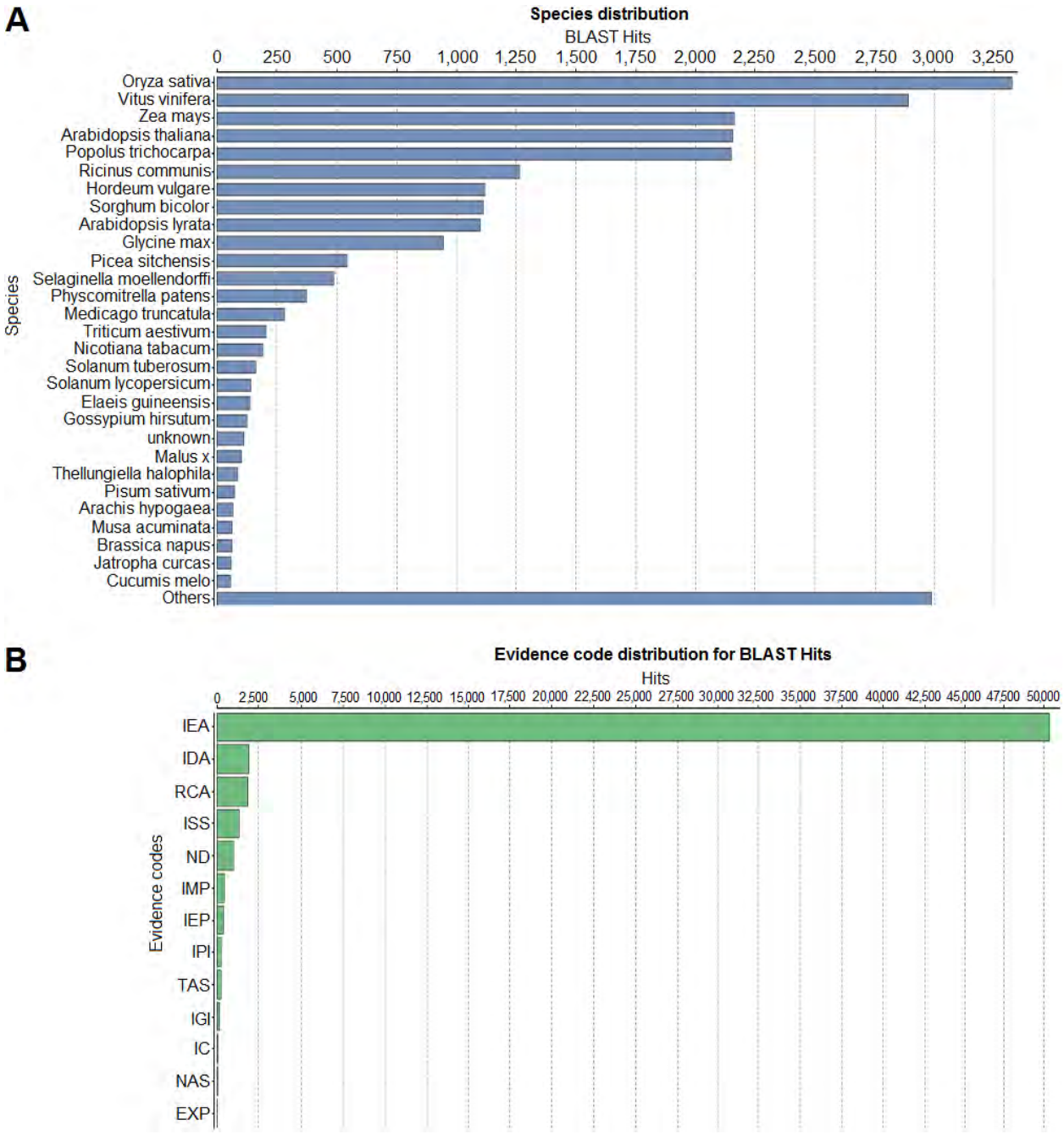


Figure 2.11. Annotation of *X. humilis* contigs. (A) Bar graph summarizing the species origins of orthologues sharing high sequence similarity to the 1268 annotated *X. humilis* peptide sequences identified by BLASTP search in Blast2GO. (B) Bar graph summarizing how the GO terms annotated for the orthologues sharing high sequence similarity to the 1268 annotated *X. humilis* peptide sequences were inferred. Evidence codes with a substantial number of hits included IEA (inferred from electronic annotation), IDA (inferred from direct assay), RCA (inferred from reviewed computational analysis) and ISS (inferred from sequence or structural similarity). A detailed description of all evidence codes can be found at http://www.geneontology.org/GO.evidence.shtml.

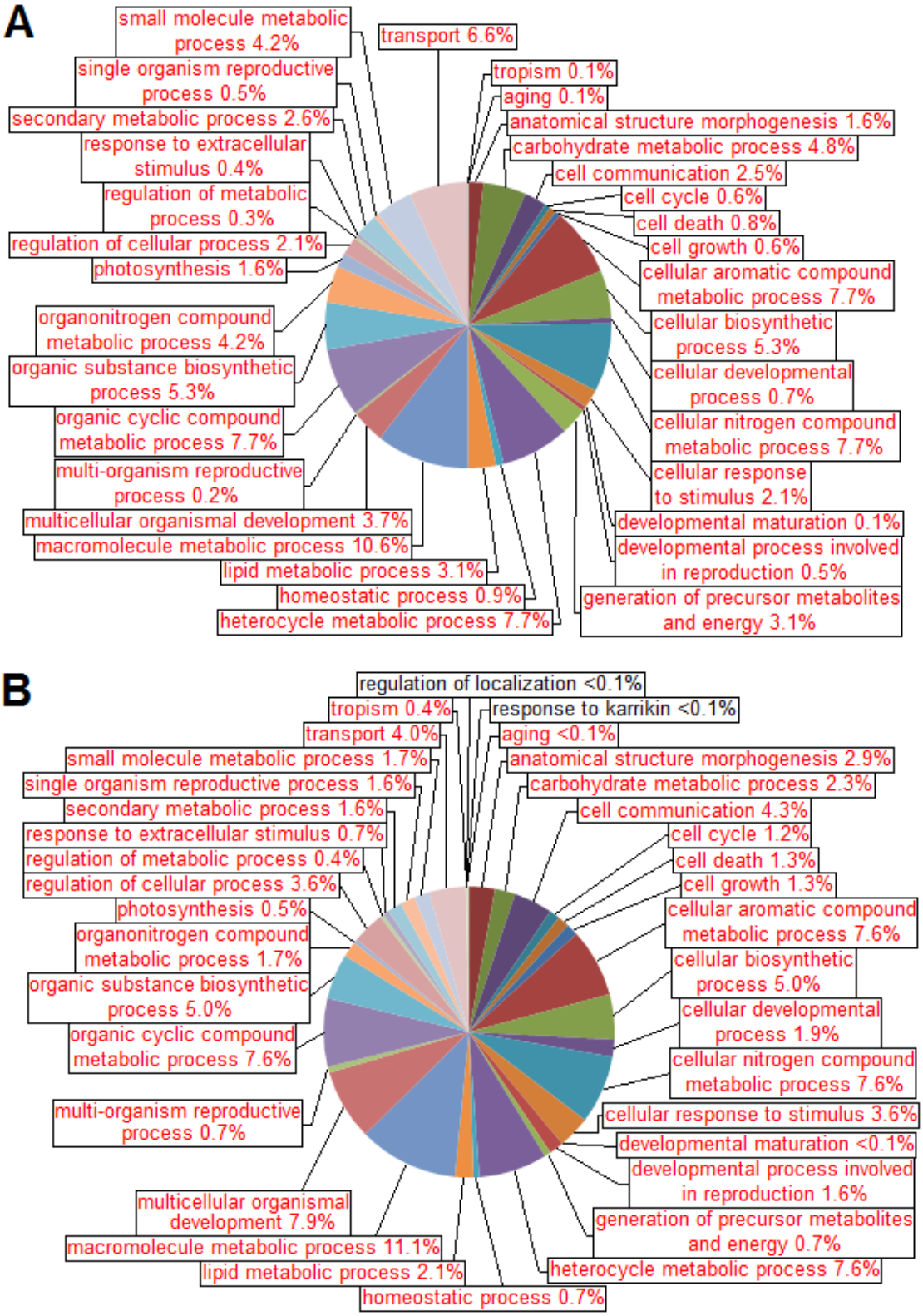


Figure 2.12. Biological process distribution of the annotated *X. humilis* and *A. thaliana* genes. Distribution of 1268 annotated *X. humilis* contigs (A) and 57495 *A. thaliana* genes (B) in the GO terms associated with a diversity of biological processes was compared. The GOSlim version of annotation was used for sequence distribution analysis. The pie chart was generated in Blast2GO under the combined graph analysis based on the level 4 GO terms. The GOSlim annotation data of *A. thaliana* genes was downloaded from B2G-FAR (version MAY2011; Götz *et al.*, 2011) at http://www.b2gfar.org/showspecies?species=3702. Terms identified and present in both plants were represented in red.

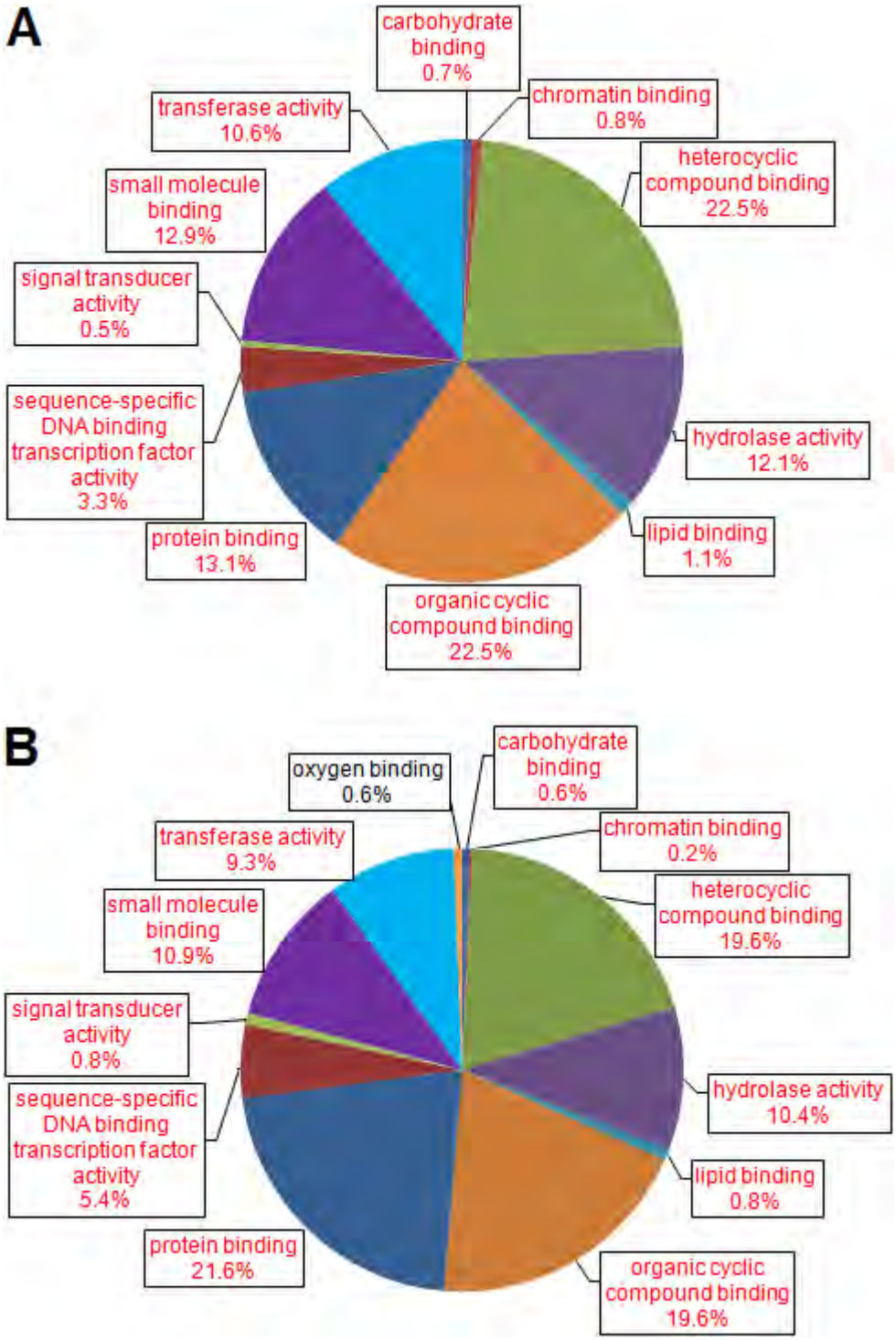


Figure 2.13. Molecular function distribution of the annotated *X. humilis* and *A. thaliana* genes. Distribution of 1268 annotated *X. humilis* contigs (A) and 57495 *A. thaliana* genes (B) in the GO terms associated with a diversity of molecular functions was compared. The GOSlim version of annotation was used for sequence distribution analysis. The pie chart was generated in Blast2GO under the combined graph analysis based on the level 3 GO terms. The GOSlim annotation data of *A. thaliana* genes was downloaded from B2G-FAR (version MAY2011; Götz *et al.*, 2011) at http://www.b2gfar.org/showspecies?species=3702. Terms identified and present in both plants were represented in red.

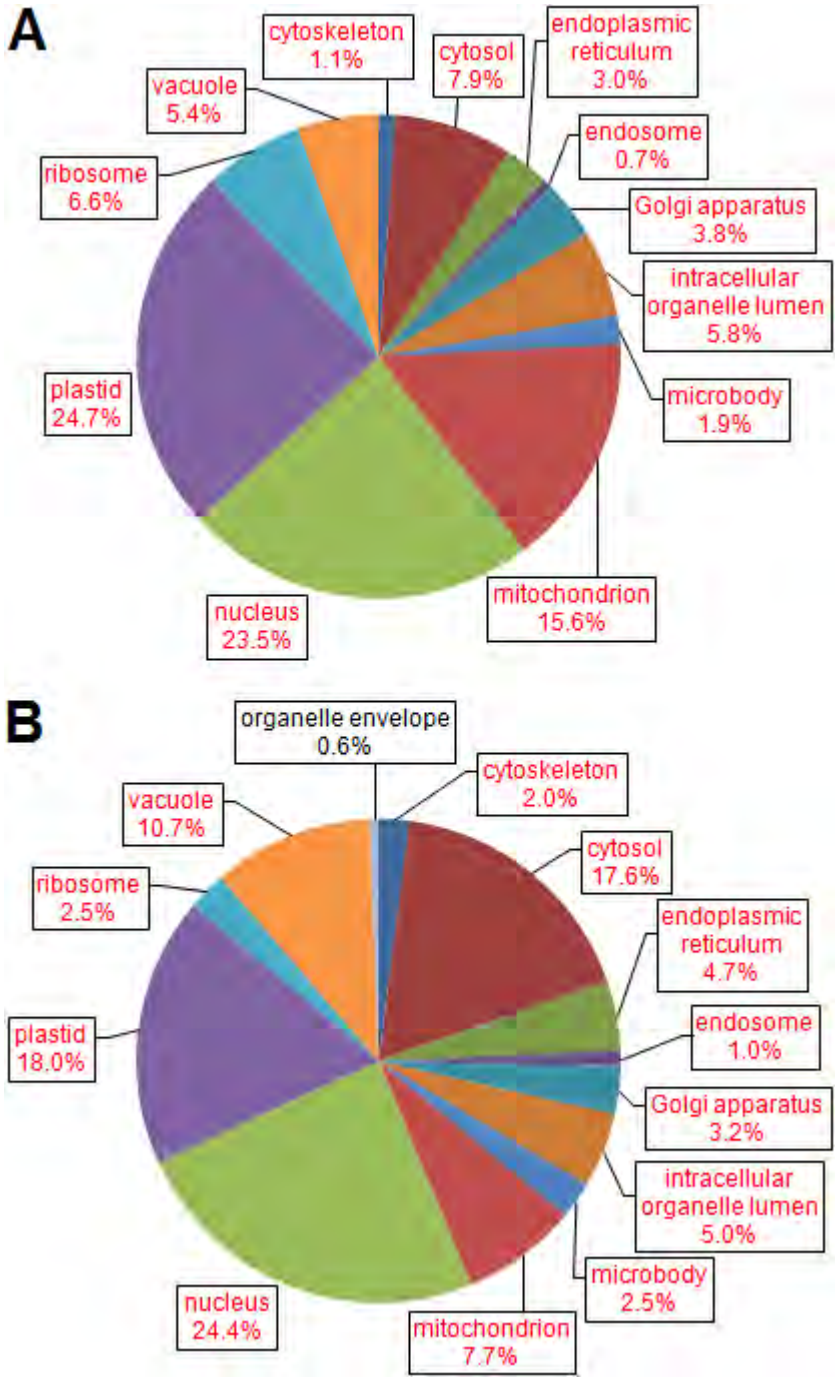


Figure 2.14. Cellular component distribution of the annotated *X. humilis* and *A. thaliana* genes. Distribution of 1268 annotated *X. humilis* contigs (A) and 57495 *A. thaliana* genes (B) in the GO terms associated with a diversity of cellular component was compared. The GOSlim version of annotation was used for sequence distribution analysis. The pie chart was generated in Blast2GO under the combined graph analysis based on the level 8 GO terms. The GOSlim annotation data of *A. thaliana* genes was downloaded from B2G-FAR (version MAY2011; Götz *et al.*, 2011) at http://www.b2gfar.org/showspecies?species=3702. Terms identified and present in both plants were represented in red.

Furthermore, most of the gene products encoded by the 1268 *X. humilis* contigs were suggested to be localized in plastid (24.7%), nucleus (23.5%), mitochondrion (15.6%) and cytosol (7.9%) (Fig. 2.14A). Although a larger proportion of cytosol localized (17.6%) and a lower proportion of mitochondrion localized genes (7.7%) were found in *A. thaliana*, the representations of the other level 8 cellular component terms remained similar in both datasets (Fig. 2.14B). No *X. humilis* contig was found annotated as organelle envelope localized genes, which comprised 0.6% of the *A. thaliana* genes.

The distribution of all 3 categories of GO terms showed that *X. humilis* microarray dataset shared similar representations of key terms with the *A. thaliana* protein dataset. This suggested that although the *X. humilis* microarray dataset is relatively small in size, but is not biased in the overall representation of biological functions.

2.3.5 Identification of LEAs, antioxidants and transcription factors in X. humilis

As the three prominent groups of genes being induced during abiotic stresses in plants, 48 contigs that were annotated as LEAs of different classes (Table 2.8), 24 as antioxidants (Table 2.9), and 93 as transcription factors (Table 2.10), were identified in the set of 1680 contigs represented on the *X. humilis* microarray slide. Expression of these 3 different groups of genes during desiccation in *X. humilis*, as well as during seed development and abiotic stress in *A. thaliana*, was specifically studied in addition to the analysis on the overall 1680 contigs, in Chapter 3.

Table 2.8. LEA contigs identified in *X. humilis*.

LEA class	No of contig	Contig ID	Contig description
1	1	XHP00221_2	embryonic abundant protein 1
2	10	XHP00016_2	RAB18 (RESPONSIVE TO ABA 18)
		XHP00109_4	RAB18 (RESPONSIVE TO ABA 18)
		XHP00182_6	dehydrin
		XHP00615_1	dehydrin-like protein dh2
		XHP00030_10	RAB18 (RESPONSIVE TO ABA 18)
		XHP00106_1	RAB18 (RESPONSIVE TO ABA 18)
		XHP00531_2	dhn9
		XHP00658_1	dehydrin
		XHP00966_1	RAB18 (RESPONSIVE TO ABA 18)
		XHP01481_1	RAB18 (RESPONSIVE TO ABA 18)
3	17	Xh_LDF_50H122	LEA protein, putative
		XHP00024_1	LEA 7
		XHP00062_6	late embryogenesis
		XHP00487_2	late embryogenesis
		XHP00641_1	LEA domain-containing protein
		XHP00011_1	late embryogenesis
		XHP00012_1	LEA domain-containing protein
		XHP00014_2	LEA protein, putative
		XHP00026_7	group 3 LEA protein
		XHP00055_2	LEA domain-containing protein
		XHP00089_3	group 3 LEA protein
		XHP00200_4	LEA domain-containing protein
		XHP00334_1	LEA domain-containing protein
		XHP00406_2	cre-lea-1 protein
		XHP01120_1	LEA family protein
		XHP01261_1	LEA domain-containing protein
		XHP00234_1	LEA protein, putative
4	5	XHP00028_2	LEA 4-5
		XHP00664_2	seed maturation protein
		XHP00665_1	LEA 18
		XHP01523_1	LEA group 1 domain-containing protein
		XHP00039_2	seed maturation protein
6	6	XHP00272_1	LEA protein, putative
		XHP00762_1	LEA protein, putative
		XHP01693_1	LEA protein, putative
		XHP01956_1_M	LEA protein, putative
		XHP00492_1	Seed maturation protein
		XHP00623_1	LEA protein, putative
7	3	XHP00066_1	drought-induced 21
		XHP00330_1	LEA protein
		XHP01037_1	ATDI21 (DROUGHT-INDUCED 21)
8	3	XHP00052_3	LEA14
		XHP01697_1	LEA 14
		XHP01570_1	LEA protein
10	3	XHP01703_1	AWPM-19-like membrane family protein
		XHP00054_3	AWPM-19-like membrane family protein
		XHP00105_1	AWPM-19-like family protein

Table 2.9. Antioxidant contigs identified in *X. humilis.*

Contig ID	Contig description
XHP00084_2	1-cysteine peroxiredoxin 1
XHP01001_1	2-Cys peroxiredoxin A
XHP00394_1	ascorbate peroxidase 2
XHP02029_1_M	ATGPX3 (GLUTATHIONE PEROXIDASE 3); glutathione peroxidase
XHP02074_1_M	catalase
XHP01613_1	catalase
XHP00362_1	copper/zinc superoxide dismutase 1
XHP01786_1	DC1 domain-containing protein
XHP02001_1_M	fructose-bisphosphate aldolase, putative
XHP01545_1	G6PD6 (GLUCOSE-6-PHOSPHATE DEHYDROGENASE 6)
XHP00994_1	glutathione peroxidase 1
XHP01166_1	glutathione peroxidase 3
XHP00604_3	glutathione peroxidase 6
XHP00257_1	glutathione S-transferase TAU 19
XHP01640_1	glutathione transferase
XHP00591_1	microsomal glutathione s-transferase, putative
XHP00471_1	MSRB2 (methionine sulfoxide reductase B 2)
XHP00763_1	NQR; binding / catalytic/ oxidoreductase/ zinc ion binding
XHP01774_1	peroxidase 52
XHP00810_1	Peroxidase superfamily protein
XHP01983_1_M	peroxiredoxin type 2, putative
XHP01155_1	thioredoxin H-type 1
XHP00107_1	Thioredoxin superfamily protein
XHP01781_1	thioredoxin-dependent peroxidase 1

Table 2.10. Transcription factor contigs identified in *X. humilis.*

Contig ID	Contig description
XHP01405_1	AGAMOUS-like 19, MADS-box transcription factor
XHP00496_2	anac083 transcription factor
XHP01060_1	ap2 domain containing protein
XHP00635_2	Arabidopsis 6B-interacting protein 1-like 2
XHP01675_1	ATBZIP53 (BASIC REGION/LEUCINE ZIPPER MOTIF 53); transcription factor
XHP00496_1	ATNAC2 (NAC DOMAIN CONTAINING PROTEIN 2); transcription factor
XHP01176_1	Basic helix-loop-helix (bHLH) DNA-binding family protein
XHP01425_1	basic helix-loop-helix (bHLH) family protein
XHP00074_1	basic region/leucine zipper motif 53
XHP00339_1	Basic-leucine zipper (bZIP) transcription factor family protein
XHP00449_2	B-Box domian protein 24
XHP00580_1	BEL1-like homeodomain 7, BLH7
XHP01446_1	BolA-like family protein
XHP01783_1	BT2 (BTB AND TAZ DOMAIN PROTEIN 2
Xh_LDR_10D023	cbf-like transcription factor
XHP01827_1	CCR4-NOT transcription complex protein, putative
XHP02014_1_M	class iii homeodomain-leucine zipper protein c3hdz1
XHP00086_1	Cox19-like CHCH family protein
XHP01047_1	cytokinin response factor 2
XHP01193_1	DDB1A (DAMAGED DNA BINDING PROTEIN 1A)
XHP01502_1	DNA binding / DNA-directed RNA polymerase

Table 2.10. (continued)

Contig ID	Contig description
XHP00761_1	DNA binding protein
XHP01825_1	DNA-binding storekeeper protein-related transcriptional regulator
XHP00877_1	EER4 (ENHANCED ETHYLENE RESPONSE 4); transcription initiation factor
XHP00400_2	ERF domain protein 11
XHP00447_2	ERF4 (ETHYLENE RESPONSIVE ELEMENT BINDING FACTOR 4); transcription repressor
XHP01410_1	ethylene-insensitive 3f
XHP01164_1	ETHYLENE-INSENSITIVE3
XHP01021_1	ethylene-responsive transcription factor 3
XHP00408_1	FPA; RNA binding
XHP02064_1_M	g-box binding factor
XHP00288_2	G-box binding factor 3
XHP00297_2	general regulatory factor 7
XHP01506_1	GRAS family transcription factor
XHP00299_1	GRAS2, SCARECROW-like 14
XHP00799_1	HAIRY MERISTEM 3
XHP01128_1	heat shock factor protein 1
XHP01426_1	heat shock factor protein hsf30
XHP01296_1	heat shock transcription factor C1
XHP00578_1	high mobility group B3
XHP01271_1	Histone-fold/TFIID-TAF/NF-Y
XHP01717_1	homeobox 1
XHP01475_1	iaa-leucine resistant3, ILR3
XHP01302_1	ILR3 (iaa-leucine resistant3); DNA binding / transcription factor
Xh_LDR_50A043	indoleacetic acid-induced protein 10
XHP01979_1_M	KIWI; DNA binding / protein binding / transcription coactivator
Xh_LDR_12F103	KNOTTED1-like homeobox gene 3
Xh_LRF_02G038	LIL3:1; transcription factor
XHP00783_1	LUG (LEUNIG); transcription repressor
XHP00328_1	mads-box transcription factor 15
XHP00684_1	MBF1B (MULTIPROTEIN BRIDGING FACTOR 1B); transcription coactivator
XHP02058_1_M	MSI4/FVE
XHP00734_1	myb domain protein 73
XHP01723_1	myb family transcription factor
XHP01785_1	myb family transcription factor
XHP00217_1	myb-like transcription factor family protein
XHP00577_1	NAC 014
Xh_LDR_51E052	NAC transcription factor
XHP00975_1	NAC transcription factor
XHP00949_1	NAD(P)-binding Rossmann-fold superfamily protein
XHP00402_1	NF-YA7
XHP00350_1	NF-YB3
XHP01329_1	NF-YC3
XHP01702_1	NRP2 (NAP1-RELATED PROTEIN 2); histone binding
XHP02004_1_M	origin of replication complex 1B
XHP01959_1_M	Oxidation-related Zinc Finger 1
XHP00775_1	PICKLE
Xh_LRF_01G108	plastid transcriptionally active 14
XHP01067_1	plastid transcriptionally active 6

Table 2.10. (continued)

Contig ID	Contig description
XHP01085_1	redox responsive transcription factor 1
XHP01134_1	sensitive to proton rhizotoxicity 1
XHP01087_1	serine acetyltransferase 2;1
XHP01127_1	single myb histone 6
XHP00286_1	stress enhanced protein 1
XHP00854_1	SWINGER
XHP01184_1	SZF1 (SALT-INDUCIBLE ZINC FINGER 1); transcription factor
XHP01131_1	TPL (TOPLESS); transcription repressor
XHP01696_1	transcription factor
XHP01538_1	transcription factor-related
XHP01022_1	transcription repressor
XHP01039_1	transcriptional regulator family protein
Xh_LDF_05G063	tubby like protein 3
Xh_LDR_51B02r	WRKY DNA-binding protein 23
XHP00630_2	WRKY DNA-binding protein 33
XHP01741_1	WRKY DNA-binding protein 51
XHP00930_1	WRKY DNA-binding protein 70
XHP00887_2	wrky transcription factor 27
XHP00630_1	WRKY33; transcription factor
XHP01046_1	YABBY2
XHP01231_1	zinc finger (C2H2 type) family protein
XHP00948_1	zinc-finger protein 2
XHP01277_1	ZML1 (ZIM-LIKE 1)
XHP00323_1	ZML1 (ZIM-LIKE 1); transcription factor

2.4. Summary

This annotation pipeline was used to map the 3105 cDNA clones printed on the microarray slide to 1680 unique *X. humilis* genes. Among which, 1268 were annotated by Blast2GO. This information was used to analyze a microarray dataset generated from desiccating *X. humilis* leaves, and to determine the functional importance of genes which are differentially regulated in response to desiccation in *X. humilis*.

Chapter 3
Microarray Gene Expression Data Analysis

3.1 Introduction

3.1.1 What is a microarray?

DNA microarray technology was invented in Stanford (Schena *et al.*, 1995; Reymond, 2001), a revolutionary technology which transforms the research of monitoring gene expression patterns from the conventional single-gene approach to allowing analysis of thousands or even tens of thousands of genes at once in a single experiment. Coupled with bioinformatics tools, microarray extends the possibilities in the field of molecular biology research (Wu. *et al.*, 2001).

A microarray is typically a glass or silicone slide, onto which DNA molecules, are arrayed and immobilized. These arrayed DNA molecules are normally referred to as spots, probes or features, and usually correspond to a single mRNA molecule, or transcript, in a genome (Causton *et al.*, 2003). For model organisms like *A. thaliana* were sequencing information are available, these features are normally synthesized and cover most of the genome. For non-model organisms like *X. humilis*, features normally consist of amplified inserts from cDNA libraries. This type of array is sometimes referred to as a boutique array where it represents only a portion of the expressed genome.

There are two main platforms of microarrays available, in which the sequences representing different genes, or features, are either directly synthesized or spotted onto the chips or slides. For example, Affymetrix arrays are constructed by synthesizing short oligonucleotide probes (25-mer) directly on microarray chips (Affymetrix, Inc., Santa Clara, CA, USA). Alternatively, long oligonucleotide probes (± 60-mers, Agilent Technologies, Inc., Santa Clara, CA, USA) or cDNA inserts can be directly spotted onto glass slides using ink-jet printing technology. In order to measure mRNA transcript abundance, these probes are hybridized with fluorescent dye labeled cDNAs, converted from RNA extracted from the tissues of interest. Non-specifically hybridized products such as labeled cDNAs binding onto non-feature regions of slide, or mismatched labeled cDNA-probe complexes, are then removed by high stringency washes. Thereafter, the hybridized spots are excited by laser and the fluorescent signals associated with each spot are captured. These readings correspond to

the amount of labeled cDNAs bound to each feature, which in turn provides a direct estimate of the levels of gene expression in the tissue of interest.

There are two approaches for generating microarray data. In Affymetrix arrays, the data is generated by single channel approach only, i.e. probes of each array or slide, are only hybridized to cDNAs labeled with one type of fluorescent dye, which usually represent transcripts isolated from a particular tissue of interest. In this approach, comparisons of gene expression levels in different biological samples are made between the data generated from different arrays or hybridizations. In spotted arrays, data can be generated by either single channel or dual channel approach, i.e. the spots are simultaneously hybridized to two groups of cDNA samples, each labeled with different fluorescent dyes. In this approach, comparisons of gene expression levels in two biological samples can be made within a single array or hybridization. The ratio of fluorescent readings between the two dyes of a spot simply indicates the relative abundance of that gene in the two biological samples tested (Quackenbush, 2001).

3.1.2 An overview of a microarray experiment

A microarray experiment usually consists of several phases (Causton *et al.*, 2003):

1. *Hypotheses and experimental design*. Here, hypotheses and objectives of a study are clearly defined, and experiments are carefully planned in details.

2. *Material processing and data collection*. This phase includes (i) array fabrication; (ii) preparation of the biological samples to be studied; (iii) extraction and labeling of the RNA from the samples; (iv) hybridization of the labeled extracts to the array information processing; and (v) scanning of the hybridized array.

3. *Information processing*. This phase starts after image of an array has been generated by scanning. The phase includes (vi) image quantitation. In which each spot is precisely located by predefined grid and its fluorescent intensity is measured; (vii) data normalization and integration. Data from different arrays is visualized to identify technical variation prior to applying an appropriate normalization method to remove it. Data from different arrays are then scaled to a comparable level; (viii) gene expression data analysis. Here differentially expressed genes are identified, and are clustered according to their expression patterns. Tests for enrichment can be used to

investigate whether any specific biological, cellular or molecular functions are under- or over-represented in genes showing similar patterns of expression.

4. *Confirmation*. For a high throughput experiment, no matter how cautious one is to minimize technical variation at the different stages of a microarray experiment, noise can still remain in the final datasets. For example, inappropriate normalization methods can introduce bias into the data, therefore, it is important to confirm the microarray results generated with alternative techniques. Microarrays are normally verified by quantitative real time PCR experiments (Weeraratna and Taub, 2007).

When designing a microarray experiment, one should carefully consider the factors in each of these stages which could have negative impacts on the final results. Randomization is by far the most important issue in microarray experiments. Random assignment of samples and random sampling of population are the physical basis for the validity of statistical inferences (Page *et al.*, 2007; Rubin, 1991). Random assignment of samples simply denotes that the simultaneous processing of all hybridizations from one experimental condition in one batch, should be avoided. The printing of cDNAs onto slides also requires randomization, and closely related cDNAs, cDNAs from a common pathway, or replicates of cDNAs should not be located close to each other. Printing of multiple copies of one cDNA at different locations on a slide is useful for identifying technical noise introduced during the course of experiment on location basis

Biological repeats are also important to take variability between experimental samples into account. The number of biological replicates for each experimental condition is often limited due to the high expense of microarray experiments. A requirement of at least 3 independent biological replicates per experimental condition has been suggested for statistical analysis (Lee *et al.*, 2000).

Another important issue to consider is the amount of total RNA required for each microarray hybridization. A typical microarray experiment requires 5 to 20 μg of total RNA per sample for labeling. What if it is not possible to extract enough RNA from a single sample? One of the solutions is to pool RNA samples extracted from different individuals to reach the required amount. However, pooling of RNA samples may result in the loss of important information from individual samples. An alternative approach to consider is linear

amplification of RNA to generate the required amount of RNA without influencing the representation of the mRNA transcript population (Van Gelder *et al.*, 1990).

3.1.3 Microarray data visualization and normalization

Various sources of random and systematic technical variation introduced during microarray experiments can mask the accurate measurement of gene expression levels, and need to be removed in a process called normalization. For example, variation in the signal detected between experiments can arise from differences in labeling, hybridization or data capturing, rather than from the biology of the samples. The identification and removal of technical variations prior to the analysis of microarray data is crucial. In microarray experiments, these systematic technical biases can be quantified by visualizing the raw data by means of various diagnostic plots such as scatter plots (Drãghici, 2003; Stekel D, 2003), histograms (Drãghici, 2003), box plots (Drãghici, 2003) or false-colour plots (Stekel, 2003; Wit and McClure, 2004).

Background noise is defined as the systematic bias resulting from the non-specific hybridization of the fluorescent dye labeled samples and false targets such as the slide surface, dust particles in close proximity to the features, or natural fluorescence of the glass or its coating. Background bias in general leads to an inaccurate over-estimation of fluorescent signal abundance for specific features (van Heerden *et al.*, 2007).

An intensity-dependent dye effect bias is a result of differences in the physical properties of the commonly used Cy3 and Cy5 dyes, which can induce a systematic bias in the overall labeling and hybridization efficiency. Dye bias can be easily visualized by plotting Cy3 data from an array on one axis, and Cy5 data on the other axis in scatter plot. An overall data distribution which deviates substantially from a line $y = x$, normally suggests the presence of such bias.

A spatial bias in the dataset can result from the uneven hybridization, washing, or scanning of different parts of a slide, such that features within a certain part of slide behave similarly i.e. high/low fluorescent signal abundance, or high/low Cy5 to Cy3 log ratio values. Such type of bias can be easily detected by false-colour plots. Another type of spatial bias comes from the inconsistent depositing of cDNA samples onto slides by each printing pin. Print-tip bias can be assessed by plotting the signals from different printing pins within a slide on a box-plot.

Data from a particular print-tip which shows a unique pattern different to data derived from the other print-tips, is often an indication of print-tip bias. However, bias due to print tips may be overlooked if the ratios of intensity are being analyzed (Reilly, 2009).

The above mentioned biases affect the measurement of gene expression levels within an array. Differences in mRNA concentration, sample labeling and hybridization efficiency, and slide scanning for each array can induce a scale bias which affects the spread of data between arrays. These scale difference can be assessed by plotting data from individual arrays in box plots.

Normalization is the process of removing such systematic variation in the datasets, while preserving the biological information. Normalization has two major purposes: (1) to correct for effects that arise from technical variation within arrays; and (2) to scale the data between arrays to a comparable level for subsequent analysis such as identification of differentially expressed genes (Smyth and Speed, 2003; Page *et al.*, 2007). Numerous normalization methods have been developed for microarray analysis, and each algorithm is designed to deal with specific systematic errors detected in the data. For example, different methods are used to correct for background noise, dye effect and spatial biases during within-array normalizations.

Background noise can be measured and removed by a number of methods such as local estimation and subtraction. Pixel intensities surrounding each defined feature are first determined, from which an average or median background level is calculated, which can then be subtracted from the feature intensities. In global background subtraction, a single measurement representing average or median background level across the entire slide is subtracted from the intensities of each feature. Often, background levels can also be determined and subtracted based on signal intensities of blank or negative control spots. However, it has been cautioned that various background subtraction methods introduce more noise, (Khojasteh *et al.*, 2005; van Heerden *et al.*, 2007), and the reliability of background subtraction is still debated in the literature.

Many normalization procedures rely on common assumptions such as (1) the expression of the majority of genes analyzed is unaltered between arrays, and (2) a relatively symmetrical distribution can be expected among the up- and the down-regulated genes identified (Calza

and Pawitan, 2010). Linear regression was one of the first algorithms used in early microarray experiments to correct for intensity-dependent dye biases. This method assumes that the majority of genes arrayed on slides are expressed at similar levels. Thus, if labeling and detection efficiencies for both Cy3 and Cy5 samples are equal, the majority of genes are expected to cluster along a straight line of slope = 1 in a scatter plot of logarithmic intensities from both samples. Linear regression determines the best fit slope of the raw data, then adjusts data values so that the slope is corrected to 1 (Causton *et al.*, 2003). However this approach is generally no longer used because it has been suggested that the intensity-dependent dye effect is non-linear (Stekel, 2003; van Heerden *et al.*, 2007). Numerous non-linear methods have subsequently been developed. Lo(w)ess (locally weighted polynomial regression), has been proposed as a normalization method to correct for intensity-dependent dye effects. During the analysis, measured $\log_2$(Cy5/Cy3) ratios for each spot are plotted against $\log_2$(Cy5*Cy3) product intensities. This MA plot often reflects intensity-specific artefacts in the measurement of the ratio, which tend to occur at weakly or extreme strongly fluorescing spots (Yang *et al.*, 2001). Because the majority of the genes are assumed to be constantly expressed, the overall $\log_2$(Cy5/Cy3) ratios of the data will be 0, independent of intensity, and data points are expected to cluster along the straight line $y = 0$. Lo(w)ess detects deviations from the expected behaviour and performs a series of local regressions across the MA plot for each data point and locally corrects the deviations of each data point.

Lo(w)ess-based normalization methods can either be applied globally to the entire dataset across the whole slide, or locally to subsets of data within each block or print-tip group on the slide. Global Lo(w)ess assumes all features on slide behave similarly, and normalizes intensity-dependent dye effects by correcting all data points on the slide. Print-tip Lo(w)ess is developed to correct for dye inconsistencies specifically caused by differences in the printing pins (van Heerden *et al.*, 2007). Unlike global Lo(w)ess, this algorithm assumes features in different blocks or print-tip groups behave differently, and normalizes the subsets of spots within each block or print-tip group. Nevertheless, it has been suggested that a minimum data size of 150 features per each print-tip group is required, for an effective and reliable Print-tip Lo(w)ess normalization (van Heerden *et al.*, 2007). Often, in addition to quality assessment, control features or known house-keeping genes on a slide can also be employed for normalization.

Spatial effects often occurs within a region on a slide, therefore local approach of normalization is generally more applicable than global approach. 2D Lo(w)ess, another Lo(w)ess-based method, can be applied to effectively remove continuous spatial trends. The algorithm normalizes data based on the x- and y-coordinates of each feature within two-dimensional false-colour plots (Smyth and Speed, 2003; van Heerden *et al.*, 2007). An alternative spatial bias correcting method, the median normalization, which is based on the central tendency of neighbourhoods of spots, is also available in the Bioconductor package, "marray" (Yang *et al.*, 2002). For each spot, the median of $\log_2$ values of spots within a predefined area (neighbourhood) centred on that spot, is calculated, the value of each spot is then adjusted accordingly (Smyth and Speed, 2003; van Heerden *et al.*, 2007; Schmidt *et al*., 2011).

Each slide or array possesses different data distribution ranges or scales due to different efficiencies in sample labeling, hybridization and scanning in each slide. In order to facilitate direct comparisons between different slides, these scale differences need to be corrected, so that the data from different arrays are positioned on the same scale. Most scale bias normalization algorithms involves adjusting the means and the spread of data to be similar (van Heerden *et al.*, 2007). Quantile normalization ensures equivalent distribution of intensity values between slides by first rank-ordering the signals from all spots in each array according to their intensity values. The ranked distributions across arrays are then compared to generate a mean value for each ranked distribution. This calculated mean then replaces the original value, and the normalized data are then rearranged back to the original ordering (Bolstad *et al*., 2003; Wernisch *et al*., 2003; van Heerden *et al.*, 2007). Quantile normalization forces the data distribution in each array across different arrays to be the same, and it is thus important to ensure that all arrays have similar data distribution patterns before applying quantile normalization to avoid introducing an artificial bias. Channel-specific implementation of quantile based algorithm is also available. This method is modified specifically for dual-channel arrays with the use of common reference in either one of the channels across all arrays (Smyth, 2005). Such modified quantile algorithm forces the distribution of reference channel data across all arrays to be the same (theoretically, they should to be identical), then adjusts the data of the other channel in each array accordingly.

With the increase in the application of microarray-based techniques, a broad spectrum of programs devoted to microarray data analysis is widely available (Schmidt *et* al., 2011).

Among them, Bioconductor operated in R environment is one of the most commonly used (Reimers and Carey, 2006; Zhang *et al*., 2009), and several normalization methods have been developed into software packages for use in R. Each algorithm is based on certain underlying assumptions on the data being normalized. Normalization changes data, so the normalization strategy chosen needs to be appropriately applied to particular data to correct for particular systematic biases (Schmidt *et* al., 2011). It is essential for one to understand both the basic principles of each normalization algorithm, and the nature of the data to be normalized, before employing any normalization approach.

3.1.4 Identification of differentially expressed genes

The major goal of most microarray experiments is to identify genes whose expressions show significant changes across different experimental conditions, with the rationale that these genes might be functionally important in the specific comparison. Many early microarray experiments identified differentially expressed genes by arbitrarily assigning a fold-change threshold, usually 2-fold, and considered genes above this threshold as significant. This approach is simplistic as it does not take into account the biological variation between experimental samples, nor does it consider genes with smaller changes in expression which might be important. Thus genes which show reproducible changes in expression of less than 2-fold with little variance, are excluded from the list, whereas genes which show expression changes of more than 2-fold, but with huge data variation, are prioritized despite their changes in expression not being statistically significant. Nowadays, the fold-change approach of identifying differentially expressed genes has been replaced by methods which use the standard deviations or variances to identify genes which show a statistically significant change in gene expression (Page *et al.*, 2007). In a typical hypothesis test, a probabilistic model is built under the null hypothesis (H_0), usually assuming a gene is not differentially expressed. The p-value, a probability is then calculated to reflect the likelihood of the null hypothesis being true given the observed data. In other words, the smaller the p-value, the less likely the observed data will agree with the null hypothesis. In this case, the null hypothesis is rejected and the alternative hypothesis (H_a), usually hypothesizing that a gene is differentially expressed, is then accepted. In biological studies, a cutoff of p-value ≤ 0.05 is commonly accepted, implying a > 95% chance that a gene's mRNA transcript abundance is different between samples (Quackenbush, 2005; Mohapatra and Krishnan, 2011). Classical parametric tests such as the one-sample t-test and two-sample t-test, and non-parametric approaches such as the Wilcoxon sign-rank test and the Mann-Whitney test, have been

widely adopted for microarray data comparing mRNA transcript abundance between two samples. With the increasing use of microarrays to perform more complex experiments, for example where the number of experimental conditions may be more than two, a more sophisticated analyses such as analysis of variance (ANOVA) (Kerr *et al.*, 2000) or general linear models are required (Edward, 2007).

In a typical microarray expression dataset, the total number of samples is always unavoidably less than the total number of genes under investigation, i.e. tens of thousands of genes versus 2 to 10 biological replicates. Performing statistical tests on such a large number of genes in parallel possesses a serious problem, as large numbers of false positives are generated. In other words, genes which have been identified as being differentially expressed based on their p-values, are in fact not truly differentially expressed (a type I error). To overcome this problem of false discovery, the p-values need to be adjusted to account for multiple testing. One approach to correct for multiple testing is to control the family-wise error rate (FWER), which reflects the probability of accumulating one or more false positive errors over a number of statistical tests. The Bonferroni method is a classical FWER control method, which simply adjusts each p-value by multiplying it by the total number of genes, or by resetting the significance level by dividing it by the total number of genes (Stekel, 2003; Cui and Churchill, 2003). The problem with the Bonferroni adjustment is that it is usually too stringent for microarray analysis with large number of genes being tested. Thus, although the Bonferroni adjustment reduces the number of false positives, it simultaneously increases the number of false negatives (a type II error). An alternative approach proposed by Benjamini and Hochberg (1995), is to control the false discovery rate (FDR), which reflects the expected or acceptable proportion of false discoveries, among the initial list of genes identified as differentially expressed. Often times, with microarray datasets comprising large number of genes with very few biological replicates, multiple testing correction by FWER or FDR results in hardly any differentially expressed genes being identified, suggesting that these methods, although statistically correct, are often very conservative and too stringent in practice (Datta and Datta, 2005). The empirical Bayes method has been shown to be an effective alternative approach for handling microarray datasets (Efron *et al.*, 2001; Scott and Berger, 2010). The empirical Bayes model estimates an *a posteriori* probability by marginal maximum likelihood, also known as type II (FDR) maximum likelihood, with minimum prior assumptions. This estimated probability is then used to determine the posterior probabilities in the multiple tested data.

Several software packages with statistical tests and multiple testing corrections have been developed for identification of differentially expressed genes in microarray data. For example, Significance Analysis of Microarrays (SAM, Tusher *et al*., 2001) incorporates an adjusted t-test modified to correct for multiple testing problems, which can be used for the comparison of two experimental conditions. Linear models for microarray data (Limma) is a software package developed in R, designed for differential expression analysis of microarray expression datasets from more than two conditions. It allows the fitting of a linear model to the expression data for each gene, and uses the empirical Bayes method with correction for multiple testing, to identify differential expressed genes (Smyth, 2004; 2005). Similar to problem of normalization, the decision on which statistical method to use to identify differentially expressed genes depends on the intrinsic properties of the data being analyzed. Recently a new adaptive procedure named ROTS (Reproducibility-Optimized Test Statistics) has been introduced (Elo *et al*., 2008; 2009), which learns an optimal statistic directly from a given data without any *a priori* knowledge required about the properties of the data, enabling the identification of differentially expressed genes via optimal statistics.

3.1.5 Challenges of analyzing X. humilis microarray data

The major objective of this project was to examine the global changes of mRNA transcript abundance in *X. humilis* leaves during desiccation via microarray technology, with the specific aims of (1) characterizing the changes in mRNA transcript abundance in leaf tissue at six different stages of water loss (100%, 80%, 60%, 40%, 20% and 5% RWC); (2) identifying different temporal and functional classes of genes that are activated or repressed during desiccation; and (3) testing the hypothesis that vegetative desiccation in *X. humilis* evolved from the activation of seed-specific late maturation transcription factors in response to abiotic stress. Setting this project up posed a number of challenges, including the *de novo* printing of microarray slides with cDNAs representing the *X. humilis* transcriptome, deciding on which normalization methods to use to correct for technical errors, and which statistical tests to use to identify differentially expressed genes.

3.2 Material and Methods

3.2.1 Microarray slide preparation

3.2.1.1 cDNA PCR Amplification

The inserts of 3105 *X. humilis* library clones (1450 clones from LD cDNA library; 1453 clones from RD cDNA library; 107 clones from LR cDNA library; and 95 clones from RR cDNA library) (Collett *et al.*, 2004) corresponding to 1709 unique cDNAs, were PCR amplified in Corning Thermowell 96-well plates (Corning Incorporated, NY, USA). For each cDNA clone, 1 µl of 25% glycerol stock culture was added to 100 µl of a PCR mixture containing 0.3 µM T3 primer, 0.3 µM T7 primer, 200 µM dNTPs, 2.5 mM MgCl2, 1X PCR reaction buffer and 0.1 µl (5U/µl) Super-therm polymerase (Southern Cross Biotech, South Africa). PCR was performed as follows: at 94 °C for 1 min; for 10 cycles at 94 °C for 20 sec, 56 °C for 20 sec, 72 °C for 2 min; for 30 cycles at 94 °C for 20 sec, 51 °C for 20 sec, 72 °C for 2 min; and at 72 °C for 7 min. Three microlitres of each PCR product was electrophoresed on a 0.8% agarose gel containing 100 ng/ml ethidium bromide for a quality check prior to purification.

3.2.1.2 cDNA PCR purification

The PCR reactions were transferred to Millipore 96-well Multiscreen PCR plates (Millipore, MA, USA) fitted on a vacuum apparatus, and allowed to pass through the filters for 10 minutes. Filters were then washed twice with 100 µl H_2O. PCR products were resuspended in 50 µl 50% DMSO with vigorous agitation for 60 minutes. Twenty microlitres of which were pipette-transferred into 384-well plates (Amersham, Germany) for slide printing. The remaining 30 µl was transferred and stored in clean Nunc 96 well U-bottomed plates (Amersham, Germany) and stored at -20 °C. Three microlitres of purified PCR products were electrophoresed on a 0.8% agarose gel containing 100 ng/ml ethidium bromide for a quality check prior to printing.

3.2.1.3 cDNA printing

The purified cDNAs, together with Lucidea controls (Amersham Biosciences, Germany) were printed onto Corning GAPS II coated slides (Corning Incorporated, NY, USA) by a MicroGrid II Arrayer (BioRobotics, Cambridge, UK) with 16 10 µm split pins in a 4x8 configuration following manufacturer's instructions. The printed slides were UV-crosslinked (900 mJ cm2), and baked at 80 °C for 2 hours. Each printed block was designed to contain a random, unbiased mixture of cDNAs originated from the LD, RD, LR and RR libraries.

3.2.2 Plant material, treatment and sample harvesting

X. humilis plants collected from the Borakalalo Game Reserve, North-West Province, South Africa, were supplied by Professor Jill Farrant from the Department of Molecular and Cell Biology (MCB), University of Cape Town. The plants were maintained in trays with good drainage under natural conditions in glasshouse at the University of Cape Town, with temperatures varied between 10 and 24 °C, and daylight intensities varied between 26.9 and 1211.1 μmol. m^{-2}. s^{-1}. Plants were watered every second day for two weeks, prior to the commencement of the desiccation treatment. A night before the start of the desiccation treatment, plants were watered thoroughly, and covered with plastic bags. The plastic covers were removed the following day and plants were allowed to dry under natural weather conditions by not watering for several days. Several single leaf samples were harvested daily at 10 am, 2 pm or 4 pm, by handpicking from random trays of plants during the course of desiccation. Each harvested leaf was immediately split into two halves by carefully tearing along the midrib. One half was immediately snap frozen in liquid nitrogen and stored at -80 °C until being used for RNA extraction. The other half was used to determine the %RWC of that harvested leaf. The half-leaf was immediately weighed on an analytical balance (model SBC33, Scaltec Instruments GmbH, Heiligenstadt, Germany) to obtain the fresh weight, and then allowed to desiccate in a Petri dish containing silicon gel, in a 70 °C oven for 3 days. After oven incubation, the dry weight of the half leaf was measured. The %RWC of the half leaf was determined by subtracting the dry weight from the fresh weight, to first obtain the absolute water content (AWC), which was then divided by 2.11 (an averaged AWC value determined from 5 leaves at full turgor), then multiplied by 100.

3.2.3 RNA purification and Microarray fluorescent probe preparation

Total RNA was extracted from each half-leaf sample using TRI-Reagent (Molecular Research Centre, Inc., USA), and purified on Qiagen RNeasy Mini kit columns (Qiagen, Germany) according to the manufacturer's instructions. The concentration and purity of the purified total RNA from each sample was quantified on a NanoDrop 1000 spectrophotometer (NanoDrop Technologies, Delaware, USA). The integrity of RNA was checked by visualization of 0.5 μg of each sample electrophoresed on a formaldehyde denaturing gel containing 1% agarose, 1X 3-(N-morpholino) propanesulfonic acid buffer (MOPS) and 6.75% formaldehyde.

One microgram of purified total RNA extracted from each harvested leaf sample was linearly amplified and labeled with Cy3 using the Ambion Allyl MessageAmp II aRNA Amplification Kit (Ambion Incorporated, Texas, USA) following the manufacturer's instructions. A common reference sample was constructed by pooling equal amounts of purified total RNA derived from a set of 6 harvested leaf samples at 100.2, 81.1, 60.7, 41.9, 21.8 and 6.1% RWC respectively. This pooled total RNA was also linearly amplified and labeled with Cy5 using Ambion Allyl MessageAmp II aRNA Amplification Kit (Ambion Incorporated, Texas, USA).

3.2.4 Microarray hybridization and washing

Cy3 labeled exacts representing transcripts from a particular RWC leaf sample were simultaneously hybridized to the arrayed probes with Cy5 labeled common reference sample in each array (Fig. 3.1). To minimize batch effects, a complete set of samples representing 6 different %RWC stages were processed as a batch, rather than processing all biological replicates of a particular %RWC stage simultaneously. Microarray slides containing arrayed cDNA clones were pre-hybridized at 42 °C for 2 hours in a 100 µl buffer containing 5X saline-sodium citrate buffer (SSC), 0.1% sodium dodecyl sulphate buffer (SDS) and 1% bovine serum albumin (BSA under lifterslips (Erie Scientific, USA) in Telechem hybridization chambers (Telechem International Inc., California, USA). After pre-hybridization, the slides were washed five times in MilliQ H_2O for 30 sec with agitation, dipped in absolute ethanol, before drying by centrifugation at 1000 g for 5 mins. A mixture comprising 1 µg Cy3 labeled experimental probe (i.e. synthesized from leaf mRNA from different RWC samples), 1 µg Cy5 labeled common reference probe, 1 µg Mouse COT1-DNA (Life Technologies, Maryland, USA), 12 µg Poly(A)-DNA (Pharmacia), 25% formamide, 5X SSC and 0.2% SDS in a 60 µl reaction was denatured at 90 °C for 3 min. Each microarray slide was subsequently hybridized to the labeled mixture under lifterslips (Erie Scientific, USA) in the dark for 16 hours at 42 °C in Telechem hybridization chambers (Telechem International Inc., California, USA). Post-hybridization washes of the microarray slides were carried out as follows: (1) 4 min with agitation at 42 °C pre-warmed buffer containing 2X SSC and 0.5% SDS; (2) 4 min with agitation at room temperature in 0.5X SSC; and (3) 4 min with agitation at room temperature in 0.05X SSC. The slides were dipped in absolute ethanol after these washes, and then dried by centrifugation at 1000 x *g* for 5 min.

3.2.5 Microarray slide scanning and data capturing

Slide scanning and data capturing were carried out using the Axon GenePix 4000 scanner and GenePix Pro 5.0 software (Molecular Devices, California, USA) following manufacturer's instructions. During preview scanning, PMT gain (Photomultiplier tube) settings for both Cy3 and Cy5 channels were adjusted to ensure that the same amount of red and green signals were acquired in each channel. Saturation thresholds were also adjusted to ensure that the majority of the feature signals were below saturation level. The slides were then scanned using the PMT gain settings optimized from preview scans. A feature-indicator grid was aligned onto the image of each scanned slide to capture fluorescent intensity data from the scanned slides. Each printed cDNA spot was identified and the foreground, background signal intensities as well as other information for each spot were quantified and extracted by the GenePix Pro 5.0 software according to manufacturer's instructions.

3.2.6 Microarray raw data visualization

The R software package (version 2.2.1) was used to plot the raw expression data associated with each spot ID. All the detailed R scripts used for visualization of the raw microarray raw data are available in Appendix (A.3.1). The background associated with each spot on each slide was visualized using the function "image" in package marray, which creates false-colour images correspond to the values of a statistic for each spot on an array. Box plots of the raw log ratios for each spot were generated to visualize the data distributions within or across arrays, as well as print-tip effect in each array. Spatial effects caused by uneven hybridizations or scanning were assessed using the function "maQualityPlots" in package arrayQuality, which generates false-colour, diagnostic plots. Data distribution of spot intensities for each dye on each array was also visualized by histograms generated from GenePix Pro 6.0 software.

3.2.7 Microarray data normalization and preprocessing

Spatial bias within each array was normalized by a median based method in package marray using function "maNorm" available in the R software package (version 2.2.1). Rquantile method with function "normalizeBetweenArrays" in limma, was used to correct for scaling differences across arrays.

The list of 3105 spot IDs (presented as clone library references) in the raw data files (.gpr and .txt) exported from GenePix, as well as in the .gal file, were replaced by the

corresponding 1680 contig group IDs identified in sequence clustering process described in Chapter 2. The normalized expression values for all technical replicates for each *X. humilis* contig present in each slide/array were merged to generate an averaged expression value for each contig. Lastly, all normalized $\log_2$ ratio values were converted from the default Cy5/Cy3 format to Cy3/Cy5 (test/reference) formatted $\log_2$ ratio values to give final normalized preprocessed microarray dataset of unique *X. humilis* contigs. The detailed R scripts used for data normalization are given in Appendix (A.3.2).

3.2.8 Microarray data storage

The output raw data files (.gpr) of all 3105 *X. humilis* library clones for all 30 hybridizations were submitted to NCBI's Gene Expression Omnibus (GEO, Edgar *et al*., 2002, Barrett et al., 2011) by Mr Gerrit Botha from the Computational Biology Group (CBIO), University of Cape Town (January 2012). The averaged, normalized expression, annotation information and differential screening results of all cDNAs corresponding to the 1680 *X. humilis* contigs identified in chapter 2, were also submitted to GEO. These data are accessible through GEO Series accession number GSE34951 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc= GSE34951).

3.2.9 Normalization on A. thaliana microarray data

Affymetrix microarray data for seed development in *A. thaliana* published by Le *et al*., 2010 (globular-stage embryos (GLOB), cotyledon-stage embryos (COT), mature green embryos (MG) and postmature green embryos (PMG) stages) and Nakabayashi *et al*., 2005 (dry seed (DS) stage) were downloaded from the Gene Expression Omnibus (GEO) at NCBI (series GSE680) and Table S7 (Nakabayashi *et al*., 2005) respectively. A time course microarray data on osmotic stress response (mannitol treatment over 24 hours) in shoot of *A. thaliana* seedlings (Kilian *et al*., 2007) was downloaded from the Nottingham Arabidopsis Stock Centre's microarray database (NASCArrays, Reference NASCARRAYS-139). The same Affymetric genechip (ATH1) was used to generate all these datasets. The raw array datasets of both the seed development and osmotic stress profiles were read into R, and normalized by quantile normalization in package limma, using function "normalizeBetweenArrays". The detailed R scripts used for the normalization of the *A. thaliana* microarray data are given in Appendix (A.3.3; A.3.4)

3.2.10 Identification of X. humilis and A. thaliana orthologues

A common set of orthologues present in the *X. humilis* and *A. thaliana* microarray datasets was identified by a reciprocal BLASTP search of the predicted peptide sequences of all 1680 *X. humilis* contig groups against the *A. thaliana* non-redundant protein database (March 2011). The 1680 contigs were mapped to 825 *A. thaliana* gene locus IDs using the mapping information between UniProt IDs and *A. thaliana* gene locus IDs provided by The Arabidopsis Information Resource (TAIR, http://www.arabidopsis.org).

3.2.11 Identification of differentially expressed genes

The data of 772 orthologues found present in both *X. humilis* and *A. thaliana* were extracted from the respective datasets, and tested for differential expression using linear modelling (Smyth, 2004) in package limma. Differential expression test on the 1680 *X. humilis* contig dataset was also carried out using linear modelling in limma.

For the limma test on the 1680 *X. humilis* contigs and the 722 orthologues, the datasets were processed by following limma guidelines provided for the common reference designed experiments. The sample contrasts built for the linear model fit was specified as 80% *vs* 100%, 60% *vs* 100%, 40% *vs* 100%, 20% *vs* 100%, and 5% *vs* 100%RWC.

For the datasets of 772 orthologues captured during *A. thaliana* seed development, as well as during osmotic stress, they were processed by following limma guidelines provided for the single-channel designed experiments. The sample contrasts built for the linear model fit was specified as PMG *vs* MG and DS *vs* MG for the seed development data; and 0.5 *vs* 0, 1 *vs* 0, 3 *vs* 0, 6 *vs* 0, 12 *vs* 0 and 24 *vs* 0hr for the osmotic stress time course data.

The detailed R command lines are provided in Appendix (A.3.3; A.3.4).

3.2.12 Analysis of common differentially expressed genes

The overlap in differentially expressed genes common to the *X. humilis* and different *A. thaliana* datasets was analyzed using Venn diagram software at SABLab, (http://sablab.net/venn.php). The statistical significance of the overlap between two groups of genes was analyzed by using software provided by Jim Lund at the Department of Biology, University of Kentucky, USA (http://nemates.org/MA/progs/representation.stats.html).

3.3 Results and Discussion

3.3.1 X. humilis microarray experimental design and fabrication

The microarray experiments were designed to investigate the changes in mRNA transcript abundance in leaf tissue at six different stages of water loss during the desiccation (Fig. 3.1) with maximum efficiency while keeping the influences of technical variations to the minimum.

Firstly, the experimental design included 5 independent biological replicates for each RWC sample to increase the statistical robustness of the dataset. Half of each desiccating *X. humilis* leaf was used for RNA extraction, and the other half for measurement of RWC to ensure that each RNA sample was associated with a precise measurement of desiccation state. RNA samples were linearly amplified and labeled using T7 RNA polymerase *in vitro* transcription approach (Van Gelder *et al.*, 1990; Pabon *et al.*, 2001), as the initial amount of total RNA was insufficient for robust labeling. Several groups have verified that this method of amplification is indeed linear (Feldman *et al.*, 2002; Polacek *et al.*, 2003; Li *et al.*, 2004), with minimum bias in the amplification efficiency towards more abundant transcripts than the ones transcribed in low levels. A common reference sample similarly amplified and labeled with Cy5, was included in the experimental design to facilitate normalization and comparison between different arrays (Fig 3.1). An equal proportion of total RNA extracted from the biological repeat 1 of each RWC samples was pooled to generate the common pool, ensuring that all possible transcripts were present in the reference sample.

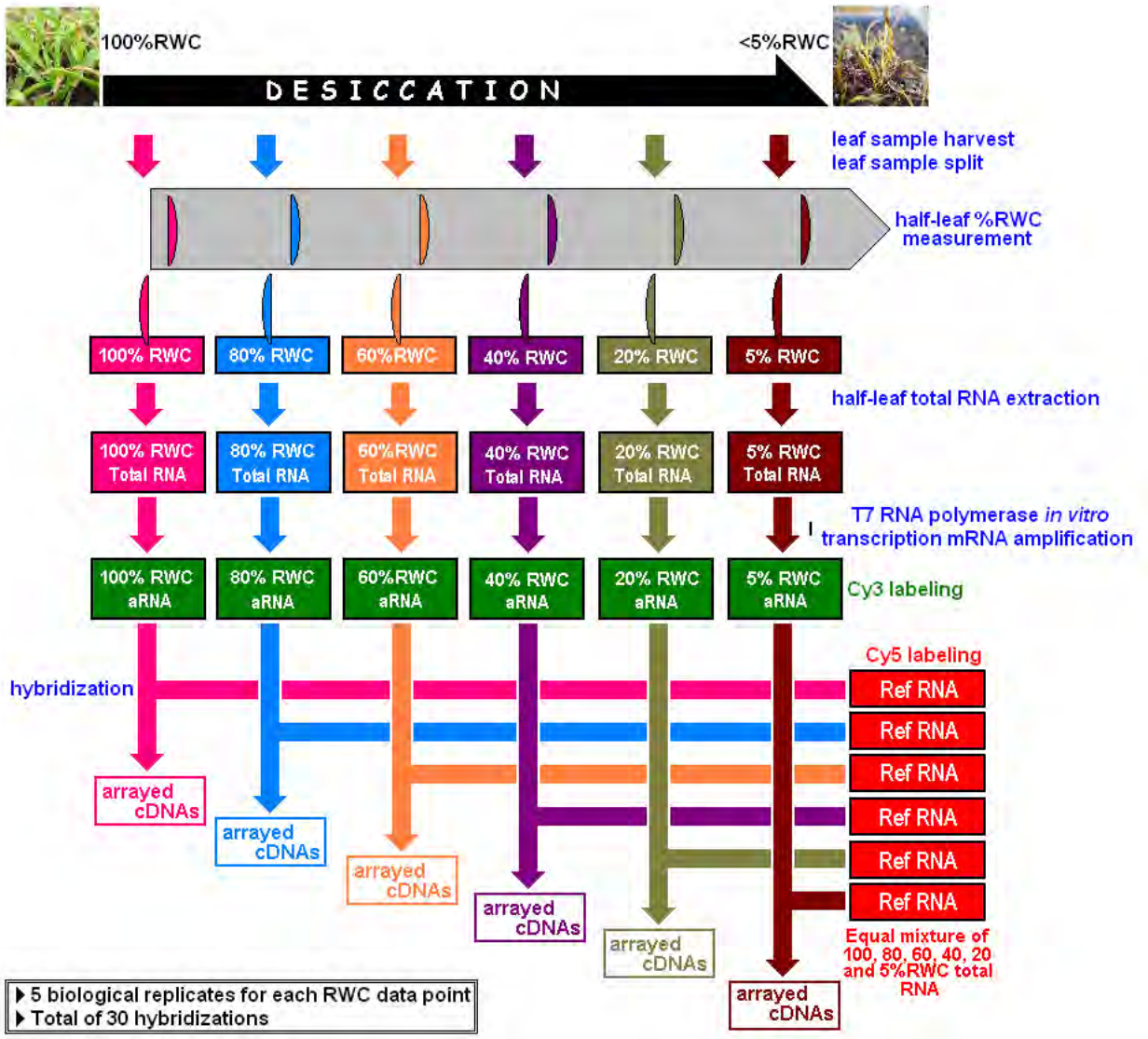


Figure 3.1. Schematic representation of microarray experimental design to investigate the changes in mRNA transcript abundance in leaf tissue at six different stages of water loss.

The overall microarray slide design is summarized in Fig. 3.2. A small set of 424 annotated *X. humilis* library clones used in the preliminary study (Collett *et al.*, 2004) were included on the printed slides for the purpose of comparing results across experiments as a control to check for the robustness of the data analysis pipeline. Each *X. humilis* library clone was printed as least four times to minimize technical variation due to printing artifacts (Fig. 3.2A).

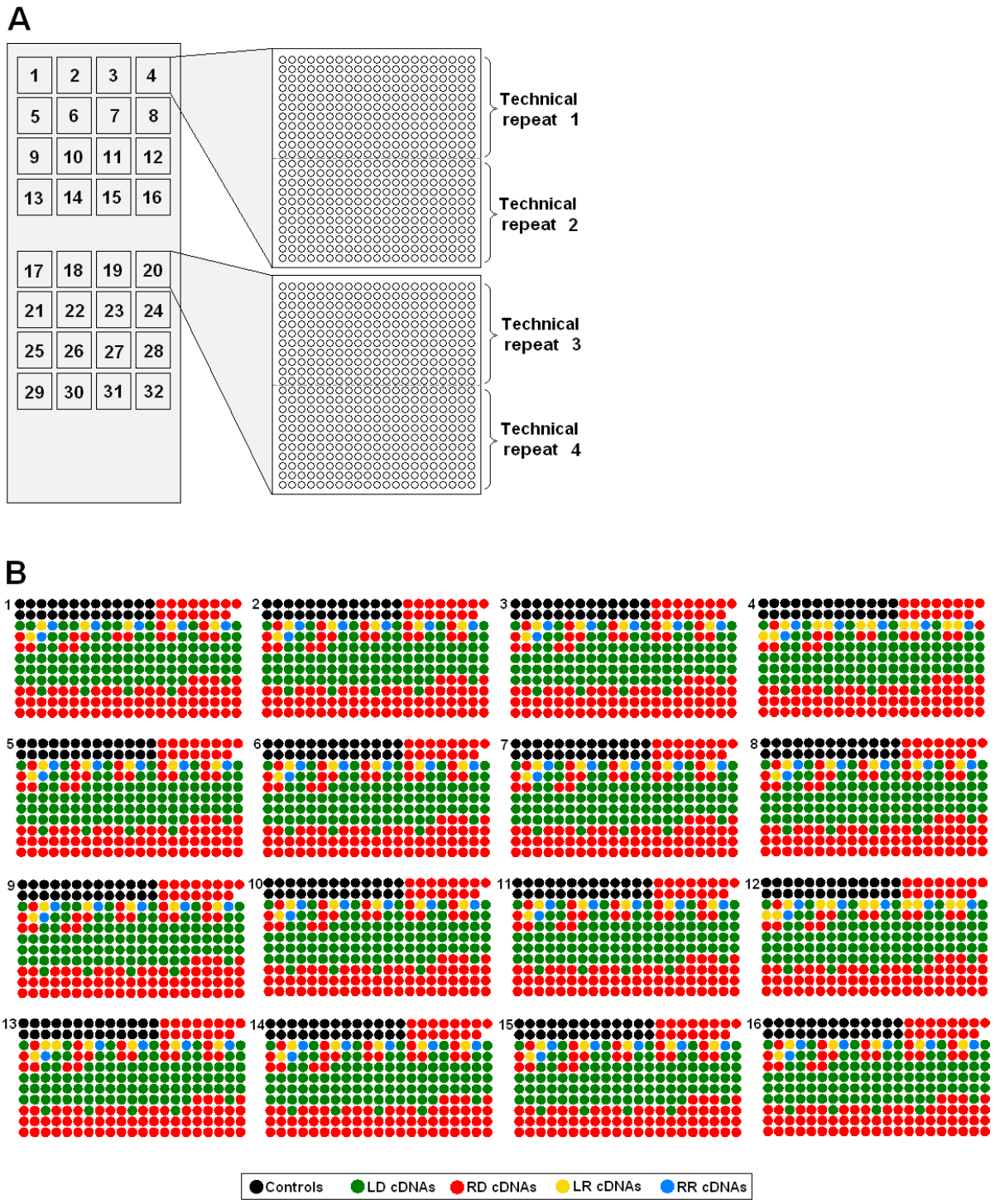


Figure 3.2. Microarray slide outline. Insert of 3105 *X. humilis* clones were PCR amplified, purified then printed onto GAPS II coated slides along with various Lucidea controls by MicroGrid II Arrayer with 16 10 µm split pins in a 4x8 configuration. (A) Each printed slide consisted of 14784 spots, arrayed in 32 blocks of 462 (21 by 22) spots. Blocks 1 to 16 were duplicate of blocks 17 to 32, and within each block, rows 1 to 11 were duplicate of rows 12 -22. (B) Schematic representation of cDNA library origins of *X. humilis* clones printed on slide showing rows 1 to 11 of blocks 1 to 16 only (a complete set of technical repeat within each slide).

In addition, control features such as calibration controls, negative controls and ratio controls were included in the print design (Fig. 3.2B). A minimum of two replicates of each control set was printed in each block. Each print block was designed to consist of features derived from all 4 *X. humilis* cDNA libraries, LD, RD, LR and RR. There was no print block, nor a large region on the slide, solely consisting of clones derived from one cDNA library, in order to prevent any false spatial effect caused by the biased source of the library clones (Fig. 3.2B). All cDNA inserts from the *X. humilis* libraries were amplified by PCR followed by purification to remove incorporated primers and dNTPs, and were checked on agarose gels prior to printing on the microarray slide (Fig 3.3).

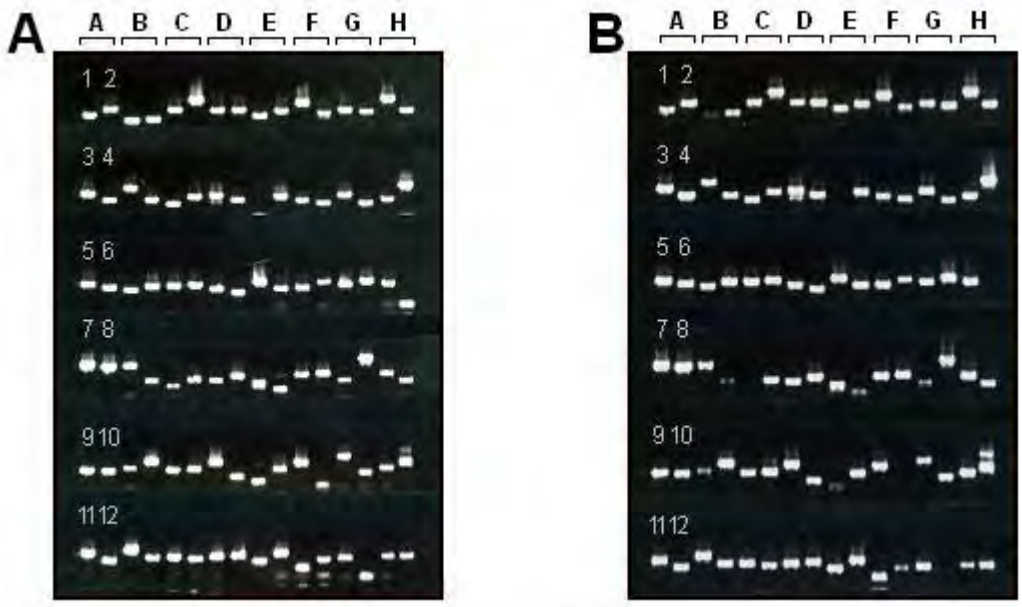


Figure 3.3. PCR amplification and purification of *X. humilis* cDNA library clone inserts. (A) The inserts of 96 clones of LD library plate 5: Xh_LD_05A01 to Xh_LD_05H12, were PCR amplified with T7 and T3 primers. (B) Purified cDNA inserts of clones Xh_LD_05A01 to Xh_LD_05H12 using Millipore 96-well Multiscreen PCR plates. Three microlitres of the purified PCR products were electrophoresed on a 0.8% agarose gel containing 100 ng/ml ethidium bromide.

3.3.2 % RWC measurement

The experimental design depended on the accurate measurement of RWC in the leaves as it was important that leaves of the same % RWC were matched for RNA extraction to minimize biological variation at each desiccation stage. The accuracy and reproducibility of the RWC measurement was thus assessed by determining the RWC of the two halves of a single leaf. Three hydrated and 3 desiccated leaves were tested (Table 3.1) and the results confirmed that the measurement of the two halves from the same leaf gave similar %RWC readings. The

RWC of half-leaves were determined from a total of 210 leaves collected during the desiccating treatment.

Table.3.1. Half leaf sample %RWC testing.

Sample No	% RWC left half-leaf	% RWC right half-leaf
Hydrated leaf no 1	94.04	96.93
Hydrated leaf no 2	96.71	94.12
Hydrated leaf no 3	93.62	93.12
Dry leaf no 1	8.12	8.27
Dry leaf no 2	6.57	6.44
Dry leaf no 3	6.68	6.55

3.3.3 Extraction and labeling of the RNA from X. humilis leaf samples

A total of 5 half-leaf samples with closely matched RWC values (Table 3.2) for each of the 6 %RWC stages, namely 100%, 80%, 60%, 40%, 20% and 5% RWC, were selected for total RNA extraction. The integrity of all RNA samples was checked by gel electrophoresis (Fig 3.4). On average, 5 to 21 µg of good quality total RNA were extracted and purified from each half leaf sample selected (Table 3.2).

Table 3.2. Summary of % RWC, purity and yield of total RNA extracted from half-leaves, as well as the labeling efficiency and amount of RNA amplified from 1µg of total RNA.

Desiccation point	Biological sample	%RWC	Harvest time of the day	Concentration (ng/µl)	Yield of total RNA (µg)	Yield of amplified RNA (µg)	Labeling efficiency (dye molecules per 1000 nucleotides)
100%RWC	Repeat 1	100.23	10am	547.90	16.40	74.45	60.55
	Repeat 2	100.42	10am	413.50	12.41	66.38	47.94
	Repeat 3	100.31	2pm	235.60	7.07	70.99	32.37
	Repeat 4	105.58	10am	165.90	4.98	83.77	45.53
	Repeat 5	109.11	10am	270.70	8.12	96.28	52.24
80%RWC	Repeat 1	81.14	2pm	347.00	10.41	73.04	45.23
	Repeat 2	82.62	2pm	174.80	5.24	90.05	44.81
	Repeat 3	80.43	2pm	361.30	10.84	69.91	37.76
	Repeat 4	81.59	2pm	544.40	16.33	100.20	44.66
	Repeat 5	80.78	2pm	342.70	10.28	99.10	49.73
60%RWC	Repeat 1	60.70	4pm	540.90	16.23	97.16	58.35
	Repeat 2	60.18	10am	464.90	13.95	100.31	47.46
	Repeat 3	58.00	10am	523.60	15.71	45.79	43.91
	Repeat 4	57.45	2pm	371.30	11.14	72.53	46.49
	Repeat 5	64.81	10am	465.40	13.96	105.17	52.92

Table 3.2. (continued)

Desiccation point	Biological sample	%RWC	Harvest time of the day	Concentration (ng/µl)	Yield of total RNA (µg)	Yield of amplified RNA (µg)	Labeling efficiency (dye molecules per 1000 nucleotides)
40%RWC	Repeat 1	41.89	10am	292.00	8..76	81.37	63.55
	Repeat 2	40.85	4pm	583.20	17.50	85.04	45.27
	Repeat 3	42.05	10am	386.00	11.58	91.34	36.93
	Repeat 4	39.65	10am	637.40	19.12	118.45	45.72
	Repeat 5	39.25	4pm	598.30	17.95	101.13	53.58
20%RWC	Repeat 1	21.78	10am	428.90	12.87	88.73	65.36
	Repeat 2	19.44	10am	262.50	7.88	69.51	44.27
	Repeat 3	19.15	2pm	328.60	9.86	105.76	38.05
	Repeat 4	20.66	2pm	413.50	12.41	116.86	48.39
	Repeat 5	22.67	10am	621.20	18.64	76.38	52.73
5%RWC	Repeat 1	6.13	10am	204.20	6.13	25.54	64.53
	Repeat 2	6.87	10am	370.80	11.12	92.69	48.55
	Repeat 3	6.73	10am	416.60	12.50	74.65	34.45
	Repeat 4	5.56	10am	703.90	21.12	121.94	51.10
	Repeat 5	4.96	10am	322.70	9.68	111.25	54.86
Reference RNA pool		mixture of 100.23 81.14 60.70 41.89 21.78 6.13	mixture of 10am 2pm 4pm	365.40	30	186.06	56.50 (labeled probe sample 1)
							58.53 (labeled probe sample 2)
							52.45 (labeled probe sample 3)
							59.08 (labeled probe sample 4)
							59.37 (labeled probe sample 5)

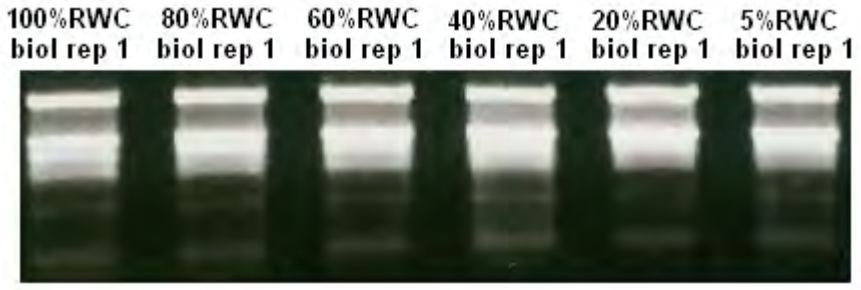


Figure 3.4. *X. humilis* leaf total RNA extraction. Total RNA samples from selected half leaf samples for biological repeat 1 were extracted using TRI-Reagent, then purified by Qiagen RNeasy Mini kit columns. Half micrograms of each purified extracted leaf total RNA samples were electrophoresed on a formaldehyde denaturing gel.

One microgram of total RNA from each leaf sample isolated, as well as from the reference RNA pool, was linearly amplified and purified. The purified amplified RNA (aRNA) samples showed a similar mRNA transcript amplification profile across all samples (Table 3.2; Fig. 3.5). Twenty micrograms of each purified aRNA sample was used in labeling reaction, which yield an average of between 7 to 25 µg of Cy dye coupled aRNA, and the labeling

efficiencies of the reactions were good, ranging from 34 to 65 Cy molecules per 1000 nucleotides (Table 3.2).

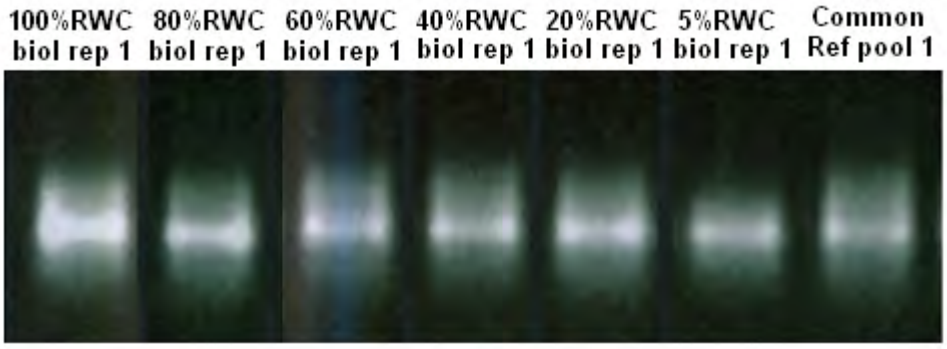


Figure 3.5. Linear amplification of *X. humilis* leaf mRNA transcripts. mRNA transcripts present in 1 µg of total RNA was linearly amplified using Ambion Amino Allyl MessageAmp II aRNA Amplification Kit. After aRNA purification, 1 µl of each reaction was electrophoresed on a formaldehyde denaturing gel. The reference RNA pool was constructed by equal pooling of 100%, 80%, 60%, 40%, 20% and 5% RWC leaf total RNA samples. The pooled mRNA populations were amplified once in a same manner as with the mRNA samples in extracted leaf total RNA samples. This amplified pooled mRNA reference sample was independently labeled in different batches of hybridization together with selected aRNA samples representing the 6 stages of RWC.

3.3.4 Microarray hybridizations and data capturing

The amplified, Cy3 labeled (green) RNA from the experimental samples were co-hybridized with amplified, Cy5 (red) labeled RNA from the reference pool for the 5 biological replicates for each of the 6 %RWC data points. After hybridization, the slides were immediately scanned, and the raw data values were captured. The scanned image for each slide revealed that the hybridization signals of the spots were robust and the level of background noise, or non-specific hybridizations, was minimal (Fig. 3.6). A clear change in the overall colour of the signals from the microarray slides, from red to green, was apparent as the %RWC of the experimental samples decreased (Fig. 3.6). This reflected a robust change in the mRNA transcript abundance of cDNA clones being assayed with many of the cDNAs arrayed on slide being induced in response to desiccation relative to the reference sample.

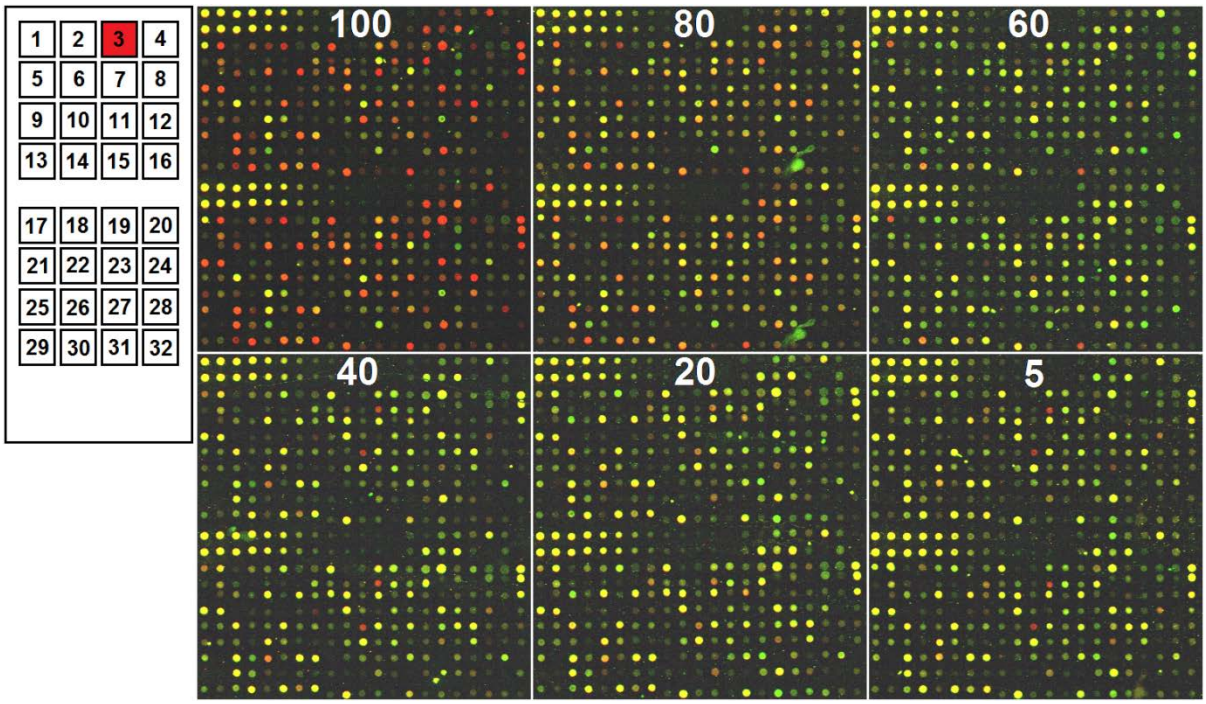


Figure 3.6. Scanned images of microarray slides representing each desiccation stage (100% down to 5%RWC) for biological repeat 2. Amplified mRNA transcripts isolated from leaf samples were labeled with Cy3 fluorescent dyes (green), and the common reference pool samples were labeled with Cy5 (red). Block 3 features of each slide image is enlarged and shown.

3.3.5 Visualization of the raw microarray data

A number of graphs were plotted to visualize the raw data and to identify the presence of any specific systematic bias present in the microarray dataset. The false-colour plots on the background intensity of green and red channels of all slides revealed that the background noise level was minimal, and that the levels of non-specific hybridization associated with each spot were very low (Fig. 3.7). Given these low levels of overall background for each microarray hybridization, it was decided to not include background subtraction in the normalization strategy, to minimize the false introduction of noise being introduced by erroneous background correction.

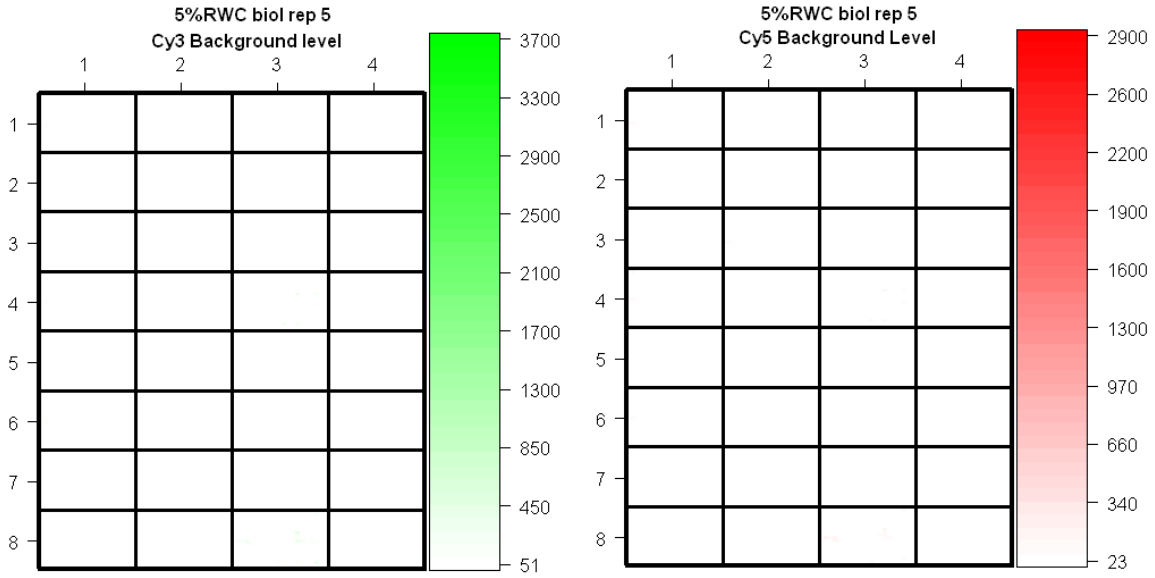


Figure 3.7. An example of background assessment from the microarray dataset for the5%RWC biological replicate 5 slide. Background intensities identified and measured in each of the 32 blocks on slide for the Cy3 (green) and Cy5 (red) channels were plotted and represented in false-colour image by marray package in R. The other slides gave similar images.

Histograms revealed that although the distributions of $\log_2$(Cy5/Cy3) values for each experimental condition were reproducible within the biological replicates, they were not similar across experimental conditions (Fig. 3.8). While data from the other RWC slides showed a more symmetrical normal distribution, the data distribution pattern from the 100%RWC slides revealed a tail of data points with large $\log_2$ (Cy5/Cy3) values. This phenomenon was consistent with the scanned images of the 100%RWC samples which had a higher proportion of red spots (Fig 3.6). This analysis showed that algorithms which assumed that raw microarray data is similarly distributed, would not be suitable for normalizing these datasets.

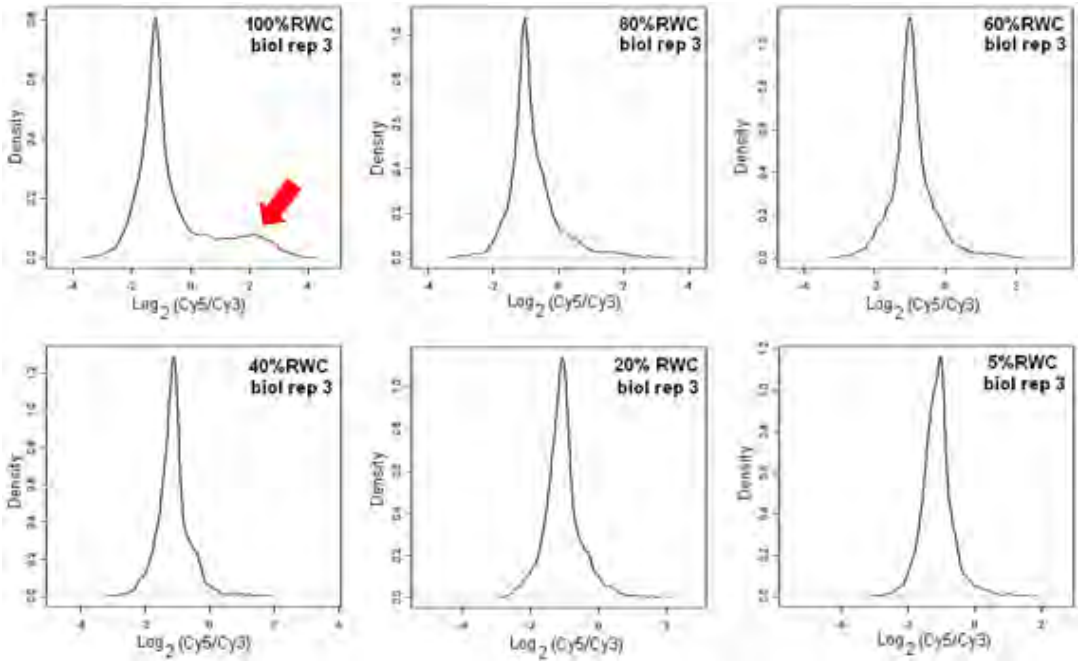


Figure 3.8. Data distribution of raw expression values. The frequency occurrence (expressed as density) of raw expression values (measured in $\log_2$ (Cy5/Cy3) from all 14784 *X. humilis* cDNA features present on each slide were plotted on histograms. The red arrow indicates the unique distribution pattern observed in the biological sample 3 of 100%RWC, revealing that there was greater representation of spots with large $\log_2$ (Cy5/Cy3) values than that observed in the other samples.

In addition to the histograms, the box plots generated from all 30 arrays (Fig. 3.9A) also showed the unique data distribution pattern for the 100%RWC samples, with a wider spread of values between the upper and lower quartiles, i.e. areas of top half of all the 100%RWC boxes were larger than the bottom half, indicating at least more than 50% of the data lay above the data median. One of the 80%RWC samples (biological replicate 1) showed a similar data distribution to the 100%RWC samples. The box plots on the raw expression data across all 30 arrays and within each array (Fig. 3.9A; Fig. 3.10A) revealed a heavy dye bias towards green (Cy3) channel, with the median $\log_2$(Cy5/Cy3) values being less than 0 for all arrays. The difference in labeling and hybridizations efficiencies of Cy3 and Cy5 is well documented. Cy3 is a less bulky molecule, and often labels and hybridizes better than Cy5 (Cox *et al.*, 2004; van Heerden *et al.*, 2007). Furthermore, presence of data scale difference across all 30 arrays which could possibly resulted from different hybridization efficiencies on each slide, was also detected (Fig. 3.10A). Such difference observed was not too serious, with data medians of 30 arrays fell between $\log_2$(Cy5/Cy3) values of 0 and -1. Nevertheless, the data median and the spreads from each array needed to be adjusted to similar levels for the purpose of conducting meaningful comparisons between the data.

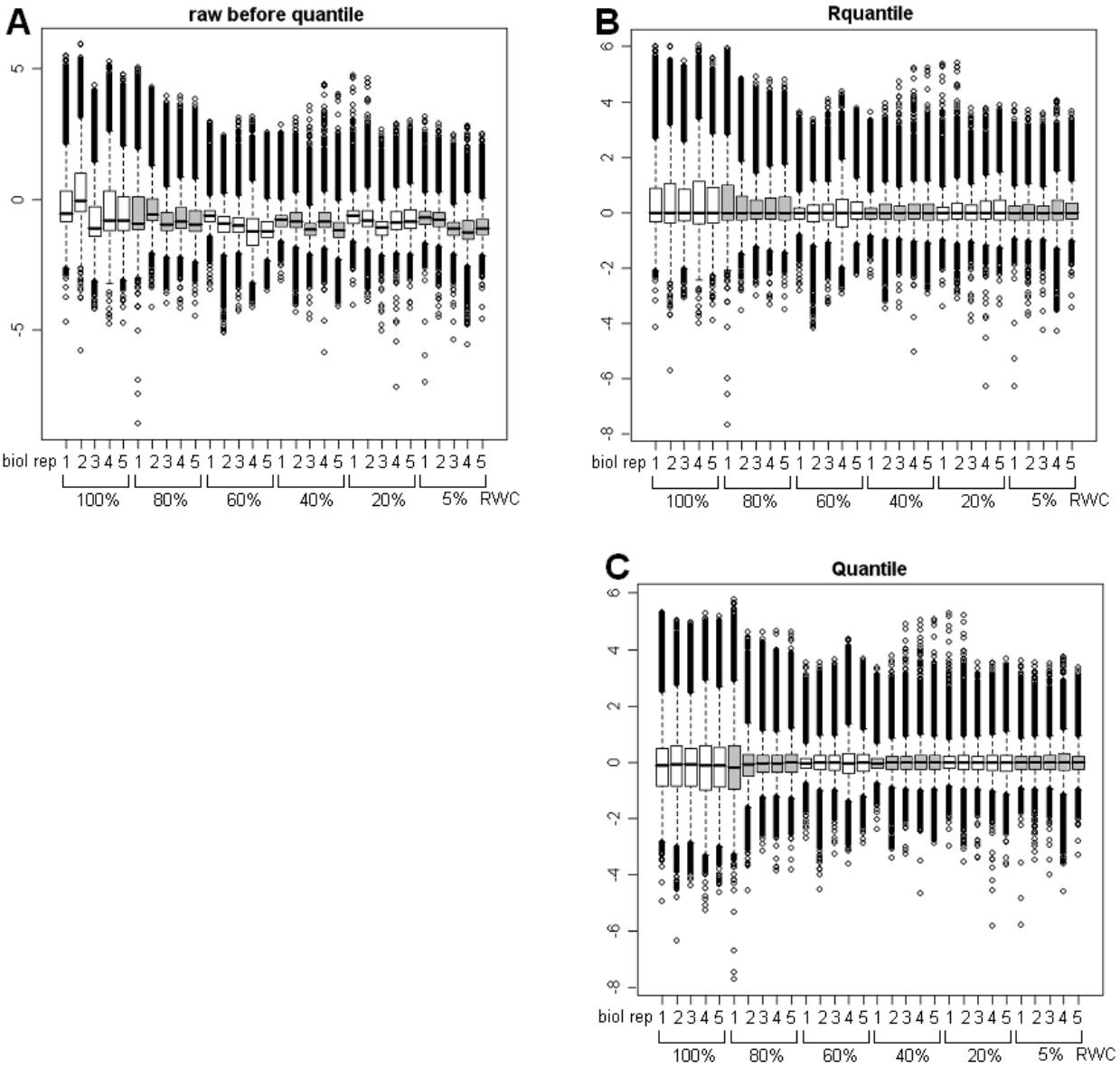


Figure 3.9. Box plot representations of the data distributions of the 30 arrays. Log_2 (Cy5/Cy3) values of features present on each slide were represented by box plots. (A) Raw data obtained after slide scanning and data capturing. (B) Raw data normalized by median normalization (spatial correction) and Rquantile normalization (scale correction). (C) Raw data normalized by median normalization (spatial correction) and quantile normalization (alternative scale correction).

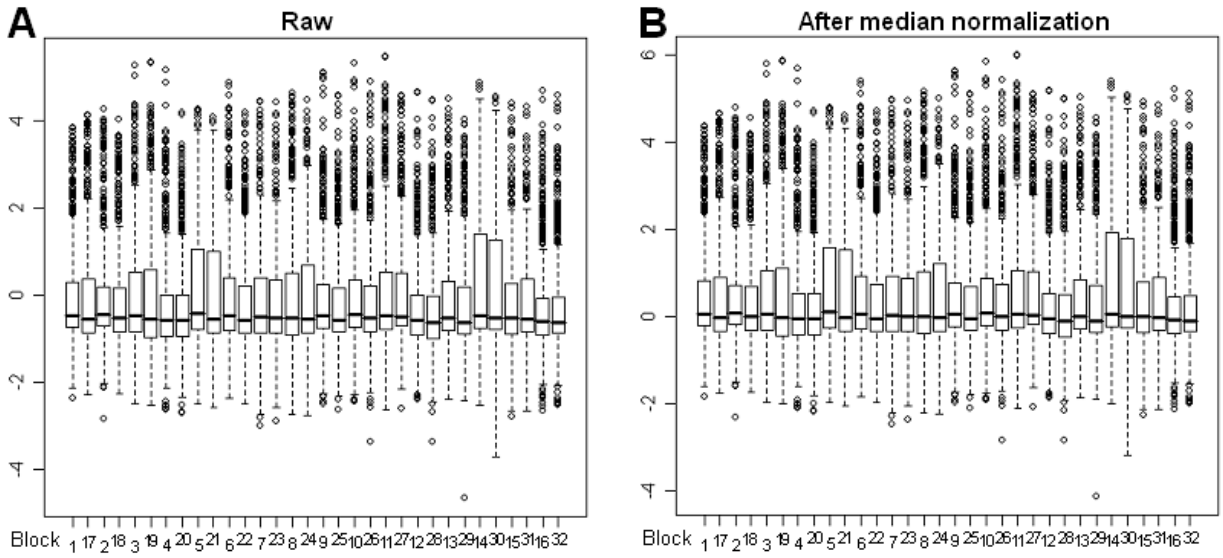


Figure 3.10. Box plot representations of the data distributions of the 32 blocks in 100%RWC biological replicate 1 array. Log_2 (Cy5/Cy3) values of features present in each block of 100%RWC biological replicate 1 array/slide were represented by box plots. (A) Raw data obtained after slide scanning and data capturing. (B) Raw data normalized by median normalization (spatial correction).

Diagnostic spatial heat maps were plotted to determine any signs of spatial bias resulting from possible from uneven hybridizations or from image scanning. Fig. 3.11 illustrates that there were spatial differences in the signals generated from biological replicates of a particular RWC data sample. The top half of the 20%RWC biological replicate 1 slide had a greater density of yellow spots, whereas the bottom of the slide showed a greater density of blue spots (Fig 3.10). These differences are likely to be a technical artifact as the spots in the top 16 blocks are a duplicate copy of the bottom 16. Different patterns of spatial effects were detected for all biological replicate samples within a particular RWC data sample. This was a clear indication of systematic error resulting from either uneven hybridization or image scanning which needed correction.

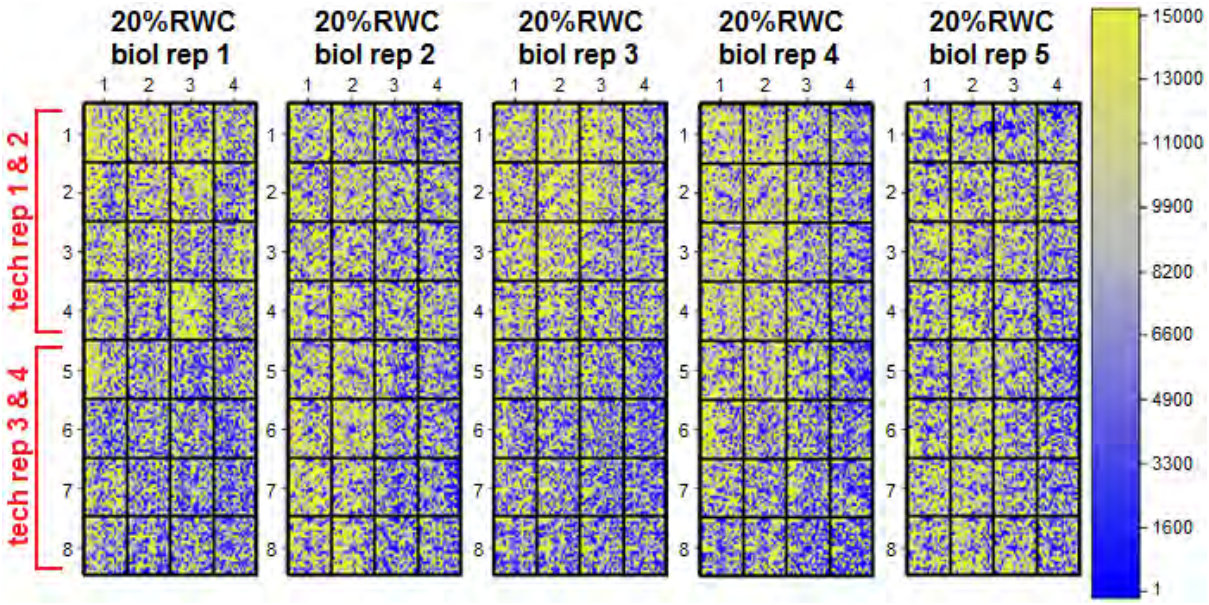


Figure 3.11. arrayQuality spatial diagnostic plots. Log_2 (Cy5/Cy3) values of features present on slides were ranked and colour coded. Bluer the spots, earlier the ranks (i.e. rank 1), lower the Cy5/Cy3 log ratio values; Yellower the spots, later the ranks (i.e. rank 15000), higher the Cy5/Cy3 log ratio values. The top 16 blocks are identical representation of the bottom 16 blocks in terms of cDNA spot identities (Fig. 3.2).

In summary, several conclusions could be drawn from the visualization of the raw data: (1) background noise levels were minimal, and no background subtraction was needed; (2) no apparent print-tip effects were detected; (3) a dye bias towards the Cy3 (green) was observed across all experimental conditions; (4) spatial biases were present in the raw data; and (5) scale difference across all 30 slides due to difference in slide hybridization efficiencies were present. In considering which normalization methods to use to remove these technical variations in the dataset, the conventional algorithms such as Global Lo(w)ess, Print-tip Lo(w)ess and 2D Lo(w)ess were excluded. The assumption that these algorithms are based on, namely that the majority of the genes present on a slide are not differentially expressed, and that the mean log ratio value should be equal to 0, is not met. The majority of the arrayed cDNAs were clones derived from Leaf Desiccation and Root Desiccation libraries, and were not randomly selected from a cDNA collection representing the full *X. humilis* genome. Thus the majority of genes arrayed on the slide were expected to be up-regulated during desiccation, with the mean log ratio value deviating from 0 for the different experimental conditions. A median based method was selected to correct for within array normalization to remove spatial biases as this algorithm is not based on the assumption that the majority of genes are not differentially expressed. Fig. 3.10 shows the box plots of data within a slide

before and after median normalization. By adjusting the neighbourhood medians to similar levels, the means of each block were also brought closer to a similar level.

Quantile normalization was excluded as a method to correct for the scale biases across all 30 arrays as the assumption that the distribution of log ratio values should be the same across all arrays was not met. Although the histograms of the *X. humilis* microarray data reflected reproducible shapes of data distribution within all biological replicates of each RWC data point, the 100%RWC dataset showed a different data distribution shape to the other RWC samples (Fig. 3.8). Instead, advantage was taken of the common reference RNA, labeled with Cy5 (red) which should have shown the same distribution of values for each slide. The Rquantile algorithm first normalized on the red (Cy5) channel data across all arrays, then adjusted the green (Cy3) channel data accordingly. Box plots comparing *X. humilis* microarray data after R quantile and after quantile normalization (Fig. 3.10), show that the biological signal of a greater proportion of down-regulated genes in the 100% RWC (i.e. a higher proportion of higher Cy5/Cy3 values) was preserved in the data after Rquantile normalization, i.e. the size proportions of the top and bottom halves of 100%RWC data boxes were retained. This biological signal was lost for data normalized by the normal quantile method (Fig 3.10).

3.3.6 Identification of differentially expressed X. humilis genes

Differentially expressed genes in the normalized microarray data for 1680 unique *X. humilis* contigs were identified in the package limma. A linear model was first fitted for each contig present in the *X. humilis* data, and a contrast matrix was specified (as described in the Material and Methods section) to compare the 100%, 80%, 60%, 40% 20% and 5%RWC sample profiles in order to identify differentially expressed candidates. Estimated coefficients and standard errors for the specified set of contrasts were calculated, and used to compute moderated t-statistics of differential expression by empirical Bayes shrinkage of the standard errors towards a common value. After correction for multiple testing across genes and contrasts, a total of 1361 contigs representing 81% of the gene set, were identified as differentially expressed in *X. humilis* leaves during desiccation treatment. This high proportion of differentially expressed genes validates our decision not to use normalization methods based on the assumption that the expression values for the majority of genes would not change across the dataset.

3.3.7 Microarray data normalization for A. thaliana seed development and osmotic stress profiles and identification of a common set of X. humilis and A. thaliana orthologues

Microarray expression data of 22414 *A. thaliana* genes during seed development, as well as during osmotic stress treatment were downloaded and included in the analysis in order to examine the overlap between genes differentially expressed during desiccation in *X. humilis* leaves, desiccation in *A. thaliana* seeds, and abiotic stress in *A. thaliana* seedlings. For the *A. thaliana* seed development data, only the later stages MG, PMG and DS in the seed developmental series were considered relevant for this analysis, as these stages captured gene expression changes from the initial phase of water loss to the phase where the seed was fully desiccated. Due to the large difference in the size of gene sets between *A. thaliana* and *X. humilis*, it was decided to perform the test of differentially expressed genes on a subset of genes containing only the orthologues found common in both data sets, to make sure that errors in multiple testing were minimized by identifying differentially expressed genes in the same sized datasets. Furthermore, an analysis of a common set of genes allowed a direct comparison of genes expressed during the osmotic stress response and seed maturation in *A. thaliana*, with desiccation in *X. humilis* leaves.

A total of 891 *X. humilis* contigs were identified as having *A. thaliana* orthologues by a reciprocal BLASTP search. The *A. thaliana* UniProt IDs associated with these orthologues could be directly mapped to 825 *A. thaliana* locus IDs, of which 772 were represented on the *A. thaliana* Affy Geneschip (ATH1).

3.3.8 Comparison of differentially expressed genes in A. thaliana seed development and osmotic stress profiles, and X. humilis desiccation series

Expression values corresponding to the set of 772 orthologues, present on both the *A. thaliana* AffyChip and *X. humilis* microarray slide were then extracted from the respective normalized datasets. A total of 615 significantly differentially expressed genes were identified in the *X. humilis* desiccation dataset, 504 in the *A. thaliana* late seed development dataset, and 423 in the *A. thaliana* seedling osmotic stress dataset. The overlap between these differentially expressed genes is summarized in Fig 3.12.

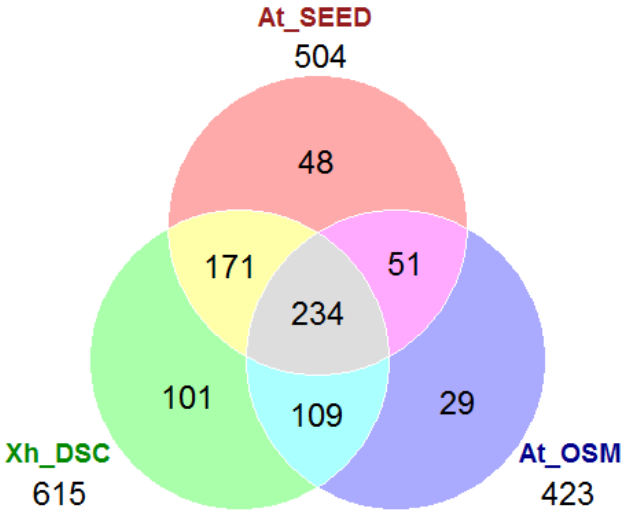


Figure 3.12. A Venn diagram showing the overlap between the differentially expressed genes identified in *X. humilis* during desiccation (Xh_DSC), and in *A. thaliana* during seed development (At_SEED) and during osmotic stress (At_OSM). Numbers underneath the expression profile name indicated the total number of differentially expressed genes identified in that expression profile from the common set of 772 genes.

The Venn diagram analysis showed the greatest degree of overlap between differentially regulated (both up- or down-regulated) genes identified in the *X. humilis* desiccation and *A. thaliana* seed series (a common set of 405 genes, with 171 being specific to desiccation) (Fig. 3.12). The statistical significance of this overlap was analyzed using a hypergeometic distribution calculator available at (http://nemates.org/MA/progs/overlap_stats.html). The software makes use of the following information namely the total number of genes present in group 1, the total number of genes present in group 2, total number of genes overlapped between group 1 and group 2, and the number of genes in the genome, to calculate a representation factor, as well as the probability, the p-value, using a hypergeometric probability formula (Bouyer *et al*., 2011). The detail of the calculations on representation factor and the p-value can be found is provided in the webpage http://nemates.org/MA/progs/representation.stats.html. In general, a representation factor of >1 indicates that the total number of genes overlapping between the two analyzed groups was more than expected, suggesting that the overlap was significant. A representation factor of <1 suggested the overlap between the two group was less than expected, thus not significant. A representation factor of 1 indicated that the two groups overlapped by the number of genes expected. This statistical test showed that there was significant overlap between differentially expressed genes across all three datasets (Table 3.3).

Table 3.3. Summary of statistical significance of common differentially expressed genes calculated using the hypergeometic distribution calculator.

Condition 1	Condition 2	R	p-value
X. humilis desiccation	*A. thaliana* seed development	29.3	< 0.000e+00
X. humilis desiccation	*A. thaliana* osmotic stress	29.6	< 0.000e+00
A. thaliana osmotic stress	*A. thaliana* seed development	30.0	< 0.000e+00

This large overlap between orthologues in the desiccating *X. humilis* leaf and *A. thaliana* seed maturation datasets, desiccation is consistent with our predication that the transcriptional profile of *X. humilis* genes during desiccation would be more similar to that during late seed development However, a substantial number of differentially expressed genes were shared between the *X. humilis* desiccation and *A. thaliana* osmotic stress datasets (343 genes) and between the *A. thaliana* seed and osmotic stress datasets (285 genes). This analysis suggests that features of vegetative osmotic stress tolerance are shared with seed maturation and desiccation.

Chapter 4

Functional enrichment analysis of the *X. humilis* microarray datasets

4.1 Introduction

Clustering analyses were applied to the 1680 *X. humilis* contigs that were differentially expressed across 6 different stages of water loss from 100% to 5% RWC to identify i) when the main switch in gene expression occurs during desiccation in *X. humilis* leaves, as well as ii) different groups (or cohorts) of co-regulated genes. Tests for functional enrichment of genes in these different cohorts were run to identify which biological processes or molecular functions are regulated during desiccation in *X. humilis*. In the introduction to this chapter, different clustering and functional enrichment algorithms currently available for the analysis of microarray data are reviewed.

4.1.1 Cluster Analysis

Cluster analysis is a popular approach for pattern classification in microarray datasets, which groups data observations (i.e. gene expression values) into different classes called clusters based on their similarities and dissimilarities (Page *et al.*, 2007; Loewe and Nelson, 2011). A number of different distance measures are available to measure the similarity between data observations including Pearson's correlation distance and Euclidean distance (Goldstein *et al.*, 2002; Stekel, 2003; Yona *et al.*, 2009). Pearson's correlation focuses on the directional similarity or whether the patterns of the expressions change in the same way, without concerning the amplitude of the expression vectors. Whereas the Euclidean distance measures the absolute distance between the expression patterns and takes the magnitude into account (Quackenbush, 2001; Do and Choi, 2007).

Cluster analysis can either be supervised or unsupervised. Supervised methods require some prior knowledge or information, i.e. existing biological information, regarding to how the genes or samples should be clustered to fit the predetermined pattern. Supervised methods are useful in clustering expression data to identify genes with expression patterns that are significantly associated with defined conditions, with the purpose of identifying genes that accurately predict a characteristic that condition. However, most of the cluster analyses in microarray studies are unsupervised, which allow exploratory grouping of genes solely based the similarities measured between genes or samples without the introduction of prior information or knowledge on genes or samples (Butte, 2002; Domany, 2003; Do and Choi,

2007). Unsupervised methods can be performed to identify groups of co-expressed genes that are likely to be involved in the same cellular processes, or are possibly regulated by the same regulatory network across all conditions studied (Eisen *et al.*, 1998; Spellman *et al.*, 1998; Heyer *et al.*, 1999). Alternatively, cluster analysis can be done to group samples with similar gene expression profiles to reveal the overall similarity of samples under the experimental conditions tested (Domany, 2003).

Several unsupervised clustering methods, including hierarchical and non-hierarchical approaches, have been applied to the analysis of microarray expression data. The most commonly used method is agglomerative hierarchical clustering (Eisen *et al.*, 1998; Quackenbush, 2001; Nugent and Meila, 2010). In this method, each gene profile (or sample profile if performing sample based analysis) is initially treated as an individual cluster. The pairwise distance matrix between all the gene profiles is first calculated, and the two gene profiles with the shortest distance are then merged to form a new cluster. This merging of clusters continues, and each time, the pair of clusters with shortest distance is combined to form a new cluster, until there is only one large cluster left at the end. There are various options for defining the distance measured between clusters if there are at least two member profiles in a cluster, such as (1) single-linkage that defines the minimum distance between the clusters by assessing the distance between the two nearest profiles in the clusters; (2) complete-linkage describes the maximum distance between the clusters by assessing the distance between the two furthest members in the clusters; and (3) average-linkage that determines the average distance by averaging all the distances between all members in the clusters. The average-linkage method is typically used for microarray expression data. (Quackenbush, 2001; Stekel, 2003; Nugent and Meila, 2009). Hierarchical clustering produces a tree-like dendrogram at the end of analysis that represents the relationship between genes and the samples or experimental conditions. The dendrogram may include a heat map, a coloured representation of gene expression matrix in the data to facilitate easier visualization and interpretation (Causton *et al.*, 2003). Although many of the hierarchical clustering methods are agglomerative, divisive versions also exists. In divisive hierarchical clustering, all data profiles are initially treated as one big cluster. With each subsequent step, the cluster that has the most overall dissimilarity between its members is split into two groups: the original cluster and the splinter group. This process of splitting clusters continues until all clusters are left with one member only (Kaufman and Rousseeuw, 1990).

An alternative, commonly used non-hierarchical method is K-means clustering (Tavazoie *et al.*, 1999; Do and Choi, 2007; Yona *et al.*, 2009). In this partitioning method, the number of clusters (k) which all gene expression patterns are intended to be classified into, is specified prior to the analysis. The gene expression patterns are randomly assigned to the k clusters, where after the centroid or medoid reflecting the centre of gravity of each cluster, is calculated. For each expression pattern, the distances between itself and the centroids of each of the k clusters are calculated. If the gene expression pattern is closest to a different cluster from the current cluster to which it belongs, it will be reassigned to that cluster. The centroids of both the "receiving" and "donating" clusters are readjusted and updated after this reassignment. This reattributing of the gene expression patterns continues until all the members within each of the k clusters are closest to the currently allocated cluster than to the centroids of the other clusters (Stekel, 2003; McLachlan *et al.*, 2008). The main disadvantage of k-means clustering is that k, the number of clusters needs to be specified before the analysis commences. One of the commonly used methods to help with the defining of k is through the analysis of the average Silhouette width scores (Rousseeuw, 1987; Kaufman and Rousseeuw, 1990). The Silhouette width is defined as (B-W)/max(B,W), where W denotes the average distance of a clustered expression pattern to all other expression patterns within its cluster, and B denotes the average distance of that expression pattern to all expression patterns in its nearest neighbouring cluster. Silhouette width scores have values which range between 1 and -1. A positive score closer to 1 indicates that this expression pattern is closer to the centre of its cluster; a score around 0 indicates that this expression pattern falls between the two compared clusters; a negative score indicates that this expression pattern may be in the wrong cluster (). An averaged Silhouette width score for a cluster can be determined by averaging the Silhouette width values calculated for all the expression patterns within the cluster, as an indication if the cluster is "well clustered". The Mclust algorithm available in the mclust package in R estimates the optimal value of k for a given expression data by analyzing the data with different k values to identify which value of k tested yields the maximum average Silhouette width score. There are several other, different methods currently available such as gap statistic (Tibshirani *et al.*, 2000), prediction-based resampling method Clest (Fridlyand and Dudoit, 2001) and many others (Pollard and van der Laan, 2005) which can be considered for the optimal estimation of k.

PAMSAM is a partitioning method, which combines a K-means type clustering method with a method for visualizing the levels of similarity between clusters via a multidimensional

scaling Sammon plot (Wit and McClure, 2004). Through a number of expand-and-collapse steps of expression pattern reattributions, PAMSAM reassigns expression patterns from cluster to cluster, attempting to increase the overall average Silhouette width scores. The final PAMSAM clustering is the one that attains the highest average Silhouette width scores (Wit and McClure, 2004), and such optimal clustering reflecting relative similarity of the clusters identified, is visualized via the multidimensional scaling method called Sammon mapping (Sammon, 1969). In PAMSAM, Sammon mapping creates a representation of clusters in two-dimensional space, that attempts to minimize the extent to which the distances between each of the clusters is distorted (Wit and McClure, 2004). In the graphical representation of PAMSAM clustering, the distances between different PAMSAM clusters are shown, and each PAMSAM cluster is represented by a graph derived from its medoid. These graphs are similar to heat maps and serve to summarize the expression pattern of genes in a cluster.

Self-organizing map (SOM) is another popular non-hierarchical partitioning method used in the analysis of microarray data, is the (Kohonen, 1995). Similar to k-means, the number and the geometric configuration of the clusters is arranged on a two-dimensional grid prior to the analysis commencing. However, unlike K-means, the centroid for each partition, or node, needs to be defined and trained before the actual assigning of gene expression profiles into the nodes begins. First, a random centroid, also known as the reference vector, is generated for each node, in a way that each node has closer centroids to its neighbouring nodes than the nodes being arrayed further on the two-dimensional grid. An expression pattern is then randomly selected from the dataset and assigned to the node with centroid that is closest to itself. Once a node has received the assigned expression pattern, its centroid is then recalculated and updated. In turn, the nodes on the grid are reorganized accordingly so that the neighbouring nodes have the closer centroids. The process of assigning an expression pattern to the closest node, and the reorganizing of the nodes continues until all expression patterns have been assigned to their closest node, and nodes with closer centroids are arrayed in close proximity on the grid. An advantage of SOM over K-means is that it does not force the number of clusters resulted to be equal to the number of starting nodes defined, because some nodes may have no expression patterns associated with them when the map is complete (Babu, 2004).

Principal components analysis (PCA) is another useful clustering method that projects a high dimensional space of microarray data onto a two-dimensional space while retaining

characteristics of data that contribute most to its variance, keeping lower-order principle components and ignoring higher-order ones (Raychaudhuri *et al.*, 2000; Quackenbush, 2001; de Haan *et al.*, 2007; Loewe and Nelson, 2011). First, the variance-covariance matrix for all data profiles present in multi-dimensional space is constructed, capturing the variability of each data profile and the extent to which it co-varies with every other profile. The principle component is defined, which represents a linear combination of the data profiles that has the maximum amount of variance. The second principal component, orthogonal to the first principal component, is then identified to have the maximum amount of the remaining variance. The process is repeated until as many principal components, orthogonal to the previous one, have been identified. These principal components determine the best ways for the clustering of the multi-dimensional data (Stekel, 2003; Loewe and Nelson, 2011).

4.1.2 Functional enrichment analysis

Clustering analysis of data from high throughput technologies such as microarrays often results in long lists of genes showing a similar expression profile. Interpreting the biological importance of these long lists is challenging. The use of databases such as Gene Ontology (GO), which uses a controlled vocabulary to describe the molecular, cellular and biological function of genes, has enabled researchers to statistically test whether particular functional terms are over- or under-represented within a list of genes of interest, compared to a reference list (Huang *et al.*, 2009). In microarray datasets, the test list will usually be a set of co-regulated, differentially expressed genes. Several authors have proposed that the total set of genes in a genome is the most appropriate reference list (Beissbarth and Speed, 2004; Martin *et al.*, 2004; Zeeberg *et al.*, 2003; Zhang *et al.*, 2004). However, there are several criticisms to this approach, with the main argument being that that only annotated genes that are present on the microarray platform should be considered as a reference list. (Khatri and Drăghici, 2005; Fresno *et al.*, 2012). Several statistical methods have been used to test for over-representation of GO terms within a list of genes (summarized in Huang *et al.*, 2009 and Dopazo, 2009), among them, Fisher's exact test is the most commonly used. Fisher's exact test takes several measures into consideration: (1) the total number of a tested GO term present in the test list, or in the reference list of genes; (2) the total number of other GO terms (excluding the tested term) present in the test or the reference lists of genes; and (3) the grand total number of GO terms (including the tested term) present in the test or the reference lists of genes, and calculates the probability of the tested GO terms being overrepresented in the test list (Barder and Enright, 2005; Fresno *et al.*, 2012). The probabilities or p-values

calculated are usually followed by the correction for multiple testing (Khatri and Drăghici, 2005; Huang *et al.*, 2009). Fisher's exact test can be one-tailed or two-tailed, to identify GO terms overrepresented, or overrepresented and underrepresented in the list of differentially expressed genes respectively, relative to the reference list.

Tools such as FatiGO (Al-Shahrour *et al.*, 2004) and Blast2GO (Conesa *et al.*, 2005; Conesa and Götz, 2008; Götz *et al.*, 2008) have been developed to allow researchers to identify enriched GO terms using Fisher's exact test, as well as the use of different methods to correct for multiple testing such as the popularly used FDR method and Bonferroni/FWER method. Another popular statistical method for GO enrichment analysis is the hypergeometrical test (Maere *et al.*, 2005). This method has been shown to be equivalent to the one-tailed Fisher's exact test (Rivals *et al.*, 2007). BiNGO (Biological Network Gene Ontology tool) is an alternative tool that offers GO enrichment test using the hypergeometrical test, as well as the FDR and Bonferroni/FWER options for multiple testing corrections (Maere *et al.*, 2005). It is available as a plugin for Cytoscape, a fast growing open-source bioinformatics software platform for visualizing molecular interaction networks with annotation integration (Shannon *et al.*, 2003; Saito *et al.*, 2012). An advantage of BiNGO over Blast2GO or FatiGO, is that it operates in Cytoscape's versatile visualization environment, which allows more flexibility in image customization that leads to the generation and export of high-quality figures.

In this chapter, PAMSAM was chosen to analyze the *X. humilis* dataset because the algorithm enables fast clustering analysis and returns results with graphs showing representative expression pattern for each cluster identified. RT-qPCR was used to measure the expression profile of a transcript identified in each PAMSAM cluster to independently validate the microarray data. Functional enrichment tests were run in BiNGO, due to the better quality of the images summarizing the results. Furthermore, clustering and functional enrichment analyses were also applied to the common set of 772 orthologues that are present in both the *X. humilis* cDNA arrays and in the *A. thaliana* ATH-1 chip arrays. The functional importance of cohorts identified in each expression profile for desiccating *X. humilis* leaves, developing seed in *A. thaliana*, and in the shoots of *A. thaliana* exposed to osmotic stress, were compared to assess the similarities between vegetative desiccation, seed development, and abiotic stress.

4.2 Material and Methods

4.2.1 Cluster analysis of X. humilis contigs

The $\log_2$ expression ratios of the 5 biological replicates of each RWC sample were averaged before performing the clustering analysis of differentially expressed contigs using the PAMSAM algorithm in the Smida package in R. The number of clusters (k value) specified and used in the PAMSAM clustering analysis was estimated by Mclust algorithm in the mclust package. For sample based clustering analysis, microarray data for the 30 samples (five per RWC condition) were kept separate. This clustering analysis was performed using the Sammon mapping and principal component analysis (PCA) methods in the Smida package, and the divisive hierarchical method (DIANA) in the cluster package in R. For the clustering analysis of LEAs, antioxidants and transcription factors, the $\log_2$ expression ratio data of contigs belong to each of the 3 gene families were extracted. Each group was separately analyzed using the PAMSAM algorithm. Detailed R scripts used for clustering analyses are supplemented in Appendix (A.3.2).

4.2.2 Cluster analysis of A. thaliana orthologues

For the clustering of the differentially expressed orthologues identified during seed development, or osmotic stress, expression values obtained from the ATH1 array of the 2 biological replicates of each seed developmental stage sample, or osmotic stress time series sample, were averaged before performing the clustering analysis of genes using the PAMSAM algorithm in the Smida package in R as described in 4.2.1. Detailed R scripts used for clustering analyses are supplemented in Appendix (A.3.3; A.3.4).

4.2.3 Functional enrichment analysis

The BiNGO 2.3 plugin for Cytoscape (http://www.psb.ugent.be/cbd/papers/BiNGO/Home.html) was used to test whether any GO terms were statistically over-represented in each PAMSAM cluster of contigs identified during desiccation in *X. humilis*. A custom annotation file for each of the 3 GO categories: biological process, cellular component, and molecular function, was created by extracting information on category-specific GO terms of the 1268 significantly annotated *X. humilis* contig peptide sequences in Blast2GO (as described in chapter 2). Hypergeometric tests were performed in BiNGO with default parameter settings, and a FDR test cutoff of p-value < 0.05 on each GO category separately. The balance of the genes in the list of 1268 annotated *X. humilis* contigs was specified as the reference set. Enrichment analysis was repeated with

cutoff of p-value < 0.1, if no over-representation of GO term was identified using cutoff of p-value < 0.05.

The annotation file for *A. thaliana* which is provided in BiNGO was used to test for functional enrichment of clusters identified during seed development or osmotic stress in *A. thaliana*. Tests for functional enrichment were carried out as described above, except that the list of common 772 orthologues identified in the *X. humilis* and *A. thaliana* array dataset, was used as the reference list. Enrichment analysis of KEGG pathways was carried out using FatiGO in BABELOMICS version 3.2 (http://babelomics3.bioinfo.cipf.es).

4.2.4 Microarray data validation: quantitative real-time PCR analysis

cDNA synthesis: mRNA transcript abundance of 24 selected contigs in 2 biological replicates of each of the leaf samples of 6 different stages of water loss (RWC100_1, RWC100_3, RWC80_1, RWC80_3, RWC60_1, RWC60_3, RWC40_1, RWC40_3, RWC20_1, RWC20_3, RWC5_1 and RWC5_3) were analyzed by quantitative real-time PCR analysis (RT-qPCR). 1.2 µg of (in volume of 10 µl) of the amplified RNA sample for each chosen RWC sample previously used in microarray experiments, was first denatured with 2 µl second round primers (Ambion Incorporated, Texas, USA) at 65 °C for 5 minutes followed by1 min chilling on ice. The denatured aRNA and second round primers were then incubated at 50 °C for 60 min in final reaction volume of 20 µl containing 500 µM dNTP mix, 1X first strand buffer, 5 µM DTT, 40 U Protector RNase Inhibitor (Roche, Germany) and 200 U SuperScript III reverse transcriptase (Invitrogen, California, USA). After which, an additional 200 U SuperScript III reverse transcriptase was added, before incubating the reaction at 50 °C overnight. The reaction was stopped by 15 min of incubation at 70 °C. cDNA samples were then treated with 2 units RNase H (Invitrogen, California, USA) at 37 °C for 30 min, then incubated at 65 °C for 5 min to stop the reaction. Two cDNA synthesis reactions were performed for each chosen RWC sample. After reverse transcription, the two reactions were pooled and quantified on a NanoDrop 1000 spectrophotometer (NanoDrop Technologies, Delaware, USA). A 50 ng/µl aliquot was made for each RWC cDNA sample. Each 50 ng/µl cDNA sample was stored in aliquots of 2 µl in individual tubes at -80 °C until needed, to avoid degradation of cDNA samples resulted from repeated freezing and thawing.

Standard curve sample preparation: Equal amount of cDNA samples from each of the RWC samples were pooled, and the concentration was quantified by NanoDrop 1000 spectrophotometer (NanoDrop Technologies, Delaware, USA). Several diluted aliquots of this pooled standard sample were made: 650, 50, 10, 1 and 0.5 ng/µl, to be used for generation of standard curves in RT-qPCR reactions. Each diluted standard sample was stored in aliquots of 2 µl in individual tubes at -80 °C until needed.

Gene-specific primer design: Gene-specific primers for each of the 24 selected contigs were designed by using the online primer designing program, Primer3 version 0.3.0 (http://frodo.wi.mit.edu; Rozen and Skaletsky, 2000). Contig sequences in FASTA format were supplied, and primer design parameters were specified to generate primers with Tm of approximately 60 °C; 18 to 25 bases in length; GC content between 40 and 60%; and amplicon of size 100 to 150 bp. The sequences of the RT-qPCR gene-specific primer pairs are summarized in Table A.4.1 in Appendix.

Quantitative real-time PCR: Quantitative real-time PCR was conducted in Rotor-Gene 2000 Real-Time Cycler (Corbett Research, Sydney, Australia). Three technical repeats were carried out for each biological repeat, in a reaction volume of 12.5 µl containing 50 ng RWC cDNA template, 200 nM forward gene-specific primer, 200 nM reverse gene-specific primer, 1X SensiMix (Quantace, London, UK) and 1X SYBR Green (Quantace, London, UK). For generation of standard curve, the reactions were performed in the same fashion but with 650 ng, 50 ng, 10 ng, 1 ng and 0.5 ng standard pool cDNA templates respectively. For the no template control (NTC) reaction, cDNA template was replaced with 1 µl H_2O. RT-qPCR was performed as follows: at 95 °C for 10 min (enzyme activation); for 50 cycles at 95 °C for 10 sec, 60 °C for 15 sec, 72 °C for 20 sec. Melting curve analysis was included after completion of RT-qPCR run Appendix (A.4.2).

Data analysis: The RT-qPCR data was captured and analyzed using Rotor-Gene 6 software. Threshold setting, standard curve generation, quantification of experimental reaction samples and export of the data as Microsoft Excel file, were carried out following manufacturer's instructions. Two normalization approaches were performed: (1) Normalization against starting amount of cDNA template: For each of the 22 contigs tested, the concentration value for all 36 samples (2 biological replicates X 3 technical replicates X 6 RWC points) was first logarithmically transformed (base 2). For each of the 12 biological replicates,

log_2[concentration] values of all 3 technical replicates were averaged. Subsequently, for each of the 6 RWC data points, Log_2[concentration] values of the 2 biological replicates were averaged. The data values were plotted and compared to the corresponding microarray results; (2) Normalization against non-differentially expressed contigs: For each of the 22 contigs tested, as well as the 2 non-differentially expressed contigs used for normalization purpose, the concentration value for all 36 samples was first logarithmically transformed (base 2). For each of the 22 tested contigs, the log_2[concentration] value of each of the 36 samples was subtracted by that of the corresponding samples of a non-differentially expressed contig (determining expression relative to the non-differentially expressed contig, equivalent to log_2[ratio]). Then the 3 technical log_2[ratio] values of each of the 12 biological replicates were averaged. Subsequently, for each of the 6 RWC data points, log_2[ratio] values of the 2 biological replicates were averaged. The data values were plotted and compared to the corresponding microarray results. The Spearman's rank coefficient between the RT-qPCR data and microarray data for each tested gene was calculated using the web based Spearman's Rank Correlation - Free Statistics and Forecasting Software (Calculator) available at http://www.wessa.net/rwasp_spearman.wasp/ (Wessa, 2012).

4.2.5 Analysis of LEA, antioxidant or transcription factor orthologue expressions identified during desiccation in X. humilis, and seed development and osmotic stress in A. thaliana

Selection of LEA orthologues: 11 of the 48 *X. humilis* LEA contigs were mapped with *A. thaliana* orthologue IDs (Chapter 3). Expression data of these orthologues were extracted from the relevant datasets (*X. humilis* desiccation profile, *A. thaliana* seed profile, or *A. thaliana* osmotic stress profile) and analyzed.

Selection of antioxidant orthologues: 15 of the 24 *X. humilis* antioxidant contigs were mapped with *A. thaliana* orthologue IDs (Chapter 3). Expression data of these orthologues were extracted from the relevant datasets and analyzed.

Selection of embryogenesis- or abiotic stress related transcription factor orthologues: 51 of the 93 *X. humilis* transcription factor contigs were mapped with *A. thaliana* orthologue IDs (Chapter 3). Amongst which, 19 orthologues were identified as transcription factors associated with abiotic stress or embryogenesis by a selection of relevant GO terms (Table 4.1) in Blast2GO. Expression data of these orthologues were extracted from the relevant datasets and analyzed.

Table 4.1. GO terms used in the identification of *X. humilis/A. thaliana* transcription factor orthologues involved in embryogenesis or abiotic stresses.

Selection by GO term	Term description
GO:0000303	response to superoxide
GO:0006970	response to osmotic stress
GO:0006972	hyperosmotic response
GO:0006979	response to oxidative stress
GO:0009266	response to temperature stimulus
GO:0009408	response to heat
GO:0009409	response to cold
GO:0009414	response to water deprivation
GO:0009555	pollen development
GO:0009567	double fertilization forming a zygote and endosperm
GO:0009651	response to salt stress
GO:0009737	response to abscisic acid stimulus
GO:0009738	abscisic acid mediated signaling pathway
GO:0009739	response to gibberellin stimulus
GO:0009788	negative regulation of abscisic acid mediated signaling pathway
GO:0009793	embryo development ending in seed dormancy
GO:0009960	endosperm development
GO:0010431	seed maturation
GO:0034605	cellular response to heat
GO:0042538	hyperosmotic salinity response
GO:0042542	response to hydrogen peroxide
GO:0048825	cotyledon development
GO:0071470	cellular response to osmotic stress
GO:2000693	positive regulation of seed maturation

4.3 Results and Discussion

4.3.1 Cluster analysis revealed a rapid shift in X. humilis transcriptome profile as leaves desiccate

Similar results were obtained for sample based clustering performed using 3 different algorithms: Sammon mapping (Fig. 4.1A), principle components analysis (PCA) (Fig. 4.1B), and divisive hierarchical analysis (DIANA) (Fig. 4.1C). The five biological repeats for the 100%, 60% RWC samples, and 20%/5% RWC samples formed discrete groups. However, the biological repeats for the 80% samples were split between 100% and 60%, and the biological repeats for the 40% RWC samples were split between the 60% and 20%/5% groupings. One of the biological repeats for the 80% RWC (81.1) clustered closely with the biological repeats for hydrated leaf tissue, three clustered together, more closely aligned with the 60% RWC, and one biological repeat (82.6) clustered between these groups. The hydrated leaf samples formed a distinctive clade on the dendrogram from the DIANA analysis (Fig. 4.1C), which included two of 80% RWC samples (81.1 and 82.6). The closeness of these samples to the hydrated leaf hybridization was apparent in the overall colour signal of this slide (Appendix Fig. A.4.3). Three of the 40% biological repeats (41.9, 39.2 and 40.9) formed a discrete group, but other two biological repeats (42.1, 39.7) overlapped with the 20% and 5% samples.

Three distinct clades are defined in the divisive clustering analysis, with samples falling into either a hydrated clade, an early desiccation or late desiccation stage (Fig. 4.1C). We predict that a major switch from the hydrated to the desiccated program of gene expression occurs around 80% RWC in *X. humilis* leaves. This switch must be fairly rapid, as two of the 80% RWC samples clustered with the hydrated leaves, and the other three grouped with the other desiccating leaf samples. A second rapid transition is apparent in the late desiccation clade, where the biological repeats for the 40% RWC samples are on separate branches. Lastly, the close clustering of 20% and 5% RWC leaf samples suggests that transcriptional activities associated with desiccation tolerance in *X. humilis* leaves may have been fully established by the time the leaves reached the stage of 20% RWC, and that no further substantial changes occur in the final stages of desiccation.

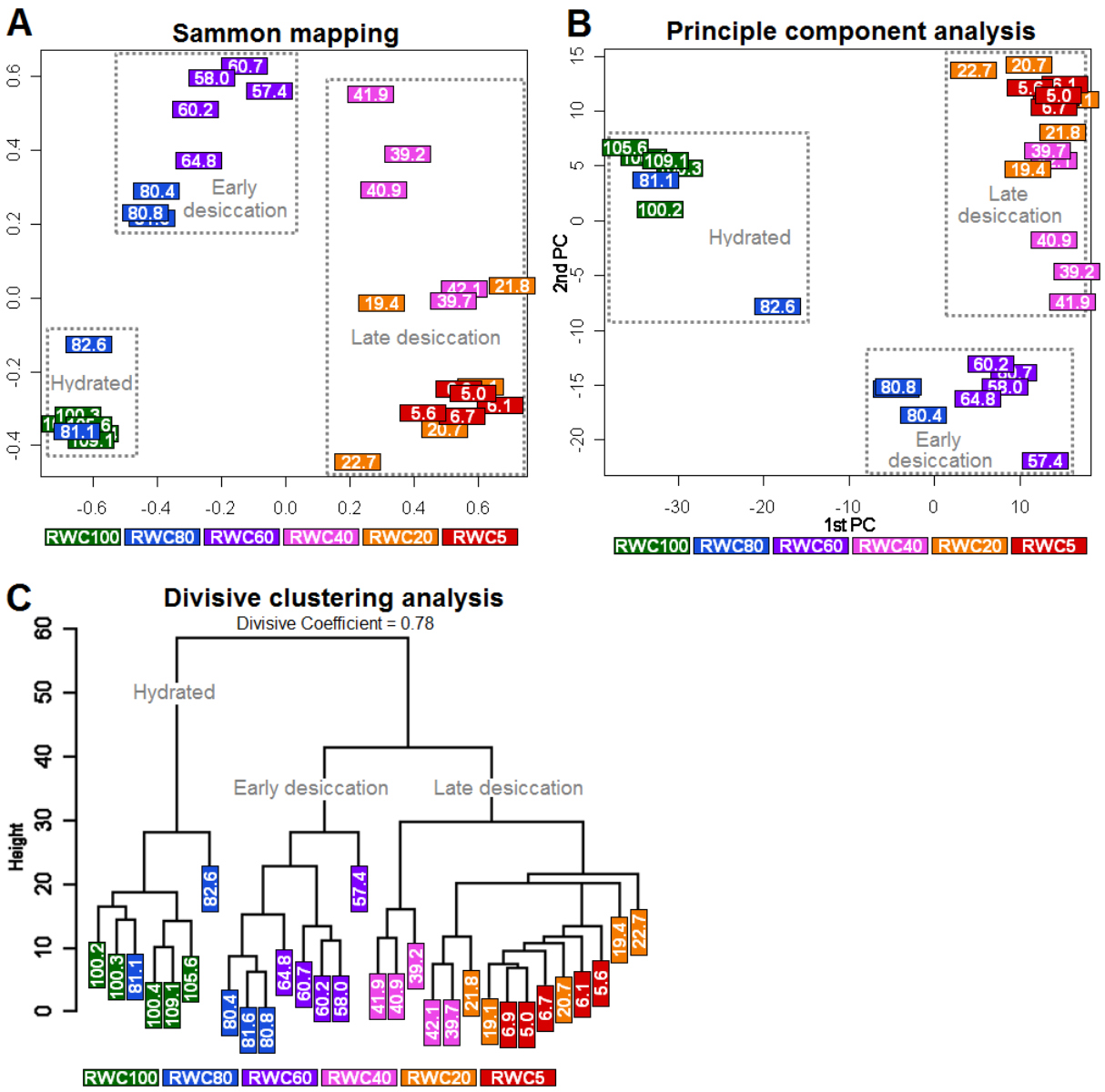


Figure 4.1. Cluster analysis of *X. humilis* leaf samples. The 5 biological repeats of *X. humilis* leaf samples representing 6 different stages of water loss: 100%, 80%, 60%, 40%, 20% and 5%RWC, were clustered based on the expression profiles of 1386 differentially expressed contigs identified using (A) Sammon mapping algorithm. Samples were projected into two-dimensional Sammon map based on their similarities and dissimilarities. The x- and y-axis of the Sammon map represented distances between the sample profiles in the two different data dimensions; (B) Principle component analysis (PCA). Samples were projected into two-dimensional PCA map based on their similarities and dissimilarities. The x- and y-axis of the PCA map represented the first and second components, or data dimensions that captured most variations among the expression profiles; (C) Divisive hierarchical analysis (DIANA). Samples were split as from one cluster to clusters containing single sample based on their similarities and dissimilarities. Height values represent diameters of the clusters prior to splitting. Divisive coefficient indicates the strength of the clustering structure found by the algorithm. The different RWC samples are coded with distinct colours, and the numbers indicated the actual %RWC figures of the leaf samples.

4.3.2 Cluster analysis identified different cohorts of genes differentially expressed in the X. humilis leaves during desiccation

The PAMSAM algorithm was used to cluster the 1361 differentially expressed *X. humilis* contigs into 7 groups (as predicted by Mclust algorithm) based on their expression patterns identified during the desiccation treatment (Fig. 4.2). Contigs whose mRNA transcript abundance declined during desiccation were grouped into clusters 2, and 3, which were more closely associated in PAMSAM two dimensional map, than the up-regulated clusters 1, 4, 5, and 6. The medoid of the down-regulated cluster 2 reached a minimum value at 60% RWC, while the medoid of the down-regulated cluster 3 only reached a minimum value at 40% RWC. Thus the mRNA transcript abundance for contigs in cluster 2 is more rapidly down-regulated in *X. humilis* leaves during desiccation than cluster 3. Although cluster 7 was grouped close to cluster 2 and 3, the medoid did not show a dramatic drop in expression during desiccation.

Four distinctive expression trends were observed among the desiccation up-regulated clusters. The mRNA transcript abundance for contigs grouped in clusters 6 and 5 showed a transient response to desiccation, with the medoid for contigs in cluster 6 reaching its maximum expression value more rapidly than the medoid in cluster 5. While the medoid for cluster 6 reached a maximum at 60% RWC, the medoid for cluster 5 only reached a maximum at 40% RWC, before declining. The mRNA transcripts for the majority of contigs in cluster 4 and cluster 5 were hardly expressed in hydrated leaves relative to the reference sample. The mediod profile for cluster 4 reached a maximum at 20% RWC, and remaining at high levels in fully desiccated leaves (5% RWC), representing a group of contigs which are specifically activated at late stages of desiccation. The medoid profile for cluster 1 differed, in that mRNA transcripts for contigs in this cluster were present at relatively high levels in hydrated leaves compared to the reference sample, and were initially down-regulated in response to desiccation. However, the mRNA transcripts for these contigs were subsequently up-regulated in the late phase of desiccation, with the medoid reaching a maximum at 5% RWC. These 4 up-regulated clusters can thus be ordered into successive responses to desiccation (6, 5, 4, then 1), on the basis of the maximum value of their medoid profile.

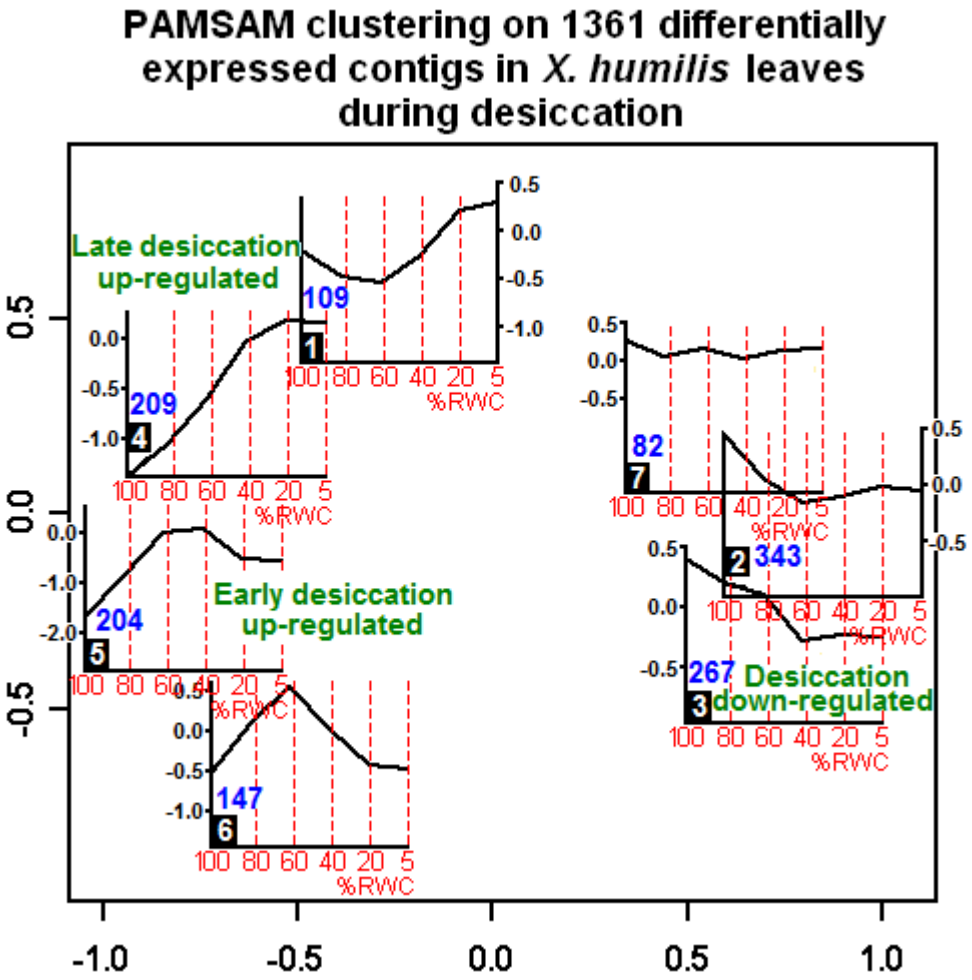


Figure 4.2. Cluster analysis of differentially expressed *X. humilis* genes. The 1361 differentially expressed contigs identified during desiccation treatment in *X. humilis* leaves were clustered into 7 clusters by the PAMSAM algorithm. These 7 medoids/clusters were partitioned and plotted in a two-dimensional Sammon map, indicating their approximate similarity to each other. Each medoid/cluster was represented by a small graph that showed the expression values in log2 ratios (y-axis of cluster graph) across the 6 different stages of water loss: 100%, 80%, 60%, 40%, 20% and 5%RWC during desiccation (x-axis of cluster graph). The x- and y-axis of the Sammon map represent distances between the medoids in the two different data dimensions. Total number of contigs in each cluster was indicated by the number above the cluster number.

4.3.3 Functional enrichment of the down-regulated clusters identified in X. humilis leaves during desiccation

GO terms associated with photosynthesis were identified as being significantly over-represented in cluster 2 by BiNGO (Fig. 4.3). These included GO terms for photosynthesis and chromophore-protein linkage (for biological process), chloroplast and thylakoids (for cellular component) and chlorophyll- and tetrapyrolle-binding (for molecular function).

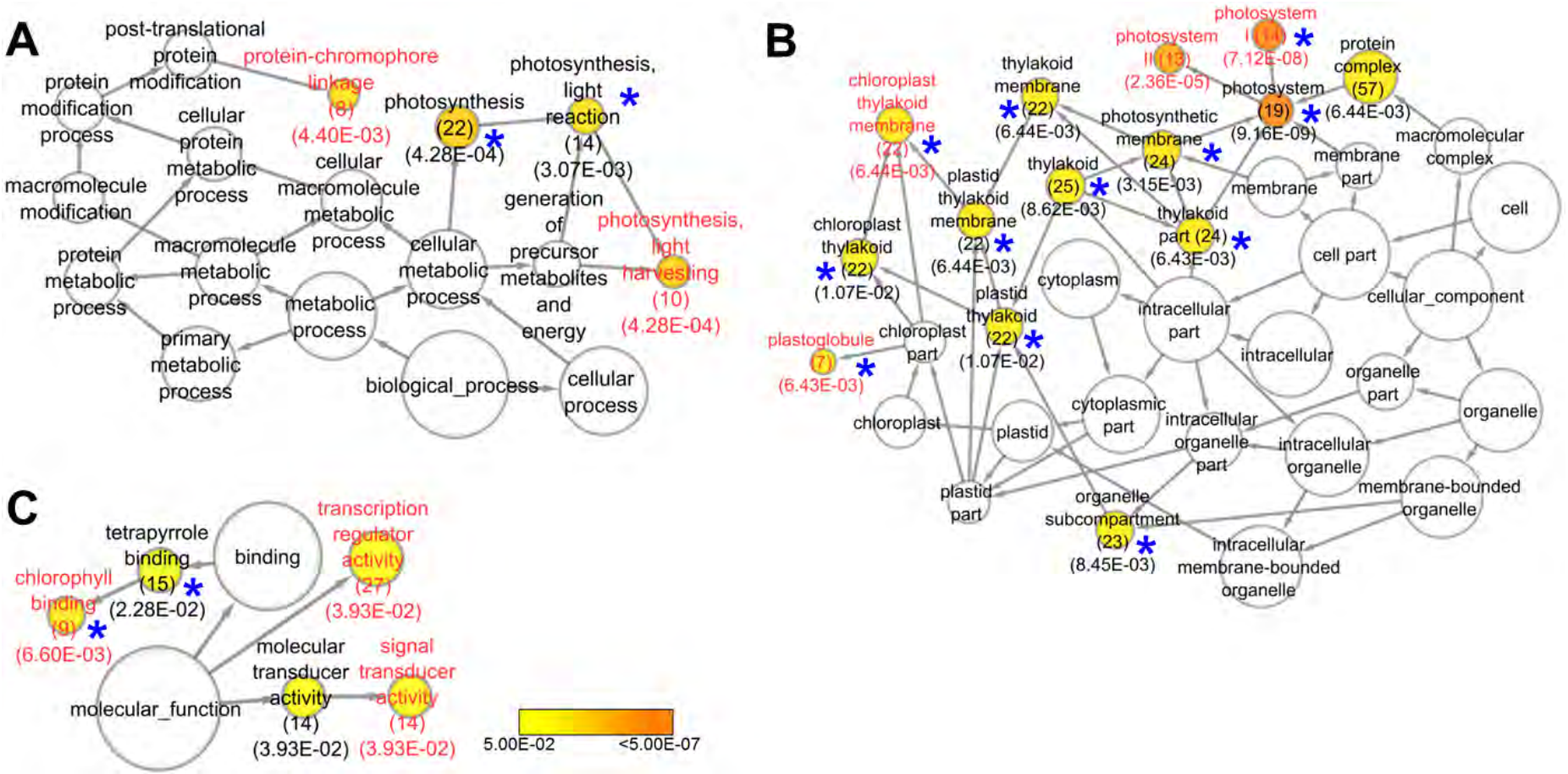


Figure 4.3. Over-representation of biological process (A), cellular component (B) and molecular function (C) ontology terms found in the set of 343 contigs from cluster 2 which were down-regulated in *X. humilis* leaves during desiccation. The enrichment of ontology terms was analyzed using hypergeometric test with a significance level of p-value= 0.05. The node size is proportional to the number of contigs in the test set being annotated to that node, and the significantly over-represented terms are presented as coloured nodes with a colour scale ranging from light yellow (p-value= 0.05) to dark orange (p-value< 5.00E-07). The two numbers shown in brackets under each significantly over-represented term indicate the number of contigs being annotated, and the p-value after FDR correction to that node respectively. The contigs annotated to the most specific terms (in red text) are summarized in Table 4.2. The blue asterisks indicate the common over-represented terms found in the set of 108 cluster 2 differentially expressed *X. humilis* orthologues.

Contigs annotated with the most specific GO terms linked to photosynthesis included several chlorophyll binding proteins, the subunits of the photosystem I and II light harvesting complexes and the early-light inducible protein (ELIP2) (contig XHP01661_1) (Table 4.2). The inclusion of contig XHP01661_1 in cluster 2 is unusual, as ELIPs have been reported to be up-regulated in plant chloroplasts during thylakoid biogenesis and under stressful conditions such as high light intensity (Hutin *et al.*, 2003). Free chlorophylls generate singlet oxygen molecules under strong light, and ELIPs are proposed to have a photoprotective function that involves binding of free chlorophyll to maintain a low level of free chlorophyll under high-light stress conditions (Hutin *et al.*, 2003). An ELIP-like gene was also reported induced in homoiochlorophyllous *C. plantagineum* upon water loss (Bartels *et al.*, 1992; Alamillo and Bartels, 1996). However, the down-regulation of XHP01661_1, ELIP2 upon desiccation is consistent with the breakdown of chlorophylls and photosynthetic apparatus in poikilochlorophyllous *X. humilis* (Collett *et al.*, 2003).

Both transcription regulator activity and signal transduction activity were identified as being significantly enriched under molecular function (Fig 4.3C). Several receptor-like kinases, MAP and protein kinases, as well as a large number of transcriptional regulators are down-regulated (Table 4.2). These include possible inducers of genes involved in photosynthesis, such as light-harvesting-like protein LIL3:1 (Xh_LRF_02G038) that plays an essential role in chlorophyll and tocopherol biosynthesis (Tanaka *et al.*, 2010; Takahashi *et al.*, 2014). A number of transcriptional repressors that are active in hydrated tissue were also found down-regulated during desiccation (Table 4.2, indicated in red). These include ethylene insensitive 4 (EIN4, XHP01625_1), ethylene responsive element binding factor 4 (ERF4, XHP00447_2), FERONIA (XHP01147_1), TOPLESS (TPL, XHP01131_1), and a zinc-finger protein 2 (XHP00948_1) (Table 4.2). EIN4 is an ethylene receptor with serine kinase activity which negatively regulates ethylene signaling (Hua *et al.*, 1998; Sakai *et al.*, 1998). Plants synthesize and accumulate phytohormone ethylene when exposed to various abiotic or biotic stress conditions (Morgan and Drew, 1997). In addition to ABA, the accumulated ethylene initiates protective responses, as well as the reduction in the plant growth. ERF4 has been reported to negatively modulate the ethylene and ABA responses in *A. thaliana* (Yang *et al.*, 2005). FERONIA encodes a plasma membrane localized receptor-like kinase, and has been reported as a negative regulator of ABA signaling, suppressing ABA response through activation of ABI2 (Yu *et al.*, 2012). In plant seeds, TPL represses the expression of root-promoting genes in the top half of the embryo, ensuring a correct differentiation of the shoot

pole during embryogenesis (Long *et al.*, 2006; Smith and Long, 2010). TPL proteins have also been reported having corepressor interaction with several transcription factors involved in abiotic stress responses, with ERF4, as well as ABI5 binding proteins (AFPs) that negatively regulates ABA signaling by promoting the degradation of ABI5 (Causier *et al.*, 2012). These suggested that ethylene and ABA signaling pathways in *X. humilis* during the early stages of desiccation may be up-regulated via the down-regulation of their negative regulators.

In *A. thaliana*, zinc-finger protein 2 (AZF2) is induced by various abiotic stresses such as drought and cold, as well as ABA or ethylene treatment. AZF has been suggested to suppress many ABA repressive genes including genes involved in photosynthesis and carbohydrate metabolism (Seki *et al.*, 2002; Sakamoto *et al.*, 2004; Kodaira *et al.*, 2011). However, the *X. humilis* AZF2 orthologue was shown to be down-regulated in response to desiccation. This down-regulation of AZF2 may be involved in the up-regulation of several photosynthesis related genes, as well as genes required for re-establishment of growth and development, whose mRNA transcripts are stored in desiccated leaves for translation upon rehydration.

Table 4.2. Mapping of contig IDs to the most specific over-represented GO terms that are significantly enriched in Cluster 2 identified in *X. humilis* leaves during desiccation.

GO ID	GO description	GO category	Corrected p-value	No of contig	Contig ID	Contig description
GO:0009765	photosynthesis, light harvesting	BP	4.28E-04	10	XHP00116_1	chlorophyll A/B binding protein 1
					XHP00708_1	Lhca2 protein
					XHP01629_1	light harvesting chlorophyll a b-binding protein
					XHP01091_1	light-harvesting chlorophyll B-binding protein 3
					XHP00345_1	light-harvesting chlorophyll-protein complex I subunit A4
					XH_LRF_02G038	LIL3:1; transcription factor
					XHP01029_1	photosystem I light harvesting complex gene 1
					XHP00987_1	photosystem I light harvesting complex gene 3
					XHP01245_1	photosystem I subunit O
					XHP00122_1	photosystem II light harvesting complex gene 2.3
GO:0018298	protein-chromophore linkage	BP	4.40E-03	8	XHP00116_1	chlorophyll A/B binding protein 1
					XHP00708_1	Lhca2 protein
					XHP01629_1	light harvesting chlorophyll a b-binding protein
					XHP01091_1	light-harvesting chlorophyll B-binding protein 3
					XHP00345_1	light-harvesting chlorophyll-protein complex I subunit A4
					XHP01029_1	photosystem I light harvesting complex gene 1
					XHP00987_1	photosystem I light harvesting complex gene 3
					XHP00122_1	photosystem II light harvesting complex gene 2.3
GO:0009522	photosystem I	CC	7.12E-08	14	XHP00116_1	chlorophyll A/B binding protein 1
					XHP00708_1	Lhca2 protein
					XHP01629_1	light harvesting chlorophyll a b-binding protein
					XHP01091_1	light-harvesting chlorophyll B-binding protein 3
					XHP00345_1	light-harvesting chlorophyll-protein complex I subunit A4
					XHP01029_1	photosystem I light harvesting complex gene 1
					XHP00987_1	photosystem I light harvesting complex gene 3
					XHP00092_1	photosystem I reaction center subunit PSI-N, chloroplast, putative / PSI-N, putative (PSAN)
					XHP00727_1	photosystem I subunit F

Table 4.2. (continued)

GO ID	GO description	GO category	Corrected p-value	No of contig	Contig ID	Contig description
GO:0009522 (continued)	photosystem I	CC	7.12E-08	14	XHP00245_2	photosystem I subunit G
					XHP00710_1	photosystem I subunit l
					XHP01245_1	photosystem I subunit O
					XHP00122_1	photosystem II light harvesting complex gene 2.3
					XHP01051_1	PSAH-1 (photosystem I subunit H-1)
GO:0009523	photosystem II	CC	2.36E-05	13	XHP00116_1	chlorophyll A/B binding protein 1
					XHP00708_1	Lhca2 protein
					XHP01629_1	light harvesting chlorophyll a b-binding protein
					XHP01091_1	light-harvesting chlorophyll B-binding protein 3
					XHP00345_1	light-harvesting chlorophyll-protein complex I subunit A4
					XHP01981_1_M	oxygen-evolving enhancer protein 3, chloroplast, putative (PSBQ1) (PSBQ)
					XHP01029_1	photosystem I light harvesting complex gene 1
					XHP00987_1	photosystem I light harvesting complex gene 3
					XHP00229_2	photosystem ii core complex proteins psby
					XHP00122_1	photosystem II light harvesting complex gene 2.3
					XHP01894_1	PS II oxygen-evolving complex 1
					XHP00819_1	PSBR (photosystem II subunit R)
					XHP00446_2	PSBW (PHOTOSYSTEM II REACTION CENTER W)
GO:0010287	plastoglobule	CC	6.43E-03	7	XHP00345_1	light-harvesting chlorophyll-protein complex I subunit A4
					XHP01029_1	photosystem I light harvesting complex gene 1
					XHP00987_1	photosystem I light harvesting complex gene 3
					XHP00727_1	photosystem I subunit F
					XHP00710_1	photosystem I subunit l
					XH_RDR_31C123	Plastid-lipid associated protein PAP / fibrillin family protein
					XHP01051_1	PSAH-1 (photosystem I subunit H-1)
GO:0009535	chloroplast thylakoid membrane	CC	6.44E-03	22	XHP01803_1	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein
					XHP00116_1	chlorophyll A/B binding protein 1
					XH_RRF_01D058	FNR2 (FERREDOXIN-NADP(+)-OXIDOREDUCTASE 2)
					XHP00708_1	Lhca2 protein

Table 4.2. (continued)

GO ID	GO description	GO category	Corrected p-value	No of contig	Contig ID	Contig description
GO:0009535 (continued)	chloroplast thylakoid membrane	CC	6.44E-03	22	XHP01629_1	light harvesting chlorophyll a b-binding protein
					XHP01091_1	light-harvesting chlorophyll B-binding protein 3
					XHP00345_1	light-harvesting chlorophyll-protein complex I subunit A4
					XH_LRF_02G038	LIL3:1; transcription factor
					XHP01981_1_M	oxygen-evolving enhancer protein 3, chloroplast, putative (PSBQ1) (PSBQ)
					XHP01029_1	photosystem I light harvesting complex gene 1
					XHP00987_1	photosystem I light harvesting complex gene 3
					XHP00092_1	photosystem I reaction center subunit PSI-N, chloroplast, putative / PSI-N, putative (PSAN)
					XHP00727_1	photosystem I subunit F
					XHP00245_2	photosystem I subunit G
					XHP00710_1	photosystem I subunit l
					XHP00122_1	photosystem II light harvesting complex gene 2.3
					XH_RDR_31C12 3	Plastid-lipid associated protein PAP / fibrillin family protein
					XHP01894_1	PS II oxygen-evolving complex 1
					XHP01051_1	PSAH-1 (photosystem I subunit H-1)
					XHP00819_1	PSBR (photosystem II subunit R)
					XHP00446_2	PSBW (PHOTOSYSTEM II REACTION CENTER W)
					XHP00290_1	Transketolase
GO:0016168	chlorophyll binding	MF	6.60E-03	9	XHP00116_1	chlorophyll A/B binding protein 1
					XHP01661_1	ELIP2 (EARLY LIGHT-INDUCIBLE PROTEIN 2); chlorophyll binding
					XHP00708_1	Lhca2 protein
					XHP01629_1	light harvesting chlorophyll a b-binding protein
					XHP01091_1	light-harvesting chlorophyll B-binding protein 3
					XHP00345_1	light-harvesting chlorophyll-protein complex I subunit A4
					XHP01029_1	photosystem I light harvesting complex gene 1
					XHP00987_1	photosystem I light harvesting complex gene 3
					XHP00122_1	photosystem II light harvesting complex gene 2.3

Table 4.2. (continued)

GO ID	GO description	GO category	Corrected p-value	No of contig	Contig ID	Contig description
GO:0004871	signal transducer activity	MF	3.93E-02	14	XH_LDR_04D043	ATPERK1 (PROLINE EXTENSIN-LIKE RECEPTOR KINASE 1)
					XH_LRF_01G058	Bax inhibitor-1 family protein
					XHP01625_1	ETHYLENE INSENSITIVE 4
					XHP01147_1	FERONIA
					XHP00872_1	heat- and acid-stable phosphoprotein*
					XHP01528_1	leucine-rich repeat transmembrane protein kinase, putative
					XHP01829_1	map kinase
					XHP00612_1	protein kinase
					XHP00542_1	protein kinase family protein
					XHP02033_1_M	Protein kinase superfamily protein
					XHP01320_1	receptor-like protein kinase-like protein
					XH_RRF_01C068	SPX domain gene 1
					XHP01743_1	translocase of the outer mitochondrial membrane 40
					XHP01268_1	unknown protein
GO:0030528	transcription regulator activity	MF	3.93E-02	27	XHP01060_1	ap2 domain containing protein
					XHP00635_2	Arabidopsis 6B-interacting protein 1-like 2
					XHP00496_1	ATNAC2 (ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 2); transcription factor
					XHP01176_1	Basic helix-loop-helix (bHLH) DNA-binding family protein
					XHP01425_1	basic helix-loop-helix (bHLH) family protein
					XHP00580_1	BEL1-like homeodomain 7, BLH7
					XH_LDR_10D023	cbf-like transcription factor
					XHP02014_1_M	class iii homeodomain-leucine zipper protein c3hdz1
					XHP01185_1	Duplicated homeodomain-like superfamily protein
					XHP00447_2	ERF4 (ETHYLENE RESPONSIVE ELEMENT BINDING FACTOR 4); transcription repressor
					XHP01625_1	ETHYLENE INSENSITIVE 4
					XHP01410_1	ethylene-insensitive 3f
					XHP01164_1	ETHYLENE-INSENSITIVE3
					XHP00872_1	heat- and acid-stable phosphoprotein
					XHP01717_1	homeobox 1
					XHP01979_1_M	KIWI; DNA binding / protein binding / transcription coactivator
					XH_LRF_02G038	LIL3:1; transcription factor
					XHP00684_1	MBF1B (MULTIPROTEIN BRIDGING FACTOR 1B); DNA binding / transcription coactivator

Table 4.2. (continued)

GO ID	GO description	GO category	Corrected p-value	No of contig	Contig ID	Contig description
GO:0030528 (continued)	transcription regulator activity	MF	3.93E-02	27	XHP01085_1	redox responsive transcription factor 1
					XHP01087_1	serine acetyltransferase 2;1
					XHP01184_1	SZF1 (SALT-INDUCIBLE ZINC FINGER 1); transcription factor
					XHP01131_1	TPL (TOPLESS); transcription repressor
					XHP01039_1	transcriptional regulator family protein
					XHP01019_1	unknown protein
					XHP01046_1	YABBY2
					XHP00948_1	zinc-finger protein 2
					XHP01277_1	ZML1 (ZIM-LIKE 1)

BP: biological process; CC: cellular component; MF: molecular function. Contigs shown in red represent transcription repressors based on their GO annotation in Blast2GO.

Contigs in the more slowly down-regulated cluster 3 was enriched with GO terms involved in processes of cellular developmental processes, and regulation of protein metabolism, and amino acid catabolism, as well as DNA postreplication repair (Fig. 4.4A). The ubiquitin-conjugating complex and UBC13-MMS2 complex, an ubiquitin conjugating enzyme complex that acts as a signal to promote error-free DNA postreplication repair, were significantly enriched cellular components for cluster 3 (Fig. 4.4B).

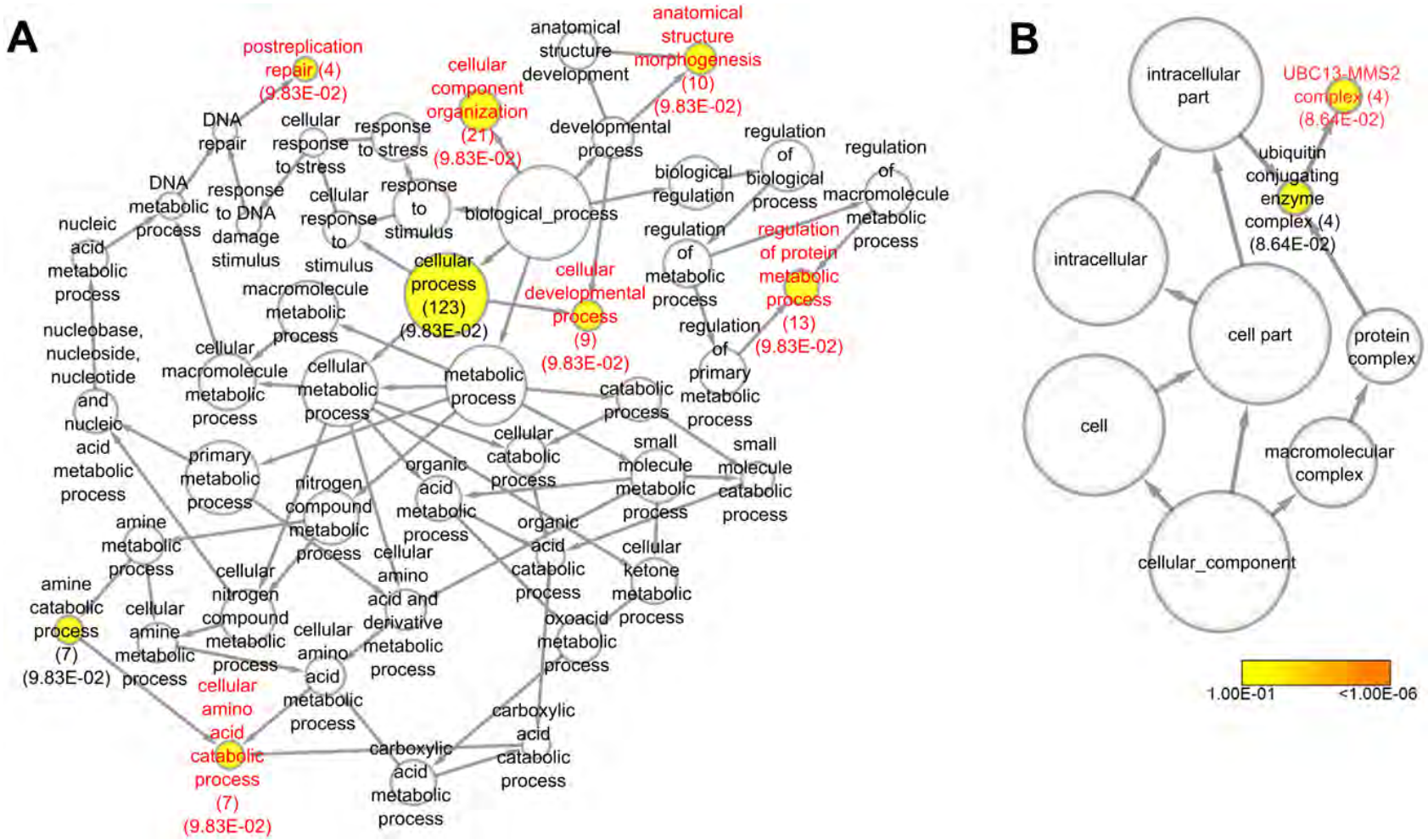


Figure 4.4. Over-representation of biological process (A) and cellular component (B) ontology terms found in the set of 267 cluster 3 differentially expressed contigs identified in *X. humilis* leaves during desiccation. The enrichment of ontology terms was analyzed using hypergeometric test with a significance level of p-value= 0.1. The node size is proportional to the number of contigs in the test set being annotated to that node, and the significantly over-represented terms are presented as coloured nodes with a colour scale ranging from light yellow (p-value= 0.1) to dark orange (p-value< 1.00E-06). The two numbers shown in brackets under each significantly over-represented term indicate the number of contigs being annotated, and the p-value after FDR correction to that node respectively. The contigs annotated to the most specific terms (in red text) are summarized in Table 4.3.

Contigs that annotated with the most specific GO terms linked to cellular development, DNA repair, and protein metabolic processes included several histone proteins (i.e. H2A and H4), and the ubiquitin-conjugating enzymes (Table 4.3). It has been reported that under stress conditions, protein ubiquitination plays an important role in the degradation of misfolded proteins as well as the repressors of the genes required for adaption to stress (Lyzenga and Stone, 2012). Degradation of target proteins via the ubiquitin proteasome system begins with the activation of ubiquitin protein by ubiquitin-activating enzyme (E1). The activated ubiquitin is then transferred to ubiquitin-conjugating enzyme (E2), and subsequently transferred to the target protein through interaction with ubiquitin ligase (E3). Additional ubiquitin molecules can be added to the mono-ubiquitinated target protein (polyubiquitination) before the degradation by proteasomes. In addition to the degradation, the ubiquitinated protein can also be destined for endocytosis, protein activation, intracellular trafficking, membrane protein internalization, vesicle sorting, DNA repair, and gene silencing depending on the topology or the degree of polyubiquitination (Pickart and Fushman, 2004; Kirkpatrick *et al.*, 2006; Kim *et al.*, 2007; Mukhopadhyay and Riezman, 2007). Whilst target protein specificity is controlled mainly by the E3, the topology of polyubiquitination is determined by E2, or E3, or the combinations of E2 and E3 (Kim *et al.*, 2007; Deshaies and Joazeiro, 2009; Rodrigo-Brenni *et al.*, 2010).

The fate of target proteins is determined by different mode of ubiquitination. As reviewed in Sadowski *et al.* (2012), monoubiquitination on the target protein regulates DNA repair, viral budding, gene expression and endocytosis; Monoubiquitnation on multiple sites of can regulate receptor endocytosis; Polyubiquitination through lysine residue 11 (K11) or K48 of ubiquitins results in proteasomal degradation; Polyubiquitination through K63 of ubiquitins can function in DNA damage tolerance, signal transduction, kinase activation and endocytosis; A linear polyubiquitin chain generated through the α-amino group of the N-terminal methionine residue results in the activation of NF-kB to initiate transcription; Branched Ub structures via K6, K27, and K48 of ubiquitins through autoubiquitination can modulate the RING E3 Ring1B ligase activity to induce histone H2A monoubiquitination. The role of K6-, K27-, K29-, and K33-linked polyubiquitin chains remain unknown.

Table 4.3. Mapping of contig IDs to the most specific over-represented GO terms that are significantly enriched in Cluster 3 identified in *X. humilis* leaves during desiccation.

GO ID	GO description	GO category	Corrected p-value	No of contig	Contig ID	Contig description
GO:0006301	postreplication repair	BP_0.1	9.83E-02	4	XHP00915_1	MMS ZWEI homologue 1
					XHP01316_1	MMS ZWEI homologue 3
					XHP01385_1	ubiquitin-conjugating enzyme 35
					XHP01255_1	ubiquitin-conjugating enzyme family protein
GO:0009063	cellular amino acid catabolic process	BP_0.1	9.83E-02	7	XHP01658_1	anti-oxidant 1, ATX1
					XHP00984_1	arginine decarboxylase 1
					XHP02013_1_M	isovaleryl- dehydrogenase
					XHP01086_1	multifunctional protein 2
					XHP02042_1_M	peroxisomal 3-ketoacyl-CoA thiolase 3
					XH_RDF_26G118	Single hybrid motif superfamily protein
					XHP00854_1	SWINGER
GO:0009653	anatomical structure morphogenesis	BP_0.1	9.83E-02	10	XHP01658_1	anti-oxidant 1, ATX1
					XH_RRF_01H098	ATP-binding cassette I19
					XHP00894_1	casein kinase I
					XHP01500_1	cyclase associated protein 1
					XHP02058_1_M	MSI4/FVE
					XHP01702_1	NRP2 (NAP1-RELATED PROTEIN 2); histone binding
					XHP00372_1	PCK1 (PHOSPHOENOLPYRUVATE CARBOXYKINASE 1)
					XHP01307_1	PYM (Partner of Y14-Mago); protein binding
					XHP00960_1	ribosomal protein S13A
					XHP01385_1	ubiquitin-conjugating enzyme 35
GO:0016043	cellular component organization	BP_0.1	9.83E-02	21	XHP01466_1	actin-related protein 4
					XH_RRF_01H098	ATP-binding cassette I19
					XHP00894_1	casein kinase I
					XHP02053_1_M	catalytic
					XHP00941_1	CLP protease proteolytic subunit 6
					XHP01500_1	cyclase associated protein 1
					XHP01647_1	dynamin-like 3
					XHP00877_1	EER4 (ENHANCED ETHYLENE RESPONSE 4); DNA binding / transcription initiation factor
					XHP00636_1	glycine-rich protein precursor
					XHP00578_1	high mobility group B3
					XHP00529_1	Histone H2A
					XHP01135_1	Histone H4
					XHP00335_1	Histone superfamily protein
					XHP02058_1_M	MSI4/FVE
					XHP01702_1	NRP2 (NAP1-RELATED PROTEIN 2); histone binding
					XHP02042_1_M	peroxisomal 3-ketoacyl-CoA thiolase 3
					XHP02018_1_M	protein arginine methyltransferase 4A
					XHP01100_1	PSBR (photosystem II subunit R)

Table 4.3. (continued)

GO ID	GO description	GO category	Corrected p-value	No of contig	Contig ID	Contig description
GO:0016043 (continued)	cellular component organization	BP_0.1	9.83E-02	21	XHP01307_1	PYM (Partner of Y14-Mago); protein binding
					XHP00960_1	ribosomal protein S13A
					XHP01127_1	single myb histone 6
GO:0048869	cellular developmental process	BP_0.1	9.83E-02	9	XHP01466_1	actin-related protein 4
					XH_RRF_01H098	ATP-binding cassette I19
					XHP00894_1	casein kinase I
					XHP01500_1	cyclase associated protein 1
					XHP02058_1_M	MSI4/FVE
					XHP01702_1	NRP2 (NAP1-RELATED PROTEIN 2); histone binding
					XHP01307_1	PYM (Partner of Y14-Mago); protein binding
					XHP00960_1	ribosomal protein S13A
					XHP01385_1	ubiquitin-conjugating enzyme 35
GO:0051246	regulation of protein metabolic process	BP_0.1	9.83E-02	13	XHP00681_1	ARM repeat superfamily protein
					XHP00324_1	elongation factor 1B alpha-subunit 1 (eEF1Balpha1)
					XHP00915_1	MMS ZWEI homologue 1
					XHP01316_1	MMS ZWEI homologue 3
					XHP00108_6	protein translation factor sui1
					XHP02051_1_M	RAB GTPase homolog E1B
					XHP01385_1	ubiquitin-conjugating enzyme 35
					XHP01567_1	ubiquitin-conjugating enzyme e2-17 kda
					XHP01255_1	ubiquitin-conjugating enzyme family protein
					XHP01082_1	ubiquitin-conjugating enzyme, putative
					XHP00185_2	ubiquitin-conjugating enzyme, putative
					XHP00659_1	ubiquitin-conjugating enzyme, putative
					XHP00677_1	ubiquiting-conjugating enzyme 2
GO:0031372	UBC13-MMS2 complex	CC_0.1	8.64E-02	4	XHP00915_1	MMS ZWEI homologue 1
					XHP01316_1	MMS ZWEI homologue 3
					XHP01385_1	ubiquitin-conjugating enzyme 35
					XHP01255_1	ubiquitin-conjugating enzyme family protein

BP: biological process; CC: cellular component; _0.1: A significance level of p-value= 0.1 was used during the hypergeometric test.

Ubiquitination of histone proteins have been identified to play critical roles in DNA repair, transcription initiation and elongation, as well as the repression of gene expressions by the polycomb repressive complex proteins (PRC) (Wang *et al.*, 2004b; Weake and Workman, 2008; Simon and Kingston, 2009; Lehmann *et al.*, 2012). The PRC1 and PCR2 proteins play a crucial role in maintaining the repression of genes in animals and plants that are not required in a specific differentiation status (Calonje, 2014). PRC1 possesses an E3 ligase activity for the monoubiquitination of histone H2A (H2Aub) and PRC2 has the histone H3 lysine 27 trimethyltransferase (H3K27me3) activity. Both histone modifications are required

for the repression of genes (Yang *et al.*, 2013; Calonje, 2014). In *A. thaliana*, the orthologues of components of PRC1 have been identified, and have been shown to have crucial functions in the stable repression of embryonic traits and regulatory genes during somatic growth (Chen *et al.*, 2010). Furthermore, it has been shown that the PRC1 mediated H2Aub is required in initiating the repression of seed maturation genes, and this repression is maintained by PRC2 mediated H3K27me3 (Yang *et al.*, 2013; Calonje, 2014). In contrast to the studies that reported up-regulation of ubiquitin-conjugating enzymes in *Glycine max* during drought stress, and during the final two stages of seed development (Jones *et al.*, 2010), as well as the report of the enhanced drought tolerance found in *A. thaliana* by overexpressing of *G. max* or *Arachis hypogaea* ubiquitin-conjugating enzymes (Wan et al., 2010; Zhou et al., 2010), the 9 contigs annotated as ubiquitin-conjugating enzyme related genes in *X. humilis* were found down-regulated in the leaves during desiccation. These 9 contigs can be categorized into 3 groups based on their functions (Table 4.4).

Table 4.4. Down-regulated ubiquitin conjugating enzymes found in *X. humilis* during desiccation.

Group	Contig ID	Best *A. thaliana* BLASTP hit	Gene name	Function
1	XHP00915_1	MMS ZWEI-like protein 1	MMSZ1	Linked to the DNA damage tolerance response in yeast. Targeting ubiquitination of Lysine 63
	XHP01255_1	MMS ZWEI-like protein 3	MMSZ3	
	XHP01316_1	MMS ZWEI-like protein 3	MMSZ3	
	XHP10385_1	ubiquitin-conjugating enzyme E2 35	UBC35	
2	XHP00677_1	Ubiquitin conjugating enzyme E2 A	UBC2	Linked directly to histone H2B monoubiquitination.
3	XHP00185_2	Ubiquitin conjugating enzyme E2 28	UBC28	Linked to target protein degradation.
	XHP00659_1	Ubiquitin conjugating enzyme E2 28	UBC28	
	XHP01082_1	Ubiquitin conjugating enzyme E2 28	UBC28	
	XHP01567_1	Ubiquitin conjugating enzyme E2 11; (2nd best hit = UBC28)	UBC11	

The first group included the 3 MMS ZWEI genes and UBC35 associated with UBC13-MMS2 complex (Table 4.3; Table 4.4) which may play a role in DNA damage responses and error-free post-replicative DNA repair via K63-based polyubiquitination reactions (Wen *et al.*, 2008) (Table 4.3). The DNA damage repair process observed in *X. humilis* may be down-

regulated in concurrence with the arrest in cell mitosis and DNA replication during desiccation.

The function of histone H2B monoubiquitination in plants is not completely understood, however it has been reported to play an essential role in chromatin remodeling in seed dormancy (Liu *et al.*, 2007; Van Zanten *et al.*, 2013). In *A. thaliana*, loss of H2B monoubiquitination resulted in altered expression levels for several dormancy-related genes (Liu *et al.*, 2007). The *X. humilis* UBC2 orthologue may specifically be involved in the H2B monoubiquitination associated in the activation of dormancy related genes, which was down-regulated to prevent the plant to enter a dormant state similarly observed in seeds, ensuring a rapidly resumption of normal functioning upon rehydration.

Absence of H2B monoubiquitination showed altered expression levels for many dormancy-related genes (Liu et al., 2007). The 4 E2 genes in cluster 3 were reported to be involved in the response to misfolded protein, as well as in the proteasome core complex assembly, which may implicate their involvement in target protein degradation (Heyndrickx and Vandepoele, 2012). These 4 E2 contigs may possibly possess a putative role responsible for the degradation or turnover of some desiccation specific proteins expressed during hydrated states in *X. humilis*.

In addition, these 9 ubiquitin-conjugating enzymes may possibly have roles in silencing the expression of desiccation specific genes in *X humilis* leaves during hydrated state. As cellular water content decreases, they are down-regulated to increase the abundance of the stress specific molecules. Furthermore, these down-regulated ubiquitin-conjugation enzymes may be specifically participated in the reaction of monoubiquitination of H2A. Decrease in H2Aub may also result in de-repression of essential stress related, or embryogenesis genes that confer desiccation tolerance in *X. humilis* leaves during desiccation.

Functional enrichment analysis of the clusters 2 and 3 desiccation down-regulated contigs suggested that during desiccation, photosynthesis and processes related to cellular development, as well as chromatin modification and protein turnover were primarily and immediately deactivated upon water loss. This deactivation was accompanied with large changes in transcriptional regulations of genes. In addition to the down-regulation of transcription activators of genes associated with normal cellular development, many

transcription repressors were also found to be down-regulated. Together with the down-regulation of ubiquitin-conjugating enzymes, the data suggested that in addition to the activation of stress specific genes, the de-repression of the ethylene and ABA-signaling pathways may also play a primary role in the onset of desiccation tolerance in *X. humilis* leaves.

4.3.4 Functional enrichment of the early up-regulated clusters identified in X. humilis leaves during desiccation

Cluster 6, representing the earliest group of desiccation up-regulated contigs, was enriched with contigs associated with ribosome structures and functions that included the biological processes of ribosome biogenesis and translation (Fig. 4.5A), the cellular components of cytosolic ribosome and ribosomal unit (Fig. 4.5B) and molecular functions involved in the structural constituent and activity of ribosome (Fig. 4.5C).

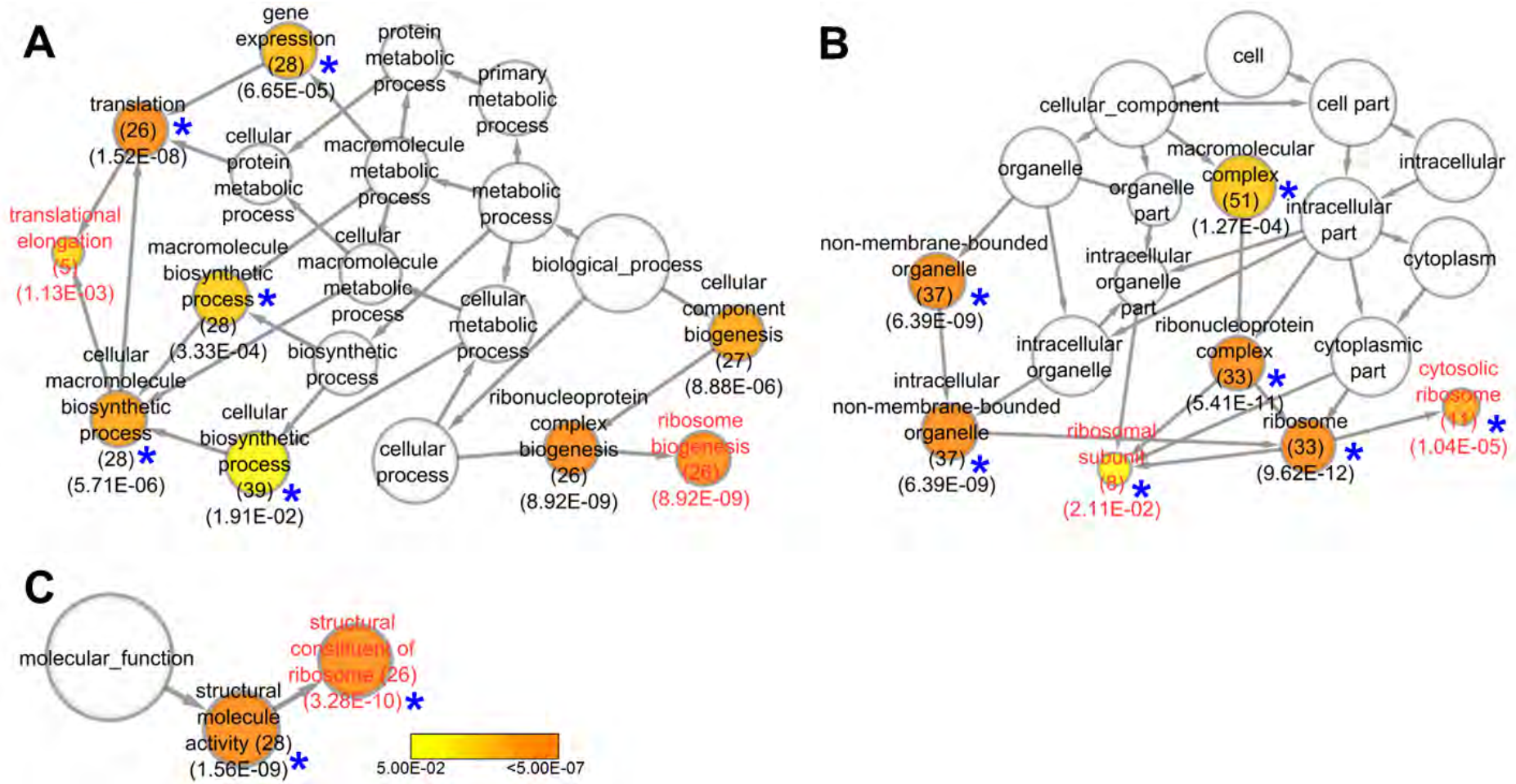


Figure 4.5. Over-representation of biological process (A), cellular component (B) and molecular function (C) terms found in the set of 147 cluster 6 differentially expressed contigs identified in *X. humilis* leaves during desiccation. The enrichment of ontology terms was analyzed using hypergeometric test with a significance level of p-value= 0.05. The node size is proportional to the number of contigs in the test set being annotated to that node, and the significantly over-represented terms are presented as coloured nodes with a colour scale ranging from light yellow (p-value= 0.05) to dark orange (p-value< 5.00E-07). The two numbers shown in brackets under each significantly over-represented term indicate the number of contigs being annotated, and the p-value after FDR correction to that node respectively. The contigs annotated to the most specific terms (in red text) are summarized in Table 4.5. The blue asterisks indicate the common over-represented terms found in the set of 85 cluster 6 differentially expressed *X. humilis* orthologues.

Contigs associated to these over-represented terms mainly included the different subunits of ribosome (Table 4.5). Proteomic analysis of leaf proteins in *X. viscosa* has suggested that activation of the desiccation response is accompanied by a switch in the proteome of Xerophyta, involving the turnover of proteins, and the simultaneous synthesis of proteins required for protection (Ingle *et al.*, 2007). A functional enrichment for the components of ribosomes and translation observed in cluster 6 was in support of this hypothesis. The rapid accumulation of ribosome-related transcripts upon desiccation suggested that the overall translation activity may have increased during the early phases of water loss. *De novo* protein synthesis of proteins involved in the protection of cellular content under desiccation would place a heavy burden on the cytoplasmic ribosomes, as the chemical environment and the energy level required by the translational processes are not optimal. Although the ribosomal protein mRNA transcripts levels declined to the level similar to or lower than that observed in the hydrated state at the late stages of desiccation, it is likely that the ribosomal proteins may be protected or stored during desiccated state. The initial stages of rehydration have been shown to be absolutely dependent on the rapid translation of stored mRNA transcripts of genes crucial for the re-establishment of normal cellular growth and functioning when water becomes available (Dace *et al.*, 1998). Lastly, the co-expressed 2 dehydrins (Class 2 LEA), XHP00531_2 and XHP00658_1, as well as the heat shock protein, XHP00429_2 found in this cluster may play a putative role in the protection and stabilization of the stored ribosomes under desiccated state.

Table 4.5. Mapping of contig IDs to the most specific over-represented GO terms that are significantly enriched in Cluster 6 identified in *X. humilis* leaves during desiccation.

GO ID	GO description	GO category	Corrected p-value	No of contig	Contig ID	Contig description
GO:0042254	ribosome biogenesis	BP	8.92E-09	26	XHP01934_1	60S acidic ribosomal protein family
					XHP01429_1	60S acidic ribosomal protein family
					XHP01489_1	60S acidic ribosomal protein P0
					XHP02016_1_M	60S acidic ribosomal protein P3 (RPP3B)
					XHP01657_1	60S ribosomal protein L13A (RPL13aD)
					XH_LDR_05F021	plastid ribosomal protein L35
					XHP00480_1	Ribosomal L29 family protein
					XHP01565_1	ribosomal protein L10 B
					XHP01558_1	Ribosomal protein L10 family protein
					XHP00958_1	Ribosomal protein L11 family protein
					XHP00753_1	Ribosomal protein L13 family protein
					XHP00348_1	Ribosomal protein L14p/L23e family protein
					XHP01769_1	ribosomal protein L18
					XHP00954_1	Ribosomal protein L22p/L17e family protein
					XH_RRF_01H088	Ribosomal protein L31e family protein
					XHP00311_1	Ribosomal protein L32e
					XHP00274_2	Ribosomal protein L6 family
					XHP01603_1	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein
					XHP00255_1	Ribosomal protein S10p/S20e family protein
					XHP01315_1	Ribosomal protein S26e family protein
					XHP01132_1	Ribosomal protein S30 family protein
					XHP00777_1	Ribosomal protein S3Ae
					XHP00771_1	Ribosomal protein S5 family protein
					XHP01317_1	Ribosomal protein S6b
					XHP00781_1	Ribosomal protein S8e family protein
					XHP01113_1	R-protein L3 B
GO:0006414	translational elongation	BP	1.13E-03	5	XHP01934_1	60S acidic ribosomal protein family
					XHP01429_1	60S acidic ribosomal protein family
					XHP01489_1	60S acidic ribosomal protein P0
					XHP02016_1_M	60S acidic ribosomal protein P3 (RPP3B)
					XHP01558_1	Ribosomal protein L10 family protein
GO:0022626	cytosolic ribosome	CC	1.04E-05	11	XHP01739_1	20S proteasome alpha subunit C1
					XHP01429_1	60S acidic ribosomal protein family
					XHP02016_1_M	60S acidic ribosomal protein P3 (RPP3B)
					XHP01657_1	60S ribosomal protein L13A (RPL13aD)
					XHP00480_1	Ribosomal L29 family protein
					XHP01558_1	Ribosomal protein L10 family protein
					XHP00753_1	Ribosomal protein L13 family protein

Table 4.5. (continued)

GO ID	GO description	GO category	Corrected p-value	No of contig	Contig ID	Contig description
GO:0022626 (continued)	cytosolic ribosome	CC	1.04E-05	11	XHP00348_1	Ribosomal protein L14p/L23e family protein
					XHP00255_1	Ribosomal protein S10p/S20e family protein
					XHP01132_1	Ribosomal protein S30 family protein
					XHP00771_1	Ribosomal protein S5 family protein
GO:0033279	ribosomal subunit	CC	2.11E-02	8	XHP01657_1	60S ribosomal protein L13A (RPL13aD)
					XHP00480_1	Ribosomal L29 family protein
					XHP00753_1	Ribosomal protein L13 family protein
					XHP00348_1	Ribosomal protein L14p/L23e family protein
					XHP00954_1	Ribosomal protein L22p/L17e family protein
					XHP00255_1	Ribosomal protein S10p/S20e family protein
					XHP01132_1	Ribosomal protein S30 family protein
					XHP00771_1	Ribosomal protein S5 family protein
GO:0003735	structural constituent of ribosome	MF	3.28E-10	26	XHP01934_1	60S acidic ribosomal protein family
					XHP01429_1	60S acidic ribosomal protein family
					XHP01489_1	60S acidic ribosomal protein P0
					XHP02016_1_M	60S acidic ribosomal protein P3 (RPP3B)
					XHP01657_1	60S ribosomal protein L13A (RPL13aD)
					XH_LDR_05F021	plastid ribosomal protein L35
					XHP00480_1	Ribosomal L29 family protein
					XHP01565_1	ribosomal protein L10 B
					XHP01558_1	Ribosomal protein L10 family protein
					XHP00958_1	Ribosomal protein L11 family protein
					XHP00753_1	Ribosomal protein L13 family protein
					XHP00348_1	Ribosomal protein L14p/L23e family protein
					XHP01769_1	ribosomal protein L18
					XHP00954_1	Ribosomal protein L22p/L17e family protein
					XH_RRF_01H088	Ribosomal protein L31e family protein
					XHP00311_1	Ribosomal protein L32e
					XHP00274_2	Ribosomal protein L6 family
					XHP01603_1	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein
					XHP00255_1	Ribosomal protein S10p/S20e family protein
					XHP01315_1	Ribosomal protein S26e family protein
					XHP01132_1	Ribosomal protein S30 family protein
					XHP00777_1	Ribosomal protein S3Ae
					XHP00771_1	Ribosomal protein S5 family protein
					XHP01317_1	Ribosomal protein S6b
					XHP00781_1	Ribosomal protein S8e family protein
					XHP01113_1	R-protein L3 B

BP: biological process; CC: cellular component; MF: molecular function.

Cluster 5, which represents the next phase of up-regulated contigs, was enriched for components related to membrane or vesicular fractions such as the microsome (Fig. 4.6). No over-represented biological process or molecular function terms were identified. Membrane fraction may possibly reflect the breakdown of the thylakoid membrane (Ingle *et al.*, 2008), or the tonoplast membrane of large vacuole during desiccation, or protein trafficking via vesicles.

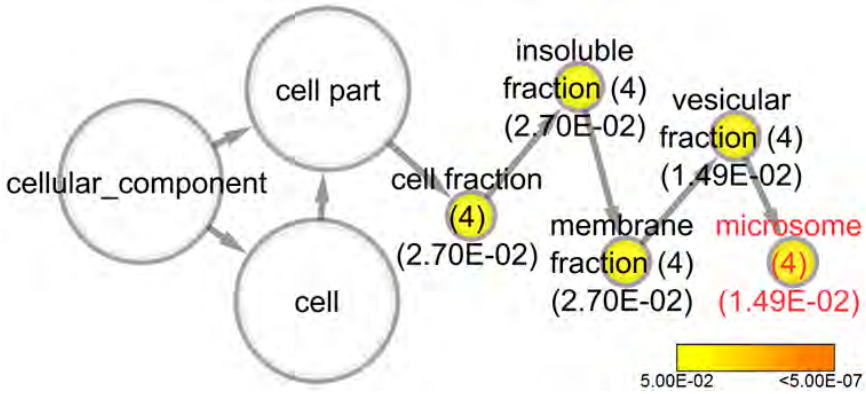


Figure 4.6. Over-representation of cellular component terms found in the set of 204 cluster 5 differentially expressed contigs identified in *X. humilis* leaves during desiccation. The enrichment of ontology terms was analyzed using hypergeometric test with a significance level of p-value= 0.05. The node size is proportional to the number of contigs in the test set being annotated to that node, and the significantly over-represented terms are presented as coloured nodes with a colour scale ranging from light yellow (p-value= 0.05) to dark orange (p-value< 5.00E-07). The two numbers shown in brackets under each significantly over-represented term indicate the number of contigs being annotated, and the p-value after FDR correction to that node respectively. The contigs annotated to the most specific terms (in red text) are summarized in Table 4.6.

Contigs associated with the most specific over-represented GO term of microsome included XHP01796_1 and XHP00076_1 (Table 4.6). XHP01796_1 encodes for *A. thaliana* seed gene 1 (ATS1), which has been reported to be an embryo-specific gene in *A. thaliana* (Nuccio and Thomas, 1999). ATS1 protein was classified as class 1 caleosin that contains calcium-binding domain. XHP00076_1 encodes for responsive to desiccation 20 protein (RD20), whose expression has been reported to be induced upon salt or drought stresses, as well as ABA treatment in *A. thaliana* (Takeshi *et al.*, 2000). RD20 protein was classified as class 3 caleosin. Calcium-binding caleosin proteins have been reported to be mainly bound to microsomal membrane fractions with low level of expression during early stages of seed development. As seeds mature, their overall level increases dramatically (Murphy *et al.*, 2000). Caleosins have an oleosin-like association with oil bodies formed and accumulated

during seed development. Together with oleosins, steroleosins and a mono layer of phospholipid membrane, they cover the entire surface of spherical triacylglycol (TAG) body to compress it, and to prevent the coalescence or aggregation with other oil bodies (Tzen and Huang, 1992). Caleosins, oleosins and steroleosins have been proposed as a product of endoplasmic reticulum (ER). They are inserted into or synthesized in ER coinciding with TAG synthesis, then they extend and bud off to give rise to oil bodies (Chen and Tzen, 2001). Although oil bodies have mainly been reported to accumulate in maturing seeds as stored energy sources for germination and subsequent growth of seedlings, their presence have also been shown in the leaf mesophyll cells of many angiosperm species (Lersten *et al.*, 2006). Oil bodies of mesophyll cells were speculated to act as intermediate storage products of photosynthesis (Lersten *et al.*, 2006). Up-regulation of the 2 caleosin contigs observed in *X. humilis* may suggest the accumulation of the oil bodies in the leaves during desiccation, which may serve as one of the stored energy sources needed for the restoration of desiccated leaves upon rehydration, or the stored intermediate products readily to be utilized to resume photosynthesis in the rehydrating leaves of *X. humilis*.

Table 4.6. Mapping of contig IDs to the most specific over-represented GO terms that are significantly enriched in Cluster 5 identified in *X. humilis* leaves during desiccation.

GO ID	GO description	GO category	Corrected p-value	No of contig	Contig ID	Contig description
GO:0005792	microsome	CC	1.49E-02	4	XHP01796_1	ATS1 (ARABIDOPSIS THALIANA SEED GENE 1); calcium ion binding
					XHP01507_1	cytochrome b5
					XHP00076_1	RESPONSIVE TO DESSICATION 20
					XHP01731_1	vacuolar protein sorting 26B

CC: cellular component.

Another contig associated with the over-represented terms is XHP01731_1, which encodes a vacuolar protein sorting (VPS) protein (Table 4.6). During maturation, plant seeds accumulate large quantities of storage proteins such as globulins and albumins. The precursors of the storage proteins are synthesized on rough ER and are sorted to protein storage vacuoles by vesicle-mediated machinery including ER, the Golgi apparatus, and the endosomes/prevacuoles, where they are converted into the mature form (Shimada *et al.*, 2006; Xiang *et al.*, 2013). VPS proteins play important roles in the delivery of vacuolar proteins synthesized on rough ER to vacuoles. In *A. thaliana*, deficiency in VPS29 was reported to

have resulted in abnormal accumulation of precursors of the storage proteins in dry seed, and had negative impact on the plant growth and development (Shimada *et al.*, 2006).

Functional enrichment analysis of the clusters 6 and 5 early desiccation up-regulated contigs suggested that during early phase of desiccation, synthesis of proteins required for protein translation, protection, as well as proteins similar to seed storage proteins or oil bodies were predominantly activated in *X. humilis*. GO terms associated with transcriptional regulation were not found among these early up-regulated contigs, which may be suggesting that the genes conferring desiccation tolerance in *X. humilis* may have predominantly be activated via de-repression (as resulted from the down-regulation of transcription repressors found in clusters 2 and 3) than being activated by newly up-regulated transcription activators.

4.3.5 Functional enrichment of the late up-regulated clusters identified in X. humilis leaves during desiccation

Levels of mRNA transcripts in Cluster 1 and 4 accumulated more slowly during desiccation, and their levels remained high in desiccated leaves. Embryonic development was the only over-represented term identified among the cluster 4 contigs (BiNGO enrichment graph not shown). Most of the contigs annotated to this GO term were LEA, or seed maturation proteins (Table 4.7). These late desiccation responsive LEA transcripts may be translated towards the end of desiccation and play important roles in the maintenance and protection of cellular integrity and macromolecule under severe water deficit. Or alternatively, they may be stably stored and translated upon rehydration to protect the cellular content from any possible damages caused by the reintroduction of water into the system.

In addition to the LEA proteins, XHP00651_1, an orthologue of *A. thaliana* endosomal sorting complexes required for transport (ESCRT) related charged multivesicular body protein/chromatin modifying protein (CHMP) was also found to be associated with the enriched term of “embryonic development” (Table 4.7). CHMP proteins are involved in the sorting and degradation of plasma membrane proteins via endosomal system, and have been shown to be essential for embryo and seedling development (Spitzer *et al.*, 2009). This may imply the correct sorting and degradation of membrane proteins may play vital roles in the survival or the recovery of plant during desiccation, or during rehydration.

Table 4.7. Mapping of contig IDs to the most specific over-represented GO terms that are significantly enriched in Cluster 4 identified in *X. humilis* leaves during desiccation.

GO ID	GO description	GO category	Corrected p-value	No of contig	Contig ID	Contig description
GO:0009790	embryonic development	BP	2.82E-02	16	XHP01047_1	cytokinin response factor 2
					XHP00665_1	LEA 18
					XHP00024_1	LEA 7
					XHP00641_1	LEA domain-containing protein
					XHP01523_1	LEA group 1 domain-containing protein
					XH_LDF_50H122	LEA protein, putative
					XHP00272_1	LEA protein, putative
					XHP00762_1	LEA protein, putative
					XHP01693_1	LEA protein, putative
					XHP00052_3	LEA14
					XHP01163_1	maternal effect embryo arrest 14
					XHP00851_1	MATERNAL EFFECT EMBRYO ARREST 31
					XHP01676_1	proton-dependent oligopeptide transport (POT) family protein
					XHP01667_1	RADICAL-INDUCED CELL DEATH1
					XHP00664_2	seed maturation protein
					XHP00651_1	SNF7 family protein

BP: biological process.

For the last group of contigs up-regulated in response to desiccation, cluster 1 was enriched with biological process terms related to biosynthesis of chlorophyll (Fig. 4.7). It has been reported that during rehydration process in *X. humilis* desiccated leaves, the partial recovery of chlorophyll biosynthesis and the electron transport system of PSII were possible without *de novo* transcription, and it was hypothesized that the mRNA transcripts responsible must have been stably stored in the desiccated state (Dace *et al.*, 1998). This was later validated by Collett *et al.* (2003), in which a significant mRNA abundance of two PSII genes, psbA and psbP, was detected in the desiccated *X. humilis* leaves.

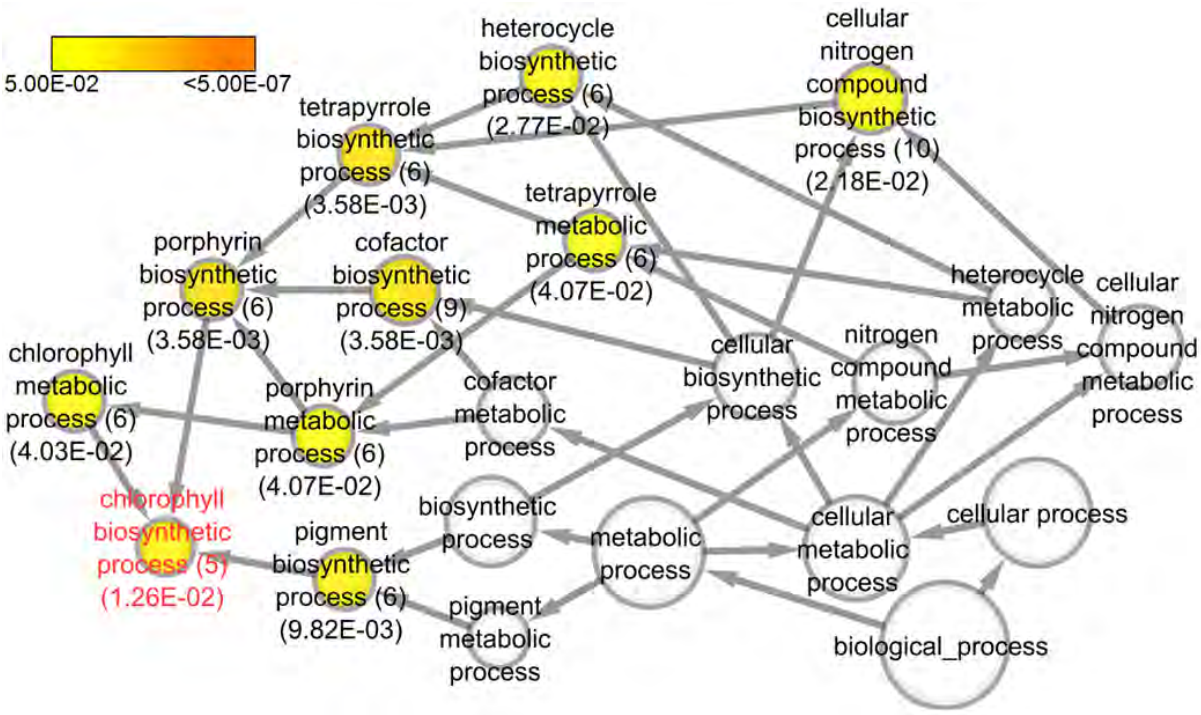


Figure 4.7. Over-representation of biological process ontology terms found in the set of 109 cluster 1 differentially expressed contigs identified in *X. humilis* leaves during desiccation. The enrichment of ontology terms was analyzed using hypergeometric test with a significance level of p-value= 0.05. The node size is proportional to the number of contigs in the test set being annotated to that node, and the significantly over-represented terms are presented as coloured nodes with a colour scale ranging from light yellow (p-value= 0.05) to dark orange (p-value< 5.00E-07). The two numbers shown in brackets under each significantly over-represented term indicate the number of contigs being annotated, and the p-value after FDR correction to that node respectively. The contigs annotated to the most specific terms (in red text) are summarized in Table 4.8.

Contigs annotated to the most specific GO term of "chlorophyll biosynthetic process" included genomes uncoupled 4 protein (GUN4) and protochlorophyllide oxidoreductase A (PORA) (Table 4.8). During photomorphogenesis, POR catalyzes the light-dependent reduction of protochlorophyllides A to chlorophyllide a, which is subsequently converted to chlorophyll. Among the 3 POR proteins identified in *A. thaliana*, PORA has been reported to be essential in photomorphogenesis and for normal plant growth and development (Paddock *et al.*, 2012). In *A. thaliana*, more than 3500 nuclear genes are predicted to encode chloroplast proteins. The developmental and metabolic status of chloroplast itself affects the expression of nuclear genes that encode chloroplast proteins. Distinct plastid-to-nucleus signals are essential for the proper expression of many nuclear photosynthetic genes. GUN4 has been reported to participate in the coupling of the expression of several nuclear genes to the functional state of the chloroplast, and is required for chlorophyll synthesis and accumulation under normal growth conditions (Larkin *et al.*, 2003; Adhikari *et al.*, 2011).

Table 4.8. Mapping of contig IDs to the most specific over-represented GO terms that are significantly enriched in Cluster 1 identified in *X. humilis* leaves during desiccation.

GO ID	GO description	GO category	Corrected p-value	No of contig	Contig ID	Contig description
GO:0015995	chlorophyll biosynthetic process	BP	1.26E-02	5	XH_RRF_01A108	Aldolase superfamily protein
					XHP00706_2	COPPER RESPONSE DEFECT 1
					XHP00378_1	GENOMES UNCOUPLED 4
					XHP00249_2	PORA; oxidoreductase/ protochlorophyllide reductase
					XHP00827_1	PORA; oxidoreductase/ protochlorophyllide reductase
GO:0043661	peribacteroid membrane	CC_0.1	8.11E-02	2	XHP01129_1	DIGALACTOSYL DIACYLGLYCEROL DEFICIENT 1
					XHP00464_1	digalactosyldiacylglycerol synthase 1

BP: biological process.

Clusters 1 and 4 late desiccation responsive contigs may represent transcripts that are responsible for cellular maintenance and protection under desiccated state, or transcripts essential for cellular restoration that are stored and immediately translated in absence of *de novo* transcription when water becomes available. Functional enrichment analysis of these contigs suggested that restoration in processes related to chlorophyll biosynthesis and photosynthesis may be an immediate primary focus when water becomes available again in *X. humilis* leaves. The resumption of processes related to cellular development may be accompanied by protection from the possibly seed specific LEA proteins.

4.3.6 Validation of X. humilis microarray expression data by quantitative real-time PCR analysis

The expression data of the 22 differentially expressed *X. humilis* contigs from the 7 PAMSAM clusters, and 2 non-differentially expressed contigs (as references used in the normalization process) were validated by quantitative real-time PCR analysis (RT-qPCR). When the RT-qPCR analysis was initiated, the sequence information of all the cDNA clones was not yet available, these 24 contigs were selected among the 424 clones that had been previously sequenced by Collett *et al*. (2004). The RT-qPCR reaction efficiencies for the 24 contigs analyzed were mostly within the recommended range of 0.8 to 1.2 (Appendix Table A.4.4). The efficiency of 5 reactions fell into the lower range of 0.6 to 1.2, which is still acceptable for experiments in which the same standard samples for standard curve generation are included in each experiment, and if only the patterns of gene expression across different conditions rather than the fold change, or relative or absolute expression levels are being determined (personal communication, Technical Assistance, Celtic Molecular Diagnostic (Cape Town, South Africa)).

Figure 4.8 summarized the comparison between expression data identified from RT-qPCR (normalized against the expression of non-differentially expressed Xh_RRF_02C098) and microarray for the 22 contigs analyzed. Similar figures showing comparison between microarray data and RT-qPCR data normalized against the expression of non-differentially expressed XHP00087_2, as well as against cDNA template input, are given in the Appendix (Fig. A.4.5; Fig. A.4.6). The profiles of the expression patterns plotted on graphs for the microarray and RT-qPCR data for 21/ 22 contigs, were highly similar.

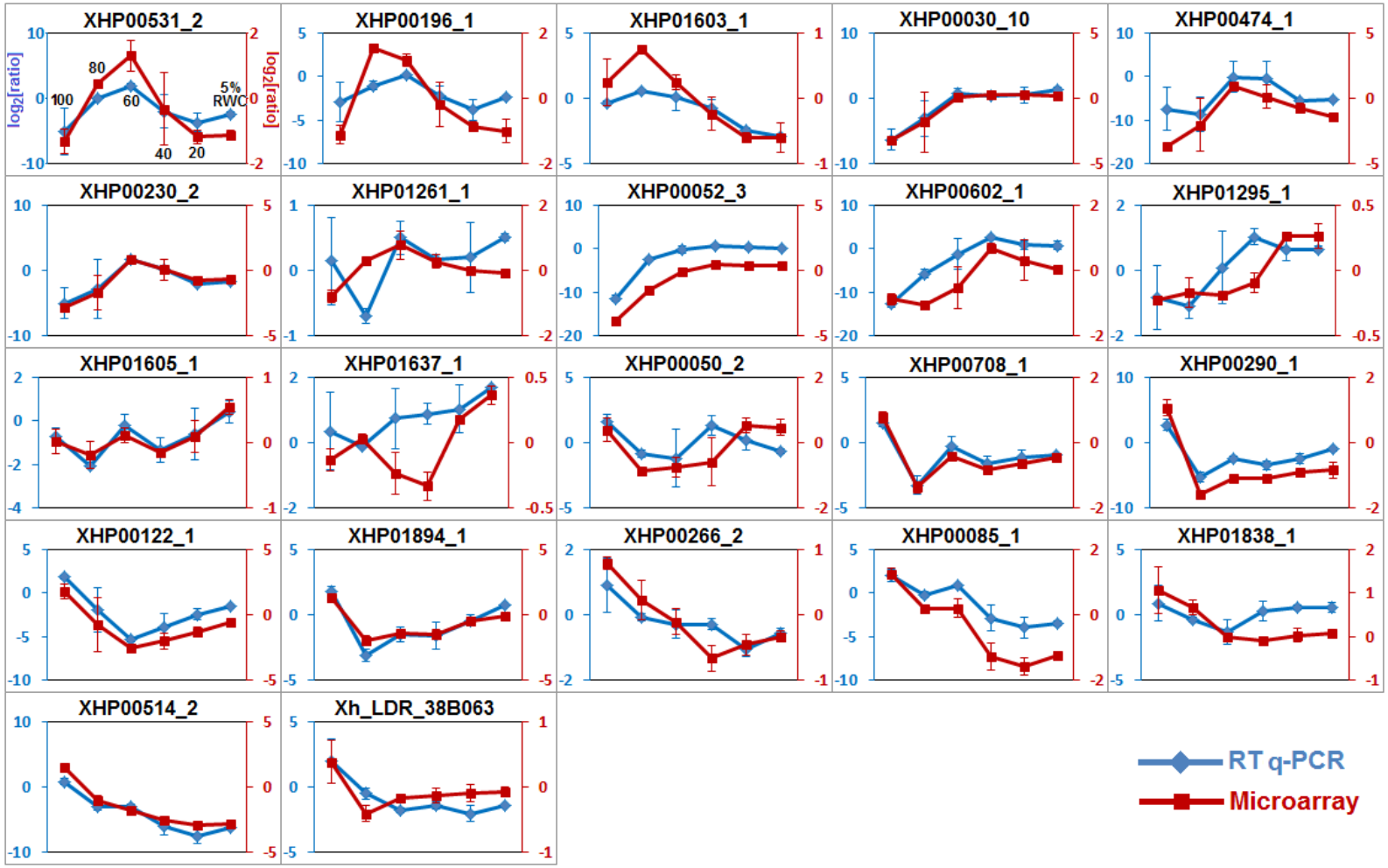


Figure 4.8. Validation of *X. humilis* microarray expression data by RT-qPCR analysis. Transitional changes in mRNA transcript abundance of 22 contigs in *X. humilis* leaves (two biological replicates) at 6 different stages of water loss during desiccation assessed by microarray analysis was validated by RT-qPCR analysis normalized against expression of Xh_RRF_02C098. Microarray expression values were represented as the averaged $\log_2$ [Cy3/Cy5] from the two biological replicates of leaf samples analyzed (red y-axis), and RT-qPCR expression values were represented as averaged $\log_2$ [ratios of contig concentration over that of Xh_RRF_02C098] (blue y-axis). Error bars represented the expression values identified in the two biological replicates of the leaf samples. Contig descriptions can be found in Appendix (Table A.4.4).

The Spearman's rank correlation coefficient test was used to statistically assess the overall correlation between the microarray data and RT-qPCR data for the 22 contigs. The test revealed that the overall correlation between expression data of the analyzed 22 contigs identified by microarray and RT-qPCR analysis normalized against (1) cDNA input amount; (2) the expression of XHP00087_2; and (3) the expression of Xh_RRF_02C098, yielded correlation coefficients of 0.85, 0.80 and 0.84 respectively (Table 4.9). It was reported that the correlation between microarray and RT-qPCR data found in the literatures ranged from -0.48 to 0.94, and a significant correlation of 0.8 was generally considered a good positive correlation (Morey *et al.*, 2006). The good agreement between the RT-qPCR and microarray expression values for the selected 22 contigs, validates the expression data captured for the 1680 contigs in *X. humilis* leaves during desiccation treatment.

Table 4.9. Summary of overall significant correlations between expression data identified by microarray and by RT-qPCR analysis with different RT-qPCR normalization approaches calculated using the online Spearman's Rank Correlation-Free Statistics and Forecasting Software (Calculator).

RT-qPCR normalization approach	Total no. of genes analyzed	Total no. of data points	Spearman's rank correlation	1-sided p-value
Norm_cDNA	22	132	0.85	<1.00E-06
Norm_NDE1 (XHP00087_2)	22	132	0.80	<1.00E-06
Norm_NDE2 (Xh_RRF_02C098)	22	132	0.84	<1.00E-06

Norm_cDNA, Norm_NDE1 and Norm_NDE2 denoted the normalization strategies using cDNA template input, expressions of XHP0087_2 and expressions of Xh_RRF_02C098 respectively.

4.3.7 Clustering of X. humilis LEA contigs in leaves during desiccation

To further characterize the desiccation responses in *X. humilis* leaves after microarray data validation, expression data of different classes of genes was examined. To focus on whether particular classes of LEAs are co-regulated during desiccation in *X. humilis* leaves, the expression values of 46 differentially expressed *X. humilis* contigs annotated as LEAs were independently clustered using PAMSAM. Mclust failed to predict an optimal division (k value) for the LEA contigs, which may have resulted from the small total number of expression data being tested. Different k values were tested in PAMSAM clustering, until a Sammon map with relatively well separated clusters was observed. The 46 LEA contigs were clustered into 3 clusters. The majority of contigs (96%) fell into the up-regulated clusters 1 and 2, which overlapped closely in SAMMON mapping space (Fig.4.9). The medoid of cluster 3 (comprising only 2 LEAs) showed little change in expression during desiccation in leaves. When the expressions of the two member contigs were examined, both contigs

appeared to be transiently down-regulated during desiccation (Fig. 4.9 inserts). Twenty three LEA contigs were grouped into cluster 2, and the medoid of this cluster showed an early desiccation response with maximum transcript abundance at 60% RWC. Although mRNA transcript levels subsequently decrease, they remained significantly higher than the hydrated state. The medoid for cluster 1 (comprising 21 LEA contigs) showed a late desiccation response, with maximum transcript abundance observed in leaves at 60% RWC.

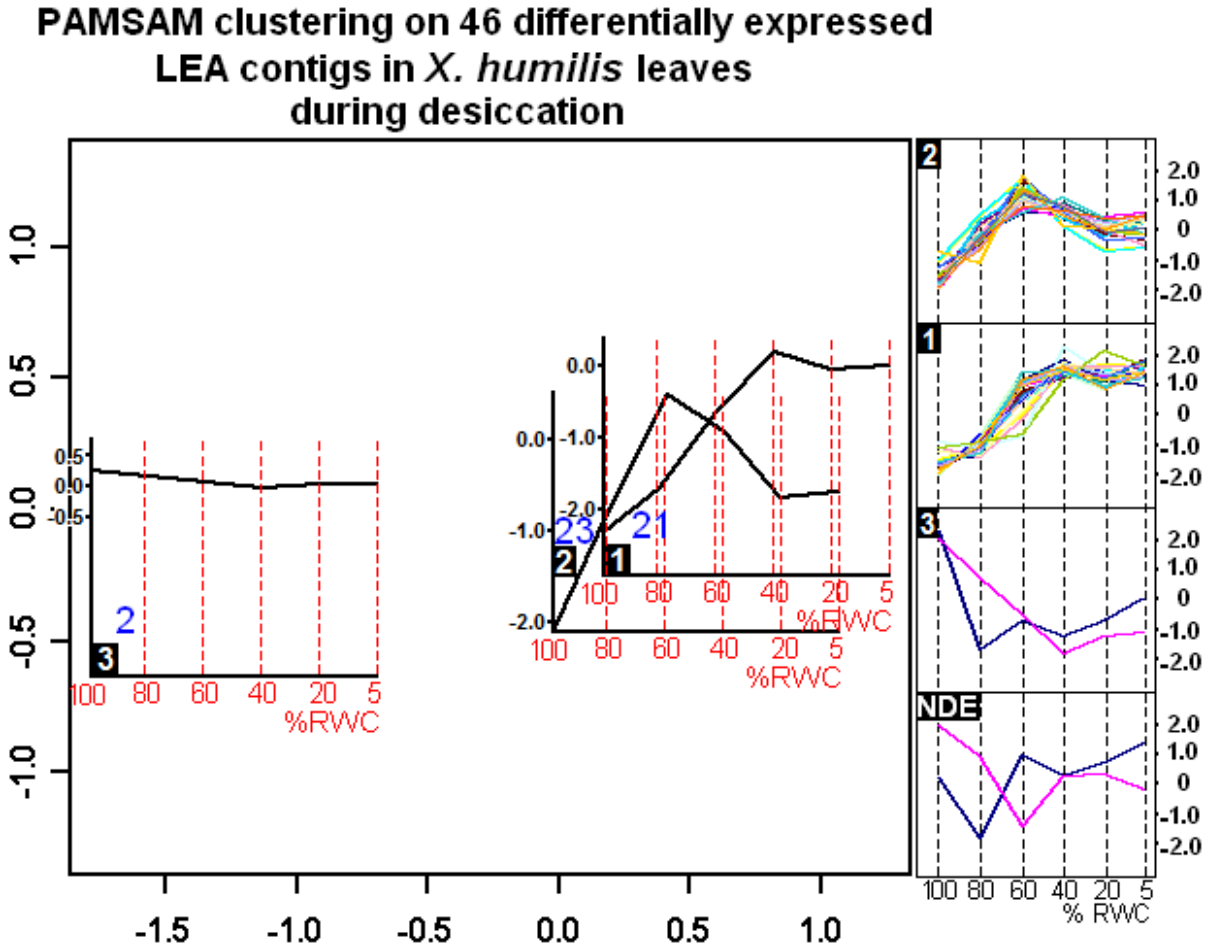


Figure 4.9. PAMSAM clustering of *X. humilis* LEAs. Expressions of 46 differentially expressed LEA contigs identified in *X. humilis* leaves during desiccation were clustered into 3 clusters by PAMSAM algorithm. The medoids/clusters were partitioned and plotted in a two-dimensional Sammon map, indicating their approximate similarity to each other. Each medoid/cluster was represented by a small graph that showed the expression values in $\log_2$ ratios (y-axis of cluster graph) across the 6 different stages of water loss: 100%, 80%, 60%, 40%, 20% and 5%RWC. The x- and y-axis of the Sammon map represent distances between the medoids in the two different data dimensions. Total number of contigs in each cluster was indicated by the number above the cluster number. The expression patterns of individual member contigs in each of the 3 clusters were shown in the inserts. The insert y-axis represents the $\log_2$ ratios after pattern standardization performed in Gene Expression Profile Analysis Suite (GEPAS), Babelomics version 3.2 (http://babelomics3.bioinfo.cipf.es). NDE: non-differentially expressed.

The 23 LEAs in cluster 2 that showed early up-regulation response during may be responsible in the protection and of cellular contents during early phase of desiccation. Three of which, XHP00334_1 (class 3), XHP00492_1 (class 6) and XHP00105_1 (class 10), whose orthologues in *A. thaliana*, At1g72100, At3g22490 and At1g04560 respectively, have been previously reported as seed specific LEAs that were only differentially regulated during seed development in *A. thaliana* (Illing *et al.*, 2005; Rodrigues *et al.*, 2010) (Table 4.10).

Maximum transcript abundance of the 25 cluster 1 contigs found in the desiccated leaves may play an important protective role during the desiccated state, or required for the rehydration phase. Class 1 LEAs are well-characterized seed-specific LEAs in *A. thaliana* (Manfre *et al.*, 2006; Tunnacliffe and Wise, 2007), and a class 1 LEA contig, XHP00221_2, fell into this cluster. In addition, a class 4 LEA contig, XHP00665_1, whose *A. thaliana* orthologue (At2g35300) has also been shown to be seed specific in *A. thaliana* (Illing *et al.*, 2005).

There was no report found in the literature regarding to the down-regulation of LEA proteins during abiotic stress responses or seed development in plants. The two down-regulated LEA contigs in cluster 3 may be involved in processes that are specifically during hydrated state in *X. humilis*, such as photosynthesis or general cellular development. When these processes are turned down in response to desiccation, the two LEAs are down-regulated in correspondence.

The expression of LEA genes during desiccation in *X. humilis* leaves did not appear to be class-dependent, because multiple classes were found in each cluster, and no class was uniquely represented by a PAMSAM cluster (Table 4.10). The distribution was unchanged even if 9 clusters were selected for the analysis (data not shown). This agreed with the experimental evidence reported by Olvera-Carrillo *et al*. (2010) that LEA genes from the same class may be differentially regulated, and a functional diversification may have occurred in the course of evolution.

Table 4.10. PAMSAM clustering of the differentially expressed *X. humilis* LEA contigs.

PAMSAM cluster	LEA class	No of contig	Contig ID	Contig description
2	2	6	XHP00030_10	RAB18 (RESPONSIVE TO ABA 18)
			XHP00106_1	RAB18 (RESPONSIVE TO ABA 18)
			XHP00531_2	dhn9
			XHP00658_1	dehydrin
			XHP00966_1	RAB18 (RESPONSIVE TO ABA 18)
			XHP01481_1	RAB18 (RESPONSIVE TO ABA 18)
	3	11	XHP00011_1	late embryogenesis
			XHP00012_1	LEA domain-containing protein
			XHP00014_2	LEA protein, putative
			XHP00026_7	group 3 LEA protein
			XHP00055_2	LEA domain-containing protein
			XHP00089_3	group 3 LEA protein
			XHP00200_4	LEA domain-containing protein
			XHP00334_1	LEA domain-containing protein
			XHP00406_2	cre-lea-1 protein
			XHP01120_1	LEA family protein
			XHP01261_1	LEA domain-containing protein
	4	1	XHP00039_2	seed maturation protein
	6	1	XHP00492_1	Seed maturation protein
	7	1	XHP00330_1	LEA protein
	8	1	XHP01697_1	LEA 14
	10	2	XHP00054_3	AWPM-19-like membrane family protein
			XHP00105_1	AWPM-19-like family protein
1	1	1	XHP00221_2	embryonic abundant protein 1
	2	4	XHP00016_2	RAB18 (RESPONSIVE TO ABA 18)
			XHP00109_4	RAB18 (RESPONSIVE TO ABA 18)
			XHP00182_6	dehydrin
			XHP00615_1	dehydrin-like protein dh2
	3	5	Xh_LDF_50H122	LEA protein, putative
			XHP00024_1	LEA 7
			XHP00062_6	late embryogenesis
			XHP00487_2	late embryogenesis
			XHP00641_1	LEA domain-containing protein
	4	4	XHP00028_2	LEA 4-5
			XHP00664_2	seed maturation protein
			XHP00665_1	LEA 18
			XHP01523_1	LEA group 1 domain-containing protein
	6	4	XHP00272_1	LEA protein, putative
			XHP00762_1	LEA protein, putative
			XHP01693_1	LEA protein, putative
			XHP01956_1_M	LEA protein, putative
	7	1	XHP00066_1	drought-induced 21
	8	1	XHP00052_3	LEA14
	10	1	XHP01703_1	AWPM-19-like membrane family protein
3	7	1	XHP01037_1	ATDI21 (ARABIDOPSIS THALIANA DROUGHT-INDUCED 21)
	8	1	XHP01570_1	LEA protein
NDE	3	1	XHP00234_1	LEA protein, putative
	6	1	XHP00623_1	LEA protein, putative

NDE: non-differentially expressed.

4.3.8 Clustering of X. humilis antioxidant contigs in leaves during desiccation

The expression values of the 19 differentially expressed contigs annotated as antioxidants were also independently clustered using PAMSAM to investigate whether particular classes of antioxidants were co-regulated during desiccation. With the failure in the optimal prediction of the k value, based on trial and error approach as discussed before, the expressions of 19 antioxidant contigs were clustered into 5 groups (Fig. 4.10). The number of antioxidant contigs were more evenly divided between the 5 clusters compared to the LEA contig dataset, and with 68% of the contigs fell into the up-regulated clusters 1, 2 and 5. The medoids of clusters 5 and 1 both showed an early up-regulation response to desiccation with maximum transcript abundance observed at 60% RWC. After which, the mRNA transcript levels of the 3 contigs in cluster 5 resumed to the level observed in the hydrated state, whereas the transcript levels of the 4 contigs in cluster 1 remained significantly higher than the hydrated state. Cluster 2 medoid (comprising 6 contigs) showed a late up-regulated response with maximum transcript abundance observed in the desiccated leaves.

The 13 antioxidants in clusters 1, 2 and 5 that showed early or late up-regulation may play important protective role against oxidative damages during the early phase of desiccation, or during the desiccated state, as well as during the rehydration phase respectively. The 6 down-regulated antioxidants from clusters 3 and 4 may have roles in minimizing the oxidative damages specifically arose from the normal metabolic and photosynthetic reactions when water is available. Once metabolism and photosynthesis have been shut down in response to water deficit, these antioxidants may also be repressed accordingly.

In contrast to the 3 up-regulated clusters, cluster 4 medoid showed an early down-regulation response to desiccation with minimum transcript abundance observed at 60% RWC, which subsequently increased and remained at a level significantly lower than the hydrated state. Cluster 3 medoid showed a late desiccation down-regulation response with minimum transcript abundance observed in the desiccated leaves.

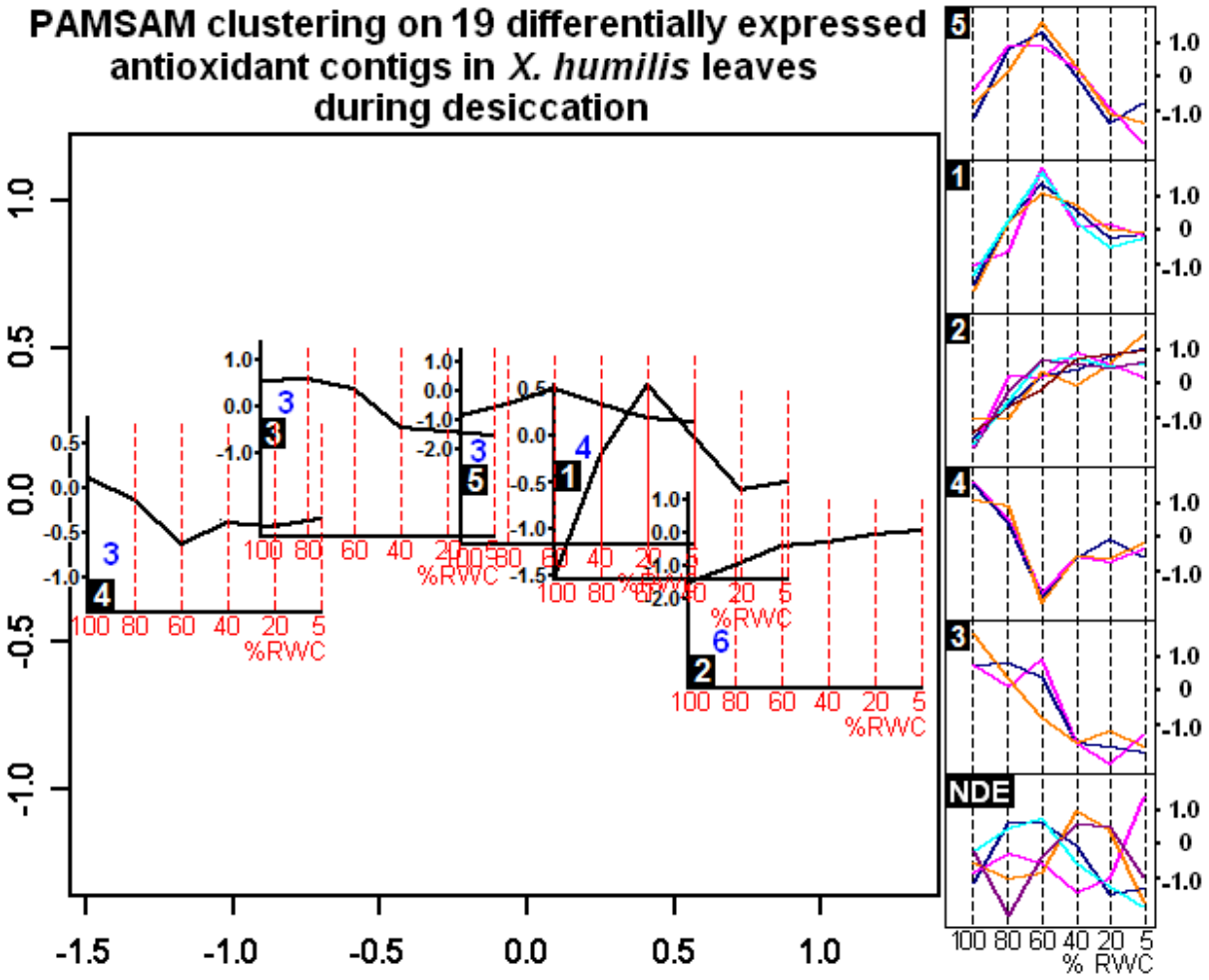


Figure 4.10. PAMSAM clustering of *X. humilis* antioxidants. Expressions of 19 differentially expressed antioxidant contigs identified in *X. humilis* leaves during desiccation were clustered into 5 clusters by PAMSAM algorithm. The medoids/clusters were partitioned and plotted in a two-dimensional Sammon map, indicating their approximate similarity to each other. Each medoid/cluster was represented by a small graph that showed the expression values in $\log_2$ ratios (y-axis of cluster graph) across the 6 different stages of water loss: 100%, 80%, 60%, 40%, 20% and 5%RWC. The x- and y-axis of the Sammon map represent distances between the medoids in the two different data dimensions. Total number of contigs in each cluster was indicated by the number above the cluster number. The expression patterns of individual member contigs in each of the 5 clusters were shown in the inserts. The insert y-axis represents the $\log_2$ ratios after pattern standardization performed in Gene Expression Profile Analysis Suite (GEPAS), Babelomics version 3.2 (http://babelomics3.bioinfo.cipf.es). NDE: non-differentially expressed.

1-cysteine peroxiredoxin 1 has been reported as a seed specific antioxidant, and is highly expressed during late seed development (Aalen, 1999; Haslekås *et al.*, 1998; 2003). XHP00084_2, a cluster 1 contig encodes for 1-cysteine peroxiredoxin 1, and was shown to be up-regulated in the leaves of *X. humilis* during desiccation (Table 4.11). XHP01001_1 from cluster 3 encodes for another cysteine peroxiredoxin protein, 2-cystine peroxiredoxin A. In contrast to the up-regulated seed specific 1-cysteine peroxiredoxin 1, 2-cystine peroxiredoxin

A was found down-regulated in *X. humilis* leaves during desiccation. It has been reported that the plastidic 2-cystine peroxiredoxin is involved in the protection of the photosynthetic apparatus against oxidative damage (Broin *et al.*, 2002). The down-regulation of XHP01001_1 may be correlated to the reduction in photosynthetic apparatus or activity in *X. humilis* upon desiccation.

The expression of antioxidant genes during desiccation in *X. humilis* leaves did not appear to be class-dependent, because multiple classes were found in each cluster, and no class was uniquely represented by a PAMSAM cluster (Table 4.11), even if 9 clusters were selected for the analysis (data not shown). For example, contigs encoding peroxidases are both up-regulated (clusters 1, 2 and 5) and down-regulated (cluster 4) during desiccation.

Table 4.11. PAMSAM clustering of the differentially expressed *X. humilis* antioxidant contigs.

PAMSAM cluster	Antioxidant class	Contig ID	Contig description
5	peroxidase	XHP01781_1	thioredoxin-dependent peroxidase 1
	thioredoxin	XHP01155_1	thioredoxin H-type 1
	unknown	XHP01786_1	DC1 domain-containing protein
1	glutathione transferase	XHP01640_1	glutathione transferase
	glutathione transferase	XHP00591_1	microsomal glutathione s-transferase, putative
	peroxidase	XHP00394_1	ascorbate peroxidase 2
	peroxiredoxin	XHP00084_2	1-cysteine peroxiredoxin 1
2	peroxidase	XHP00994_1	glutathione peroxidase 1
	peroxidase	XHP01166_1	glutathione peroxidase 3
	peroxidase	XHP00604_3	glutathione peroxidase 6
	peroxiredoxin	XHP01983_1_M	peroxiredoxin type 2, putative
	thioredoxin	XHP00107_1	Thioredoxin superfamily protein
	unknown	XHP01545_1	G6PD6 (GLUCOSE-6-PHOSPHATE DEHYDROGENASE 6)
4	CAT	XHP01613_1	catalase
	peroxidase	XHP01774_1	peroxidase 52
	peroxidase	XHP00810_1	Peroxidase superfamily protein
3	CAT	XHP02074_1_M	catalase
	peroxiredoxin	XHP01001_1	2-Cys peroxiredoxin A
	SOD	XHP00362_1	copper/zinc superoxide dismutase 1
NDE	glutathione transferase	XHP00257_1	glutathione S-transferase TAU 19
	unknown	XHP00471_1	MSRB2 (methionine sulfoxide reductase B 2)
	unknown	XHP00763_1	NQR; binding / catalytic/ oxidoreductase/ zinc ion binding
	unknown	XHP02001_1_M	fructose-bisphosphate aldolase, putative
	peroxidase	XHP02029_1_M	ATGPX3 (GLUTATHIONE PEROXIDASE 3); glutathione peroxidase

CAT: catalase; SOD: superoxide dismutase; NDE: non-differentially expressed.

4.3.9 Clustering of X. humilis transcription factor contigs in leaves during desiccation

The expression values of the 78 differentially expressed contigs annotated as transcription factors were also independently clustered using PAMSAM to investigate if there are (1) any transcription repressors active in hydrated tissue being down-regulated during desiccation; (2) any transcription activators being up-regulated during early phase of desiccation; and (3) any transcription factors being up-regulated late, and are stored in desiccated tissue await for translation upon rehydration. These transcription factor contigs were clustered into 7 groups (Mclust failed, k value was determined by trial and error approach) (Fig. 4.11).

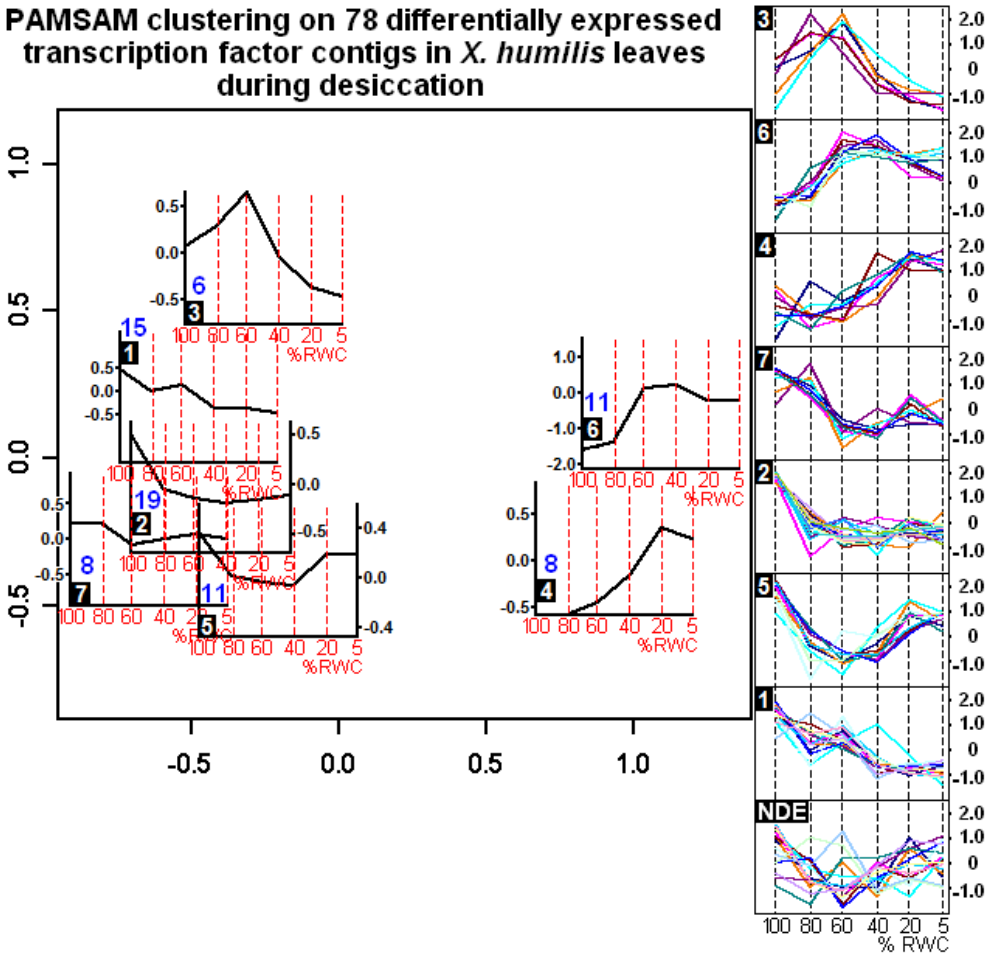


Figure 4.11. PAMSAM clustering of *X. humilis* transcription factors. Expressions of 78 differentially expressed transcription factor contigs identified in *X. humilis* leaves during desiccation were clustered into 7 clusters by PAMSAM algorithm. The medoids/clusters were partitioned and plotted in a two-dimensional Sammon map, indicating their approximate similarity to each other. Each medoid/cluster was represented by a small graph that showed the expression values in $\log_2$ ratios (y-axis of cluster graph) across the 6 different stages of water loss: 100%, 80%, 60%, 40%, 20% and 5%RWC. The x- and y-axis of the Sammon map represent distances between the medoids in the two different data dimensions. Total number of contigs in each cluster was indicated by the number above the cluster number. The expression patterns of individual member contigs in each of the 7 clusters were shown in the inserts. The insert y-axis represents the $\log_2$ ratios after pattern standardization performed in Gene Expression Profile Analysis Suite (GEPAS), Babelomics version 3.2 (http://babelomics3.bioinfo.cipf.es). NDE: non-differentially expressed.

Thirty two percent of the transcription factors were identified to be induced by desiccation, which fell into clusters 3, 4 and 6. Although cluster 3 contigs were clustered more closely with the down-regulated clusters of 1, 2, 5 and 7 on the Sammon map, and has an overall down-regulation trend, the 6 transcription factors in this cluster were shown to be transiently up-regulated, early-on during desiccation. Three of these contigs with maximum transcript abundance observed at 80% RWC, and the other 3 at 60% RWC, which then subsequently decreased to a level lower than in the hydrated state (Fig. 4.11 inserts).

Early up-regulated transcription factors found in cluster 3 included two heat shock transcription factors XHP01296_1 and XHP01128_1, a family transcription factors widely found induced upon heat and other abiotic stresses (Guo *et al.*, 2008) (Table 4.12). The up-regulation of XHP01228_1 (maximum abundance observed at 60% RWC) may be induced by XHP01296_1 (maximum abundance observed at 80% RWC), and involved in the downstream regulation of target genes when RWC reaches 60%.

Table 4.12. PAMSAM clustering of the differentially expressed *X. humilis* transcription factor contigs.

PAMSAM cluster	Contig ID	Contig description	Additional annotation information
3	Xh_LDR_51E052	NAC transcription factor	MF: DNA binding
	XHP00297_2	general regulatory factor 7	MF: protein domain specific binding
	XHP00975_1	NAC transcription factor	MF: DNA binding
	XHP01128_1	heat shock factor protein 1	MF: sequence-specific DNA binding transcription factor activity
	XHP01296_1	heat shock transcription factor C1	BP: response to stress
			MF: sequence-specific DNA binding transcription factor activity
	XHP01702_1	NRP2 (NAP1-RELATED PROTEIN 2); histone binding	BP: cell differentiation
			BP: nucleosome assembly
			MF: histone binding
			MF: chromatin binding
6	XHP00288_2	G-box binding factor 3	MF: protein dimerization activity
	XHP00402_1	NF-YA7	CC: CCAAT-binding factor complex
	XHP00496_2	anac083 transcription factor	BP: regulation of transcription, DNA-dependent
	XHP00949_1	NAD(P)-binding Rossmann-fold superfamily protein	BP: oxidation-reduction process
			CC: chloroplast stroma
			MF: coenzyme binding
	XHP01271_1	Histone-fold/TFIID-TAF/NF-Y	MF: sequence-specific DNA binding
	XHP01426_1	heat shock factor protein hsf30	BP: transcription, DNA-dependent
	XHP01538_1	transcription factor-related	MF: protein binding
	XHP01723_1	myb family transcription factor	MF: sequence-specific DNA binding transcription factor activity
	XHP01741_1	WRKY DNA-binding protein 51	MF: sequence-specific DNA binding transcription factor activity
	XHP01959_1_M	Oxidation-related Zinc Finger 1	MF: sequence-specific DNA binding transcription factor activity
	XHP02064_1_M	g-box binding factor	MF: protein dimerization activity
4	XHP00086_1	Cox19-like CHCH family protein	MF: chromatin binding
	XHP00630_1	WRKY33; transcription factor	BP: response to water deprivation
			BP: response to salt stress
			BP: response to cold
			BP: response to heat
	XHP00734_1	myb domain protein 73	MF: DNA binding
	XHP01047_1	cytokinin response factor 2	BP: root development
			BP: transcription factor import into nucleus
			BP: leaf development
			BP: cotyledon development
	XHP01231_1	zinc finger (C2H2 type) family protein	MF: sequence-specific DNA binding transcription factor activity
	XHP01783_1	BT2 (BTB AND TAZ DOMAIN PROTEIN 2	BP: response to gibberellin stimulus
			BP: response to carbohydrate stimulus
			BP: embryo sac development
			BP: circadian rhythm
			BP: response to auxin stimulus

Table 4.12. (continued)

PAMSAM cluster	Contig ID	Contig description	Additional annotation information
4 (continued)	XHP01783_1 (continued)	BT2 (BTB AND TAZ DOMAIN PROTEIN 2	BP: positive regulation of telomerase activity
			BP: pollen development
			BP: regulation of response to stress
			MF: histone acetyltransferase activity
	XHP01785_1	myb family transcription factor	BP: response to gibberellin stimulus
			BP: response to salt stress
			BP: response to abscisic acid stimulus
	XHP01827_1	CCR4-NOT transcription complex protein, putative	BP: response to wounding
			BP: response to biotic stimulus
			BP: RNA modification
			MF: poly(A)-specific ribonuclease activity
7	XHP00578_1	high mobility group B3	MF: chromatin binding
			MF: structural constituent of chromatin
	XHP00684_1	MBF1B (MULTIPROTEIN BRIDGING FACTOR 1B); DNA binding / transcription coactivator	BP: response to ethylene stimulus
			MF: transcription coactivator activity
	XHP01067_1	plastid transcriptionally active 6	BP: transcription from plastid promoter
			CC: chloroplast
	XHP01085_1	redox responsive transcription factor 1	MF: sequence-specific DNA binding transcription factor activity
	XHP01446_1	BolA-like family protein	CC: chloroplast
			MF: transcription regulator activity
	XHP01717_1	homeobox 1	MF: sequence-specific DNA binding transcription factor activity
	XHP01825_1	DNA-binding storekeeper protein-related transcriptional regulator	BP: regulation of transcription, DNA-dependent
	XHP01979_1_M	KIWI; DNA binding / protein binding / transcription coactivator	MF: transcription coactivator activity
2	Xh_LDR_10D023	cbf-like transcription factor	MF: sequence-specific DNA binding transcription factor activity
	Xh_LDR_50A043	indoleacetic acid-induced protein 10	MF: protein dimerization activity
	Xh_LRF_02G038	LIL3:1; transcription factor	BP: photosynthesis, light harvesting
	XHP00299_1	GRAS2, SCARECROW-like 14	BP: regulation of transcription, DNA-dependent
	XHP00496_1	ATNAC2 (ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 2); transcription factor	BP: leaf senescence
			BP: response to oxidative stress
			BP: response to salt stress
			MF: sequence-specific DNA binding transcription factor activity
			MF: protein heterodimerization activity
	XHP00580_1	BEL1-like homeodomain 7, BLH7	MF: sequence-specific DNA binding transcription factor activity
	XHP00887_2	wrky transcription factor 27	MF: DNA binding
	XHP01021_1	ethylene-responsive transcription factor 3	BP: transcription, DNA-dependent
	XHP01060_1	ap2 domain containing protein	MF: sequence-specific DNA binding transcription factor activity

Table 4.12. (continued)

PAMSAM cluster	Contig ID	Contig description	Additional annotation information
2 (continued)	XHP01087_1	serine acetyltransferase 2;1	BP: response to cold
			BP: cellular response to sulfate starvation
			BP: cysteine biosynthetic process from serine
			BP: response to cadmium ion
			MF: serine O-acetyltransferase activity
	XHP01164_1	ETHYLENE-INSENSITIVE3	MF: transcription regulator activity
	XHP01176_1	Basic helix-loop-helix (bHLH) DNA-binding family protein	BP: oxidation-reduction process
			MF: 2-alkenal reductase [NAD(P)] activity
	XHP01184_1	SZF1 (SALT-INDUCIBLE ZINC FINGER 1); transcription factor	MF: sequence-specific DNA binding transcription factor activity
	XHP01410_1	ethylene-insensitive 3f	MF: transcription regulator activity
	XHP01425_1	basic helix-loop-helix (bHLH) family protein	BP: response to abscisic acid stimulus
			BP: oxidation-reduction process
			MF: 2-alkenal reductase [NAD(P)] activity
	XHP02014_1_M	class iii homeodomain-leucine zipper protein c3hdz1	MF: sequence-specific DNA binding transcription factor activity
	XHP00447_2	ERF4 (ETHYLENE RESPONSIVE ELEMENT BINDING FACTOR 4); transcription repressor	BP: induced systemic resistance, jasmonic acid mediated signaling pathway
			BP: response to abscisic acid stimulus
			BP: negative regulation of ethylene mediated signaling pathway
			BP: negative regulation of transcription, DNA-dependent
			MF: transcription repressor activity
	XHP00930_1	WRKY DNA-binding protein 70	BP: induced systemic resistance, jasmonic acid mediated signaling pathway
			MF: transcription repressor activity
			MF: oxidoreductase activity
	XHP00948_1	zinc-finger protein 2	BP: hyperosmotic salinity response
			BP: response to water deprivation
			BP: embryo development ending in seed dormancy
			BP: response to abscisic acid stimulus
			MF: transcription repressor activity
5	XHP00217_1	myb-like transcription factor family protein	MF: DNA binding
	XHP00449_2	B-Box domian protein 24	BP: photomorphogenesis
			BP: response to salt stress
			BP: response to karrikin
	XHP00635_2	6B-interacting protein 1-like 2	MF: sequence-specific DNA binding transcription factor activity
	XHP00761_1	dna binding protein	MF: transcription regulator activity
	XHP00799_1	HAIRY MERISTEM 3	BP: regulation of transcription, DNA-dependent
	XHP01039_1	transcriptional regulator family protein	MF: transcription regulator activity
	XHP01046_1	YABBY2	BP: abaxial cell fate specification
			MF: sequence-specific DNA binding transcription factor activity

Table 4.12. (continued)

PAMSAM cluster	Contig ID	Contig description	Additional annotation information
5 (continued)	XHP01277_1	ZML1 (ZIM-LIKE 1)	MF: sequence-specific DNA binding transcription factor activity
	XHP01506_1	GRAS family transcription factor	BP: regulation of transcription, DNA-dependent
	XHP01675_1	ATBZIP53 (BASIC REGION/LEUCINE ZIPPER MOTIF 53); transcription factor	MF: sequence-specific DNA binding transcription factor activity
	XHP01131_1	TPL (TOPLESS); transcription repressor	BP: xylem and phloem pattern formation
			BP: response to auxin stimulus
			MF: transcription repressor activity
1	Xh_LDF_05G063	tubby like protein 3	MF: sequence-specific DNA binding transcription factor activity
	Xh_LDR_12F103	KNOTTED1-like homeobox gene 3	BP: positive regulation of transcription, DNA-dependent
	XHP00400_2	ERF domain protein 11	MF: sequence-specific DNA binding transcription factor activity
	XHP00877_1	EER4 (ENHANCED ETHYLENE RESPONSE 4); DNA binding / transcription initiation factor	MF: transcription initiation factor activity
	XHP01022_1	transcription repressor	BP: metabolic process
			CC: chloroplast
	XHP01127_1	single myb histone 6	BP: nucleosome assembly
	XHP01134_1	sensitive to proton rhizotoxicity 1	BP: response to acidity
	XHP01302_1	ILR3 (iaa-leucine resistant3); DNA binding / transcription factor	MF: sequence-specific DNA binding transcription factor activity
	XHP01475_1	iaa-leucine resistant3, ILR3	MF: sequence-specific DNA binding transcription factor activity
	XHP01502_1	DNA binding / DNA-directed RNA polymerase	BP: purine nucleobase metabolic process
			BP: pyrimidine nucleobase metabolic process
			CC: chloroplast
			MF: DNA-directed RNA polymerase activity
	XHP01696_1	transcription factor	MF: sequence-specific DNA binding transcription factor activity
	XHP02058_1_M	MSI4/FVE	BP: flower development
			BP: leaf morphogenesis
			MF: histone acetyltransferase activity
	XHP00854_1	SWINGER	BP: endosperm development
			BP: regulation of gene expression by genetic imprinting
			BP: protein methylation
			MF: histone-lysine N-methyltransferase activity

Table 4.12. (continued)

PAMSAM cluster	Contig ID	Contig description	Additional annotation information
1 (continued)	XHP00775_1	PICKLE	BP: negative regulation of abscisic acid mediated signaling pathway
			BP: response to gibberellin stimulus
			BP: regulation of lateral root development
			BP: cell proliferation
			BP: response to auxin stimulus
			BP: negative regulation of transcription, DNA-dependent
			MF: nucleoside-triphosphatase activity
	XHP01193_1	DDB1A (DAMAGED DNA BINDING PROTEIN 1A)	BP: negative regulation of photomorphogenesis
			BP: negative regulation of transcription, DNA-dependent
NDE	Xh_LDR_51B02r	WRKY DNA-binding protein 23	BP: response to auxin stimulus
	Xh_LRF_01G108	plastid transcriptionally active 14	CC: plastid chromosome
	XHP00074_1	basic region/leucine zipper motif 53	MF: protein heterodimerization activity
	XHP00286_1	stress enhanced protein 1	BP: ovule development
			MF: chlorophyll binding
	XHP00323_1	ZML1 (ZIM-LIKE 1); sequence-specific DNA binding / transcription factor/ zinc ion binding	MF: sequence-specific DNA binding transcription factor activity
	XHP00328_1	mads-box transcription factor 15	MF: sequence-specific DNA binding transcription factor activity
	XHP00339_1	Basic-leucine zipper (bZIP) transcription factor family protein	MF: protein dimerization activity
	XHP00350_1	NF-YB3	MF: sequence-specific DNA binding
	XHP00408_1	FPA; RNA binding	BP: positive regulation of flower development
			BP: embryo development
	XHP00577_1	NAC 014	BP: multicellular organismal development
	XHP00630_2	WRKY DNA-binding protein 33	MF: sequence-specific DNA binding transcription factor activity
	XHP00783_1	LUG (LEUNIG); transcription repressor	BP: flower development
			MF: transcription repressor activity
			MF: protein heterodimerization activity
	XHP01329_1	NF-YC3	MF: sequence-specific DNA binding
	XHP01405_1	AGAMOUS-like 19, MADS-box transcription factor	MF: sequence-specific DNA binding transcription factor activity
	XHP02004_1_M	origin of replication complex 1B	BP: double fertilization forming a zygote and endosperm
			MF: double-stranded methylated DNA binding

NDE: non-differentially expressed; BP: Biological process; MF: Molecular function; CC: Cellular component. Contigs shown in red represent transcription repressors based on their GO annotation in Blast2GO.

The medoid of cluster 6 (comprising 11 contigs) showed a later up-regulation response to desiccation with maximum transcript abundance observed at 40% RWC. Although the mRNA transcript abundance subsequently decreased, it was still maintained at a level substantially higher than that observed in the hydrated state. The mediod of cluster 4 showed the latest up-regulation, with maximum transcript abundance observed at 20% RWC, and

which remained high in the desiccated leaves. It is possible that the mRNA transcripts in Cluster 4 are stored in desiccated leaves, and are immediately translated upon rehydration to activate genes required for recovery from desiccation, and repair of cellular damage. These 8 mRNA transcripts are largely absent from hydrated leaves, and thus are unlikely to be important for the reactivation of photosynthesis and metabolism. These include an orthologue of BTB and TAZ domain protein 2 (BT2) (XHP01783_1) (Table 4.12). BT2 expression is affected by various abiotic stress conditions such as cold and ROS, and plays an essential role during gametogenesis, as well as throughout the plant development in *A. thaliana* (Mandadi *et al.*, 2009; Robert *et al.*, 2009). BT2 has also been reported to suppress ABA-mediated inhibition of germination (Mandadi *et al.*, 2009), thus it is highly likely to play an essential role in reactivating the genes required for normal development from their inhibition.

Cluster 7 medoid (comprising 8 contigs) showed an early transient down-regulation response with minimum transcript abundance level observed at 60% RWC, then subsequently increased and maintained at a level lower than the hydrated state. However, two of the contigs from this cluster, XHP01067_1 (plastid transcriptionally active 6) and XHP01446_1 (BolA-like family protein) have shown an initial up-regulation trend of expression (Fig. 4.11 inserts; Table 4.12). Cluster 2 and cluster 5 medoids also showed an early down-regulation response, but with minimum transcript abundance observed at the later 40% RWC. The abundance of the 19 contigs in cluster 2 remained relatively stable throughout the remaining course of desiccation. Whereas for the 11 contigs in cluster 5, the abundance subsequently increased to a level relatively lower than the hydrated state. Many transcription repressors are found in clusters 2 and 5, these include the two discussed (Section 4.3.3) TPL (XHP01131_1), and a zinc-finger protein 2 (XHP00948_1) that has been annotated with GO terms associated with water deprivation response and embryo development (Table 4.12).

Cluster 1 medoid showed a late down-regulation response with minimum transcript abundance observed in the desiccated leaves. The transcript abundance of these contigs in showed initial decrease upon desiccation except for XHP01502_1 (DNA binding / DNA-directed RNA polymerase) that showed an initial up-regulation trend of expression upon desiccation (Figure 4.11 inserts; Table 4.12). Transcription repressors were also identified in cluster 1 contigs, these include XHP00775_1 (PICKLE) and XHP00857_1 (SWINGER), the two regulators involved in the repression of seed traits during seed germination. PICKLE is a CHD3-type (chromodomain/helicase/DNA-binding domain) chromatin-remodeling factor. In

A. thaliana, loss of function in PICKLE was shown to fail in the suppression of LEC1 and FUS3, the transcription factors involved during early embryogenesis, which in turn caused abnormally persistent expression of embryogenesis genes in adult plant (Henderson *et al.*, 2004; Li *et al.*, 2005; Belin and Lope-Molina, 2008). SWINGER is a core component of a large protein complex called polycomb repressive complex 2 (PRC2) that binds to chromatin and catalyzes the deposition of trimethylation of Lysine 27 on histone H3 (H3K27me3) to repress gene expressions (Wang *et al.*, 2006; Schuettengruber *et al.*, 2007; Schwartz and Pirrotta, 2008; Bouyer *et al.*, 2011). It has been shown that PRC2 is required to switch from embryonic to seedling phase, and mutant seeds with defect in PRC2 showed enhanced dormancy and germination defects. Many genes controlling seed maturation and dormancy are marked by H3K27me3, and a defect in PCR2 functions results in up-regulation of these genes (Bouyer *et al.*, 2011). Connections between PICKLE and SWINGER have also been looked at and reported that PICKLE-dependent genes were enriched for H3K27me3 epigenetic mark, and lack of PICKLE caused reduced H3K27me3 levels at selected loci of the genes controlling seed maturation and dormancy (Zhang *et al.*, 2008; Aichinger *et al.*, 2009).

Sixty-eight percent of the differentially expressed transcription factors were identified to be down-regulated during desiccation in *X. humilis*. Furthermore, there was no known transcription repressor identified in the up-regulated clusters of 3, 4 and 6. Together with the functional term related to transcription regulator activity identified in the down-regulated contigs (Fig. 4.2, cluster 2; Fig. 4.3), these again may implicate that the genes conferring desiccation tolerance in *X. humilis* may have predominantly be activated via de-repression than being activated by newly up-regulated transcription activators.

4.3.10 Clustering of a common set of 772 orthologues during desiccation in X. humilis and during seed development and abiotic stress in A. thaliana

In addition to the characterization of desiccation response in *X. humilis* leaves by clustering of co-expressed LEA, antioxidants and transcription factor contigs, the expression of a common set of 772 orthologues in *X. humilis* and *A. thaliana* was analyzed, clustered and compared in order to further investigate the relationship between the seed maturation and abiotic stress response to osmotic stress (in a desiccation sensitive plant) to vegetative desiccation (in a desiccation tolerant plant). In *X. humilis* leaves, 615 orthologues (80%) were found to be differentially expressed in response to desiccation, and were grouped into 7 clusters by PAMSAM (Fig. 12A). The medoids of these 7 clusters were similar to those derived from the full set of 1361 differentially expressed contigs (Fig. 4.2).

These clusters of orthologues were significantly enriched for the similar functional terms as the clusters of contigs. For example, among the 108 desiccation down-regulated orthologues from cluster 2, terms related to photosynthesis were over-represented with terms such as “photosynthesis”, “photosynthesis light reaction”, “photosystem”, “photosystem I” and “chlorophyll binding” overlapped with the terms identified in cluster 2 contigs (Fig. 4.3, terms indicated by blue asterisks).

Furthermore, terms related to ribosome were found over-represented in both the 85 cluster 6 orthologues and the 147 cluster 6 contigs, with overlapping terms such as “translation” and “structural constituent of ribosome” (Fig. 4.5, terms indicated by blue asterisks). This analysis suggests that the subset of 772 orthologues comprised a reasonable representation of the original 1680 *X. humilis* contigs.

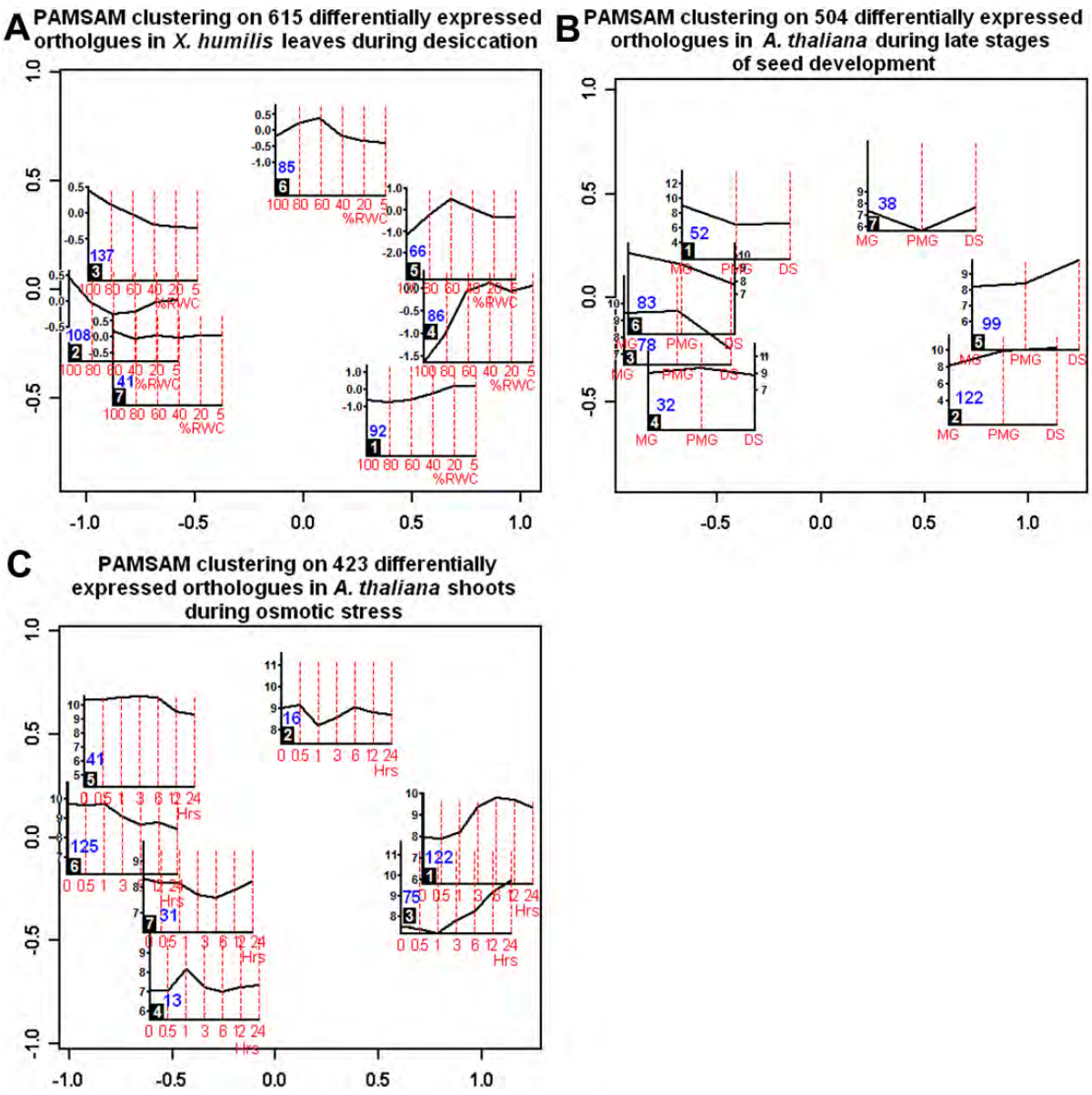


Figure 4.12. PAMSAM clustering on differentially expressed orthologues identified in *X. humilis* and *A. thaliana*. The 615, 504 and 423 differentially expressed orthologues identified during desiccation in *X. humilis* leaves (A), during seed development in *A. thaliana* (B) and in *A. thaliana* shoots during osmotic stress (C) respectively were clustered into 7 clusters by PAMSAM algorithm. The medoids/clusters were partitioned and plotted in a two-dimensional Sammon map, indicating their approximate similarity to each other. Each medoid/cluster was represented by a small graph that showed the expression values in $\log_2$ ratios (y-axis of cluster graph) across the 6 different stages of water loss: 100%, 80%, 60%, 40%, 20% and 5%RWC (A); or in $\log_2$ intensity values (y-axis of cluster graph) across the 3 selected stages of seed development: MG (mature green embryo), PMG (post mature green embryo) and DS (dry seed) (B); or in $\log_2$ intensity values across the 7 different time point during mannitol treatment (C). Total number of contigs in each cluster was indicated by the number in blue text above the cluster number.

In *A. thaliana*, 504 orthologues (65%) were identified as being differentially expressed during seed development from the matured green stage (MG) through the post mature green (PMG) to the dry seed stage (DS). Only these 3 late stages of seed development process were considered in the analysis, because desiccation tolerance is only acquired between these 3 late stages. In order to ensure a fair comparison between the *A. thaliana* and *X. humilis* datasets, these 504 differentially expressed orthologues were also clustered into 7 groups (Fig 4.12B).

Clusters 2, 4, 5 and 7 represent differentially expressed orthologues identified during the last stages of seed development. Although cluster 4 was clustered with the down-regulated clusters 1, 3 and 6, its medoid showed an up-regulation trend expression with initial increase in transcrirpt abundance at PMG, which was subsequently decreased and maintained in the dry seed at a level similar to that observed in MG. The medoids of clusters 2 and 5 both showed a consistent up-regulation response, with a prominent increase of abundance observed between MG and PMG in the 122 Cluster 2 orthologues, and between PMG and DS in the 99 orthologues in cluster 5. The Cluster 7 medoid showed an initial decrease in transcript abundance at PMG, before increasing to a level relatively higher the MG in the dry seeds. No over-represented functional term was identified in these 4 clusters of seed maturation up-regulated orthologues.

For the down-regulated clusters, although a prominent decrease in transcript abundance level of cluster 3 orthologues was observed in dry seed, similar to cluster 4 orthologues, cluster 3 medoid showed an initial increase in transcrirpt abundance at PMG. Cluster 1 medoid showed a decrease in transcript abundance at PMG, which subsequently elevated and maintained at a level significantly lower than MG in the dry seeds. Functional terms related to photosynthesis such as “photosynthesis, light harvesting” and “chlorophyll binding” were identified to be enriched in the 52 cluster 1 orthologues (Fig. 4.13). Similar terms were identified in the down-regulated cluster 2 orthologues found in *X. humilis* leaves during desiccation, which also shared a similar pattern of expression (Fig. 4.3; Fig. 4.12A).

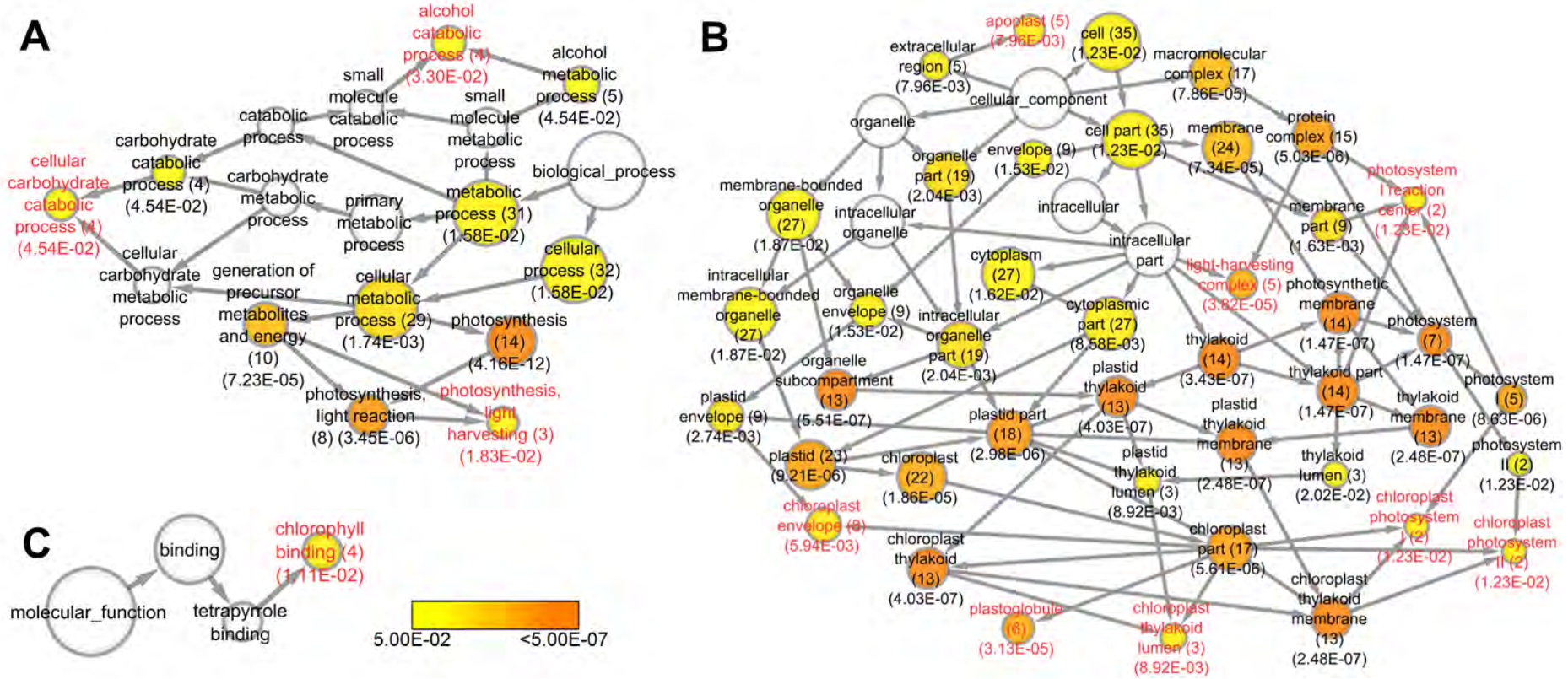


Figure 4.13. Over-representation of biological process (A), cellular component (B) and molecular function (C) terms found in the set of 52 cluster 1 differentially expressed orthologues identified in *A. thaliana* during seed development. The enrichment of ontology terms was analyzed using hypergeometric test with a significance level of p-value= 0.05. The node size is proportional to the number of contigs in the test set being annotated to that node, and the significantly over-represented terms are presented as coloured nodes with a colour scale ranging from light yellow (p-value= 0.05) to dark orange (p-value< 5.00E-07). The two numbers shown in brackets under each significantly over-represented term indicate the number of contigs being annotated, and the p-value after FDR correction to that node respectively.

Cluster 6 medoid showed a consistent down-regulation response during the late stages of seed development. Ribosome related functional terms such as “translation” and “structural constituent of ribosome” were identified as over-represented in this cluster (Fig. 4.14). Similar terms were identified in the early transiently up-regulated cluster 6 orthologues found in *X. humilis* leaves during desiccation (Fig. 4.5, terms indicated by blue asterisks), but did not show the similar pattern of expression with the initial increase in transcript level observed between 100 and 60%RWC (Fig. 4.12A).

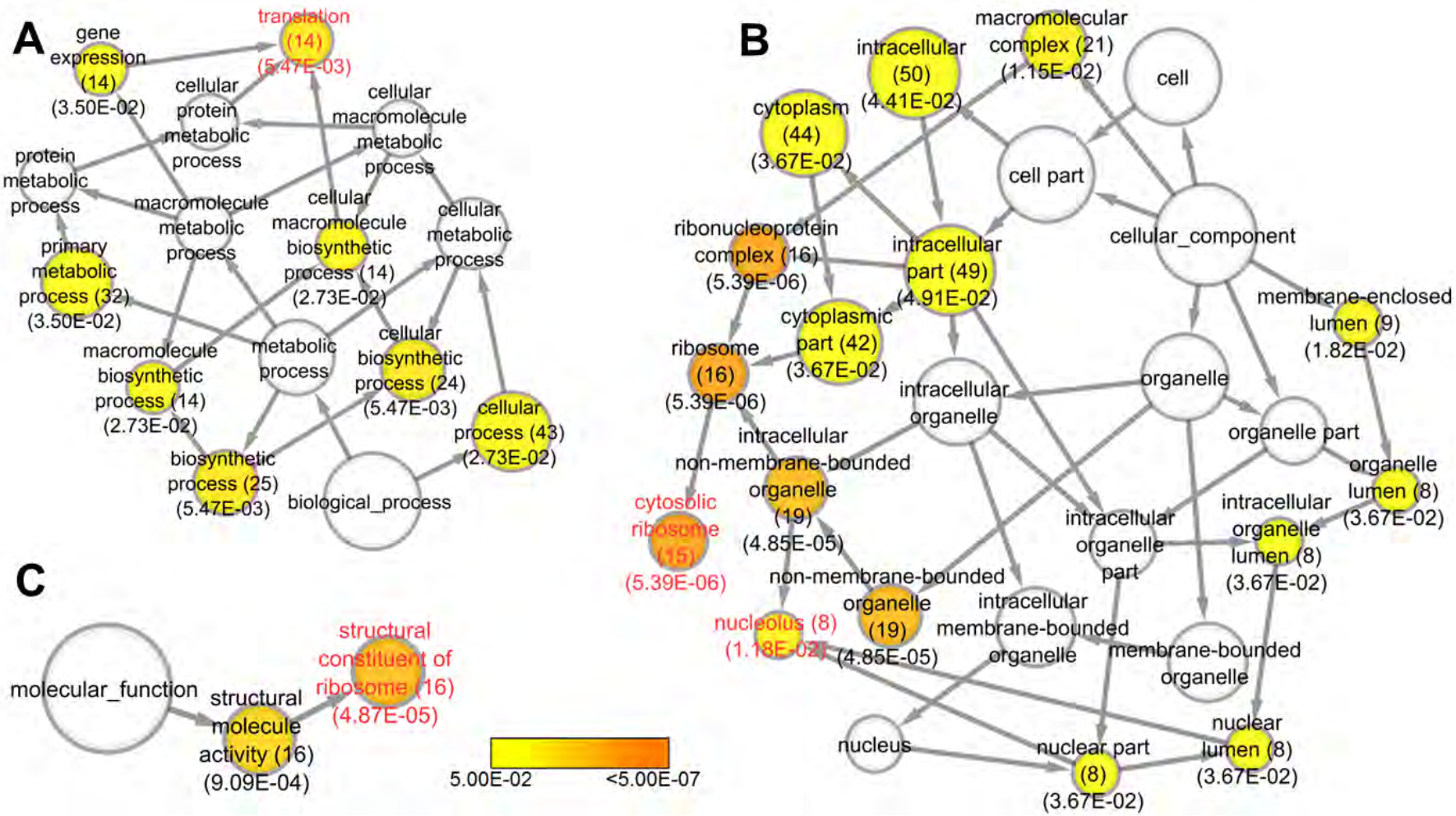


Figure 4.14. Over-representation of biological process (A), cellular component (B) and molecular function (C) terms found in the set of 83 cluster 6 differentially expressed orthologues identified in *A. thaliana* during seed development. The enrichment of ontology terms was analyzed using hypergeometric test with a significance level of p-value= 0.05. The node size is proportional to the number of contigs in the test set being annotated to that node, and the significantly over-represented terms are presented as coloured nodes with a colour scale ranging from light yellow (p-value= 0.05) to dark orange (p-value< 5.00E-07). The two numbers shown in brackets under each significantly over-represented term indicate the number of contigs being annotated, and the p-value after FDR correction to that node respectively.

Fewer orthogues (55%) were identified as being significantly differentially expressed during osmotic stress in *A. thaliana*. The 428 differentially expressed orthologues found in *A. thaliana* shoots subjected to a 24-hour course of mannitol treatment were clustered into 7 groups (Fig 4.12C). Although cluster 4 was clustered with the down-regulated clusters 6 and 7 on the Sammon map, its medoid actually showed an early up-regulation trend of expression. The transcript abundance increased after 30 minutes of mannitol treatment and reached maximum level after 1 hour, which subsequently decreased and reached minimum level after 6 hours of treatment. The abundance increased again and maintained at a level similar to that in the untreated plants at the end of the mannitol treatment. Medoids of cluster 1 and 3 showed apparent osmotic up-regulated trend of expression. Maximum transcript abundance level was observed after 6 hours of mannitol treatment in the 122 cluster 1 orthologues, which subsequently decreased and maintained at a level significantly higher than that observed before the treatment. Cluster 3 medoid (comprising 75 orthologues) showed an initial decrease in transcript abundance after an hour of mannitol treatment, which then consistently increased and reached maximum level after 24 hour of the treatment.

Cluster 5 medoid showed an overall down-regulated expression in response to osmotic stress. However, a slow increase in the transcript abundance level was observed during the first 6 hours of mannitol treatment, before being decrease towards the end of the treatment. The over-represented terms identified from the 41 orthologues were functional terms related to photosynthesis such as "photosynthesis, light harvesting" and "chlorophyll binding" (Fig. 4.15). Similar terms were identified in the down-regulated cluster 2 orthologues found in *X. humilis* leaves during desiccation (Fig. 4.3, terms indicated by blue asterisks) and cluster 1 ortholgues that were down-regulated towards the end of seed maturation (Fig. 4.13). However, *X. humilis* cluster 2 orthologues and *A. thaliana* seed cluster 1 orthologues showed different expression patterns during desiccation and maturation respectively, they all showed an initial up-regulation before the expression was down-regulated towards the end of desiccation or maturation (Fig. 4.12A; Fig. 4.12B).

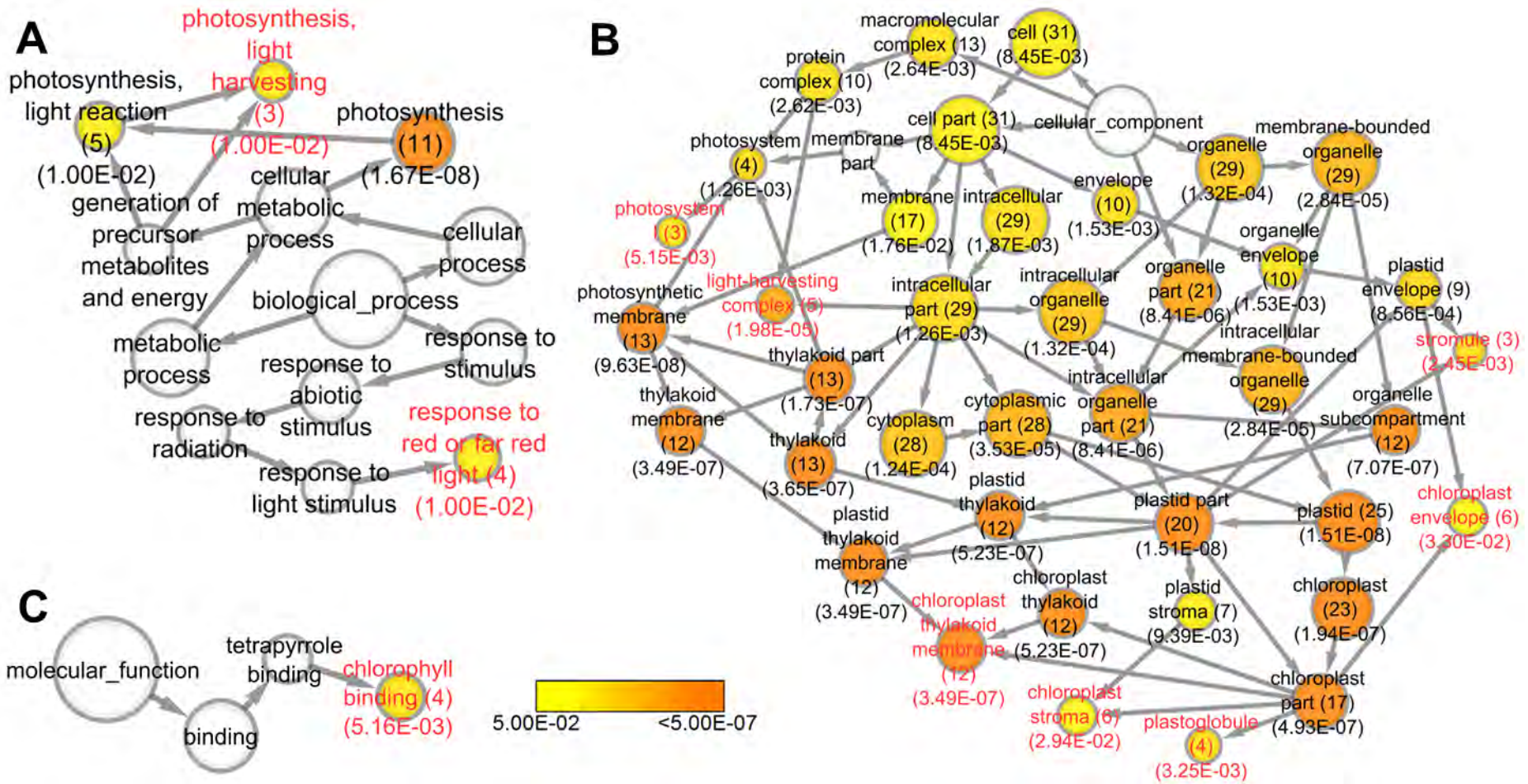


Figure 4.15. Over-representation of biological process (A), cellular component (B) and molecular function (C) terms found in the set of 41 cluster 5 differentially expressed orthologues identified in *A. thaliana* shoots during osmotic stress. The enrichment of ontology terms was analyzed using hypergeometric test with a significance level of p-value= 0.05. The node size is proportional to the number of contigs in the test set being annotated to that node, and the significantly over-represented terms are presented as coloured nodes with a colour scale ranging from light yellow (p-value= 0.05) to dark orange (p-value< 5.00E-07). The two numbers shown in brackets under each significantly over-represented term indicate the number of contigs being annotated, and the p-value after FDR correction to that node respectively.

Clusters 2 medoids showed transient expressions in response to osmotic stress. After an initial up-regulation during the first 30 minutes of mannitol treatment, the transcript abundance decreased and reached minimum level after 60 minutes of mannitol treatment. The abundance subsequently increased and reached maximum level after 6 hours of treatment, then decreased to a level just below the unstressed level by the end of the mannitol treatment.

Clusters 6 and 7 medoids (comprising 125 and 31 orthologues respectively) showed down-regulation response to osmotic stress. A prominent decrease in transcript abundance was observed after 60 minutes of the mannitol treatment in both clusters. Whilst the abundance in cluster 6 decreased consistently throughout the rest of the treatment, a subsequent increase of abundance was observed in cluster 7 after 6 hours, and maintained at a level similar to that in the untreated plants at the end of the mannitol treatment. Ribosome related terms such as "translation" and "structural constituent of ribosome" were identified as over-represented in the 125 cluster 6 orthologues (Fig. 4.16).

There was no enriched functional terms identified in clusters 1, 2, 3, 4 and 7. The functional enrichment results identified from the down-regulated clusters of orthologues suggested that photosynthesis, as well as ribosome biogenesis and translation were commonly down-regulated during desiccation in *X. humilis*, as well as in *A. thaliana* during the maturation phase of seed development, and in abiotic stress response to mannitol treatment.

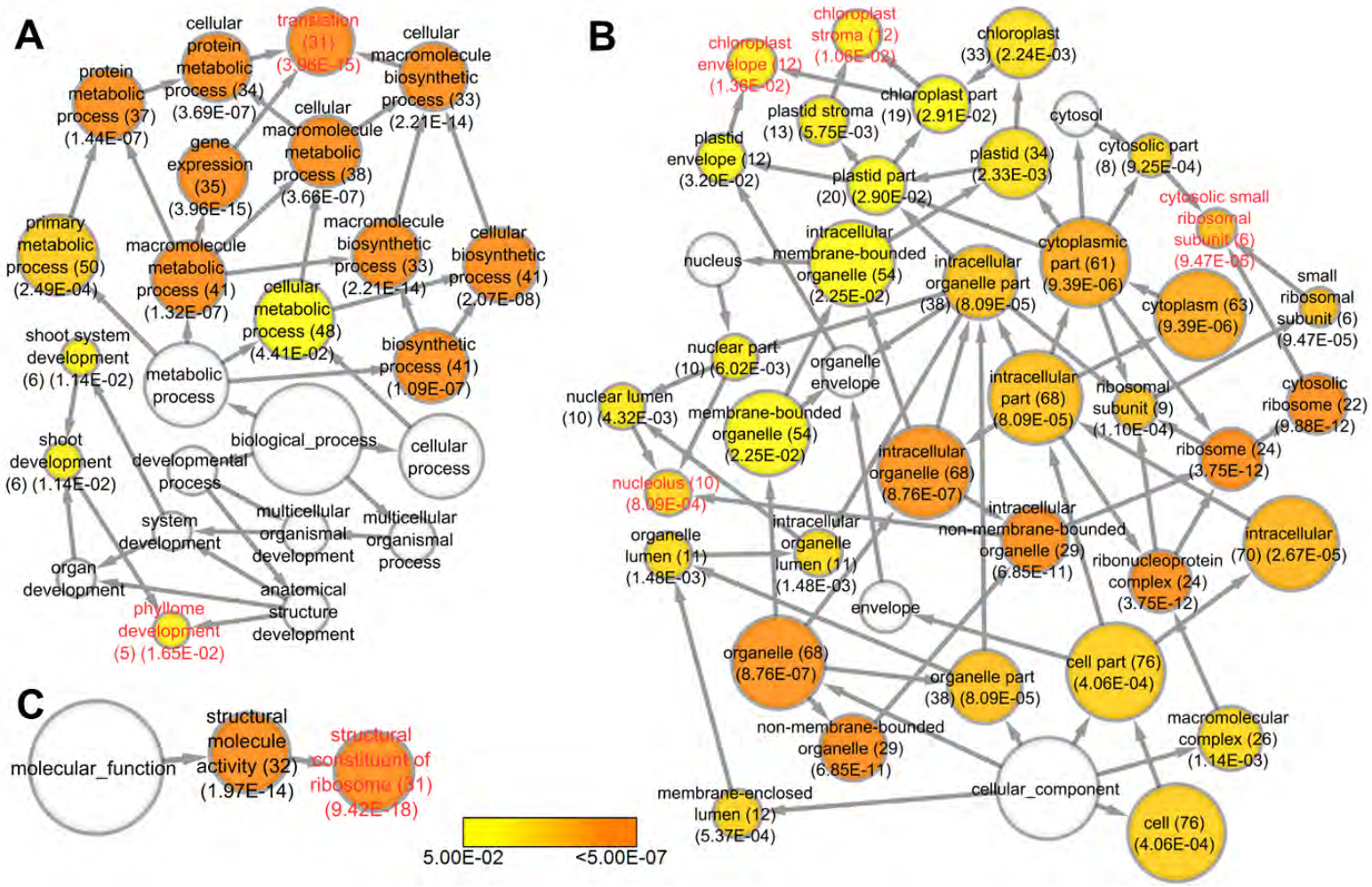


Figure 4.16. Over-representation of biological process (A), cellular component (B) and molecular function (C) terms found in the set of 125 cluster 6 differentially expressed orthologues identified in *A. thaliana* shoots during osmotic stress. The enrichment of ontology terms was analyzed using hypergeometric test with a significance level of p-value= 0.05. The node size is proportional to the number of contigs in the test set being annotated to that node, and the significantly over-represented terms are presented as coloured nodes with a colour scale ranging from light yellow (p-value= 0.05) to dark orange (p-value< 5.00E-07). The two numbers shown in brackets under each significantly over-represented term indicate the number of contigs being annotated, and the p-value after FDR correction to that node respectively.

Venn analysis was performed on the induced, as well as repressed orthologues identified in *X. humilis* leaves during desiccation and in *A. thaliana* during seed development and osmotic stress, in order to assess the transcriptome overlap and similarity between the 3 expression profiles. The list of orthologues from the clusters whose medoid showed prominent overall up- (clusters 1, 4 and 5 from the desiccation profile; clusters 2, 5 and 7 from the seed maturation profile; clusters 1 and 3 from the osmotic profile) or down-regulation (clusters 2 and 3 from the desiccation profile; clusters 1, 3 and 6 from the seed maturation profile; clusters 5 and 6 from the osmotic profile) trend were compared. The analysis showed a significant overlap between genes up-regulated during desiccation in *X. humilis*, and during seed maturation and the response to osmotic stress in *A. thaliana*. A common set of 103 orthologues were up-regulated in both *X. humilis* leaves during desiccation and in *A. thaliana* during the maturation phase of seed development (Fig. 4.17A).

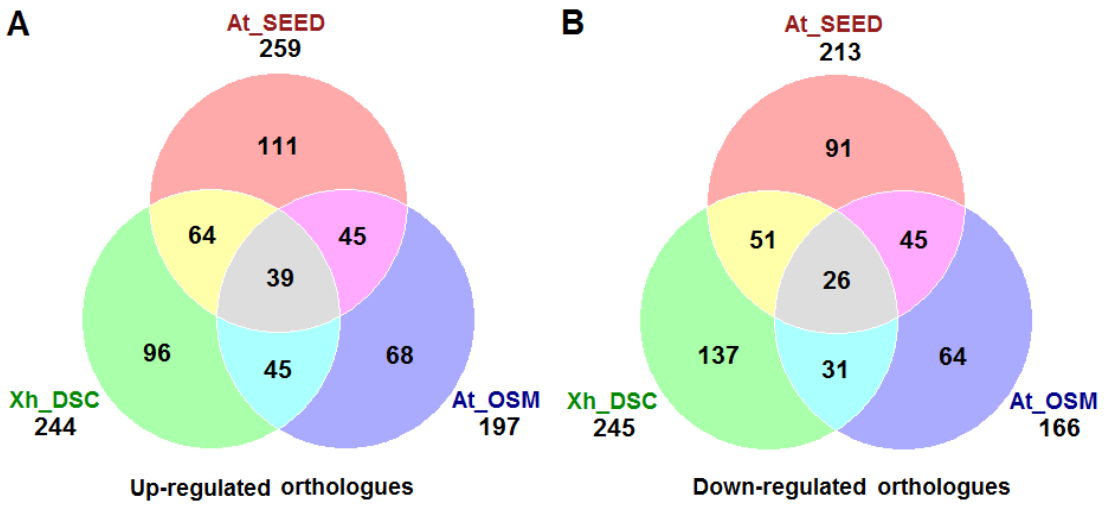


Figure 4.17. Venn diagram showing the overlap between the up-regulated orthologues (A) and down-regulated orthologues (B) identified in *X. humilis* during desiccation (Xh_DSC), and in *A. thaliana* during seed development (At_SEED) and during osmotic stress (At_OSM). Numbers underneath the expression profile name indicated the total number of up- or down-regulated orthologues identified in that expression profile from the common set of 772.

A common set of 84 orthologues were found being up-regulated in both *X. humilis* during desiccation, and in shoots of *A. thaliana* during osmotic stress. Furthermore, a common set of 84 orthologues were found being up-regulated in both seed development and osmotic stress in *A. thaliana*. The statistical significance of common up-regulated orthologues between the three profiles was calculated using the hypergeometic distribution calculator as described in chapter 3, and showed that all overlaps of up-regulated orthologues were significant (Table 4.13).

Table 4.13. Summary of statistical significance of common up-regulated genes calculated using the hypergeometic distribution calculator.

Profile 1 (no. of orthologue)	Profile 2 (no. of orthologue)	No. of genes overlapped	Representation factor	p-value
Xh_DSC (244)	At_SEED (259)	103	37.1	<1.21E-140
Xh_DSC (244)	At_OSM (197)	84	39.7	<2.21E-116
At_SEED (259)	At_OSM (197)	84	37.1	<8.99E-114

Xh_DSC (desiccation response in *X. humilis* leaves); At_SEED (*A. thaliana* seed development); At_OSM (osmotic stress response in *A. thaliana* shoots).

Similar observations were also found in the analysis on the lists of down-regulated orthologues identified from the 3 expression profiles (Fig. 4.17B; Table 4.14). The overlap analysis on up- and down-regulated orthologues identified from the 3 expression profiles revealed that there was significant similarity between the transcriptome induced or repressed between seed development, osmotic stress, and vegetative desiccation. This was supported in the findings reported by Wohlbach *et al.*, (2008), in which a plasma membrane histidine kinase (ATHK1) has been shown to involve in the water stress response during early vegetative stages of plant growth, as well as in the regulation of desiccation processes during seed development. However, no functional terms were identified in these significant overlaps of up- or down-regulated orthologues, except for the 26 commonly down-regulated orthologues found between *X. humilis* desiccation response, and seed maturation and osmotic stress responses in *A. thaliana*, where functional terms associated with photosynthesis were identified as significantly over-represented (data not shown).

Table 4.14. Summary of statistical significance of common down-regulated genes calculated using the hypergeometic distribution calculator.

Profile 1 (no. of genes)	Profile 2 (no. of genes)	No. of genes overlapped	Representation factor	p-value
Xh_DSC (245)	At_SEED (213)	77	33.6	<3.52E-99
Xh_DSC (245)	At_OSM (166)	57	31.9	<3.855E-71
At_SEED (213)	At_OSM (166)	71	45.7	<1.07E-102

Xh_DSC (desiccation response in *X. humilis* leaves); At_SEED (*A. thaliana* seed development); At_OSM (osmotic stress response in *A. thaliana* shoots).

4.3.11 Assessment of seed traits in vegetative desiccation tolerance in X. humilis based on the expression patterns of LEA, antioxidant and transcription factor orthologues observed during vegetative desiccation in X. humilis, seed development and osmotic stress in A. thaliana

The overlap analysis on the up- or down-regulated orthologues identified from the 3 expression profiles revealed significant similarities between the transcriptome prominently induced or repressed between seed development, osmotic stress, and vegetative desiccation (Fig. 4.17; Table 4.13; Table 4.14). The results suggested that several of the biological pathways are shared across these responses. In order to gain a more detailed insight into the possible differences in the transcriptome induced or repressed between the 3 responses, the overlap was characterized further by comparing the expression patterns of specific classes of genes: LEAs, antioxidants and transcription factors from the 3 expression profiles.

4.3.11.1 LEAs specific to seed development in A. thaliana were shown to be up-regulated in X. humilis leaves during desiccation

Eleven LEA contigs were identified in the common set of 772 orthologues. The expression patterns of these 11 LEAs identified during desiccation in *X. humilis*, and during seed development and osmotic stress in *A. thaliana* were compared to investigate the functional overlap between these genes which are known to be important for desiccation tolerance (Table 4.15). While At1g01470, At4g15910, At2g35300 and At5g06760 (classes 8, 7, 4 and 4 respectively) all showed up-regulation trend, the remaining LEAs showed uncorrelated expression patterns in all 3 profiles. Two class 3 LEAs, At3g17520 and At1g52690, were shown to be responsive to vegetative desiccation and osmotic stress responses, but not to seed maturation response. At3g53040 (class 3) was shown to be responsive solely during seed maturation, whilst being constantly expressed in *X. humilis* and absent in *A. thaliana* shoots during osmotic stress.

The 3 previously discussed seed specific LEAs (At1g72100, At3g22490 and At1g04560, Section 4.3.7) indeed showed seed specific expression trend in the analysis. In contrast to the two LEAs showing prominent up-regulation during seed maturation, At1g04560 (class 10) showed a slight initially up-regulation at stage PMG before decreasing towards the end of maturation, a profile similar to that found in *X. humilis*. At1g72100 (class 3 LEA) and At3g22490 (class 6) showed up-regulation during seed maturation, but no expression at all in *A. thaliana* shoots during osmotic stress. LEA orthologues At3g53040 (class 3) and

At2g35300 (class 4) have also been reported as seed specific LEAs in *A. thaliana* (Illing *et al.*, 2005; Rodrigues *et al.*, 2010). At3g53040 showed no expression and an up-regulation in *A. thaliana* during osmotic stress and seed maturation respectively, and was shown to be constantly expressed in *X. humilis* leaves during desiccation. However, the expression of At2g35300 did not appear to be seed specific as revealed in this analysis, it showed up-regulation in all 3 expression profiles.

Table 4.15. Expression of *X. humilis* LEA othologues in *A. thaliana* during seed development and osmotic stress.

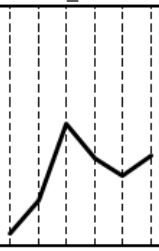
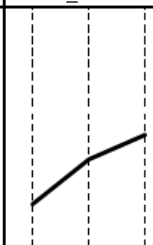
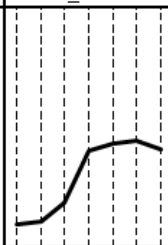
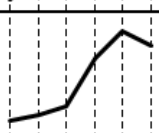
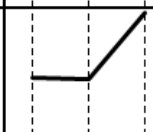
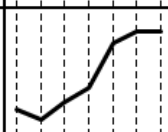
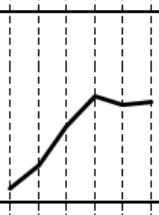
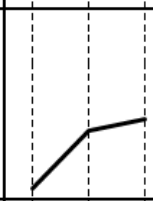
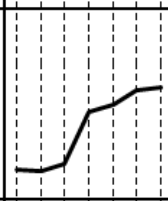
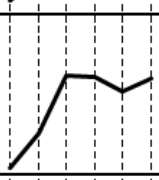
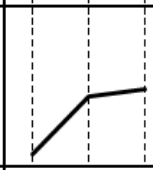
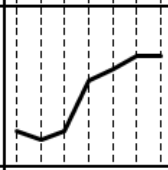
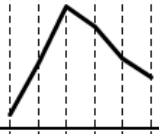
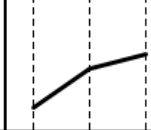
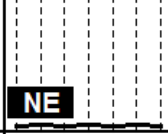
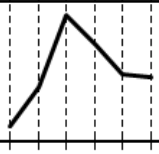
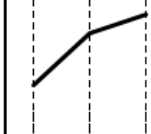
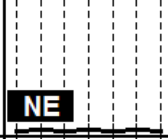
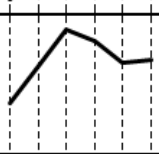
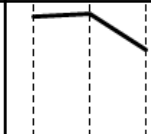
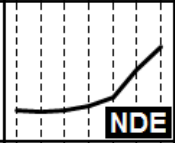
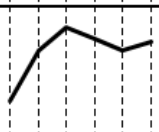
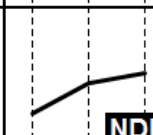
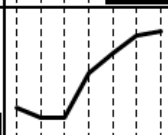
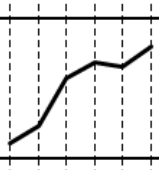
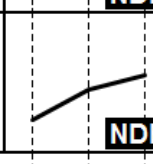
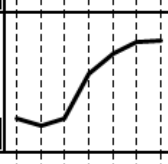
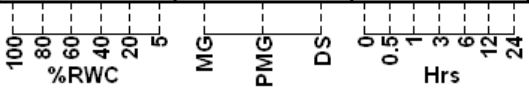
Ortholog ID	Ortholog description	LEA class	Expression Xh_DSC	Expression At_SEED	Expression At_OSM	Additional annotation information
At1g01470	**LEA 14**	8				BP: defense response to fungus BP: embryo development ending in seed dormancy BP: heat acclimation BP: response to cold BP: response to desiccation BP: response to water deprivation BP: response to wounding
At4g15910	**drought-induced 21**	7				BP: embryo development BP: response to water deprivation CC: chloroplast
At2g35300	**LEA 18**	4				BP: embryo development ending in seed dormancy BP: pollen tube development BP: response to osmotic stress BP: response to water deprivation BP: seed germination
At5g06760	**LEA 4-5**	4				BP: embryo development ending in seed dormancy BP: response to cold BP: response to osmotic stress BP: response to water deprivation
At1g72100	**LEA domain-containing protein**	3				BP: embryo development ending in seed dormancy BP: seed dormancy process CC: chloroplast
At3g22490	**Seed maturation protein**	6				BP: embryo development ending in seed dormancy
At1g04560	**AWPM-19-like family protein**	10				CC: extracellular region
At3g17520	**LEA family protein**	3				BP: embryo development ending in seed dormancy BP: fatty acid catabolic process
At1g52690	**LEA 7**	3				BP: cell wall modification involved in abscission BP: embryo development ending in seed dormancy

Table 4.15. (continued)

Ortholog ID	Ortholog description	LEA class	Expression Xh_DSC	Expression At_SEED	Expression At_OSM	Additional annotation information
At2g46140	**LEA protein**	8				BP: embryo development ending in seed dormancy BP: response to desiccation CC: apoplast CC: plasma membrane
At3g53040	**LEA protein, putative**	3	NDE		NE	BP: embryo development ending in seed dormancy BP: floral organ abscission



Xh_DSC (desiccation response in *X. humilis* leaves); At_SEED (*A. thaliana* seed development); At_OSM (osmotic stress response in *A. thaliana* shoots); NDE (non-differentially expressed). BP: Biological process; MF: Molecular function; CC: Cellular component. Additional annotation information was obtained from the *A. thaliana* GO Slim annotation file (version 2013/08/20) accessible at: ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/ATH_GO_GOSLIM.txt

The up-regulation or constant expression of the known seed specific LEAs observed in *X. humilis* leaves during desiccation, as well as their absence in *A. thaliana* shoot during osmotic stress, provided preliminary evidence to the hypothesis that vegetative desiccation tolerance observed in *X. humilis* may have derived from the activation of traits specific to orthodox seed development.

4.3.11.2 Seed specific antioxidant was found up-regulated in X. humilis leaves during desiccation

Investigation on the expression of the 15 antioxidant orthologues in the common set of 772 during *X. humilis* desiccation response, and seed development and osmotic stress in *A. thaliana*, showed that At1g48130, 1-cystine peroxiredoxin 1, (XHP00084_2) was indeed seed specific. It was highly expressed in the maturing *A. thaliana* seeds, but not expressed in the vegetative tissue in response to osmotic stress in *A. thaliana* (Table. 4.16). At3g11630, an orthologue of 2-cystine peroxiredoxin A (XHP01001_1) was shown to be down-regulated in *X. humilis* during desiccation, as well as during seed development and osmotic stress in *A. thaliana*. This agreed with the specific involvement of 2-cystine peroxiredoxin in the protection of functioning photosynthetic apparatus against oxidative damage.

Table 4.16. Expression of *X. humilis* antioxidant othologues in *A. thaliana* during seed development and osmotic stress.

Ortholog ID	Ortholog description	Expression Xh_DSC	Expression At_SEED	Expression At_OSM	Additional annotation information
At1g65820	microsomal glutathione s-transferase, putative				BP: ER to Golgi vesicle-mediated transport
At4g11600	glutathione peroxidase 6				BP: chromatin assembly or disassembly BP: circadian rhythm BP: response to salt stress
At3g11630	2-Cys peroxiredoxin A				BP: cuticle development BP: detection of biotic stimulus BP: MAPK cascade BP: negative regulation of defense response BP: photosynthesis, light reaction BP: protein targeting to membrane BP: response to cold BP: very long-chain fatty acid metabolic process CC: chloroplast CC: thylakoid MF: oxidoreductase activity MF: protein binding
At1g08830	copper/zinc superoxide dismutase 1				BP: cellular response to sucrose stimulus BP: gene silencing by miRNA BP: gluconeogenesis BP: glycolysis BP: removal of superoxide radicals BP: response to salt stress
At2g25080	glutathione peroxidase 1				CC: chloroplast
At3g52960	Thioredoxin superfamily protein				BP: defense response to bacterium CC: chloroplast CC: plant-type cell wall MF: oxidoreductase activity
At3g51030	thioredoxin H-type 1			NDE	BP: cell redox homeostasis BP: glycerol ether metabolic process BP: positive regulation of catalytic activity MF: electron carrier activity MF: enzyme activator activity MF: protein disulfide oxidoreductase activity
		100 80 60 40 20 5 %RWC	MG PMG DS	0 0.5 1 3 6 12 24 Hrs	

Table 4.16. (continued)

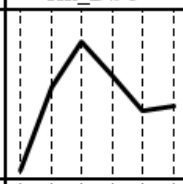
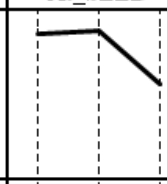
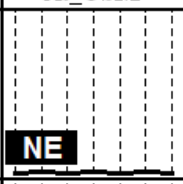
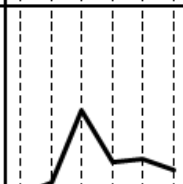
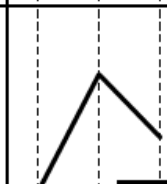
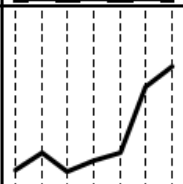
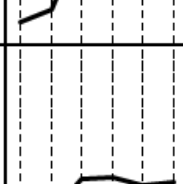
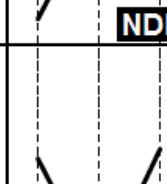
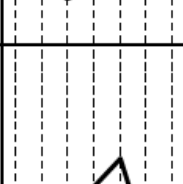
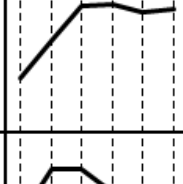
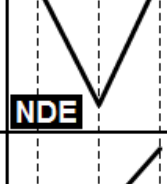
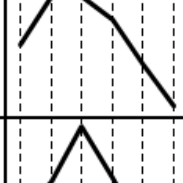
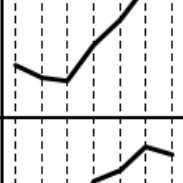
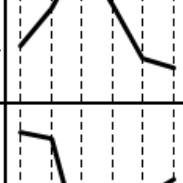
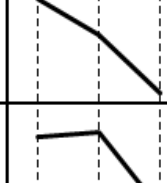
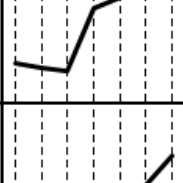
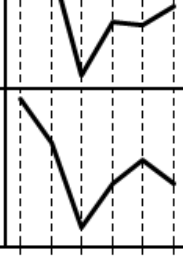
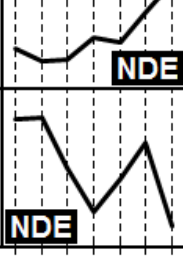
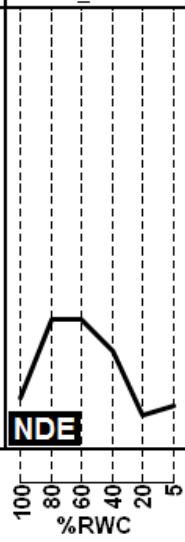
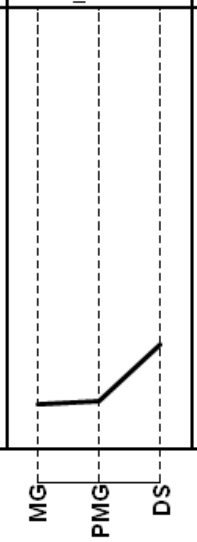
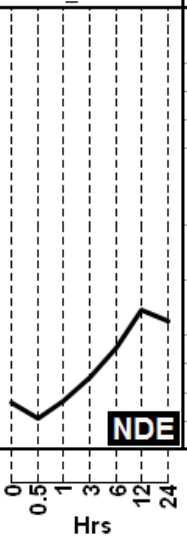
Ortholog ID	Ortholog description	Expression Xh_DSC	Expression At_SEED	Expression At_OSM	Additional annotation information
At1g48130	1-cysteine peroxiredoxin 1				BP: maintenance of seed dormancy BP: response to desiccation MF: oxidoreductase activity MF: thioredoxin peroxidase activity
At3g09640	ascorbate peroxidase 2				BP: protein folding BP: response to heat BP: response to high light intensity BP: response to hydrogen peroxide MF: heme binding
At2g43350	glutathione peroxidase 3				BP: abscisic acid mediated signaling pathway BP: cellular response to water deprivation BP: response to hydrogen peroxide CC: endosome CC: Golgi apparatus CC: mitochondrion
At1g65980	thioredoxin-dependent peroxidase 1				CC: chloroplast CC: plasma membrane MF: oxidoreductase activity
At1g60420	DC1 domain-containing protein				BP: cell redox homeostasis BP: regulation of pollen tube growth MF: oxidoreductase activity MF: protein-disulfide reductase activity
At5g05340	peroxidase 52				CC: cell wall CC: Golgi apparatus MF: heme binding MF: protein binding
At2g37130	Peroxidase superfamily protein				BP: defense response to fungus MF: heme binding

Table 4.16. (continued)

Ortholog ID	Ortholog description	Expression Xh_DSC	Expression At_SEED	Expression At_OSM	Additional annotation information
At1g78380	glutathione S-transferase TAU 19				BP: cellular response to water deprivation BP: fatty acid beta-oxidation BP: gluconeogenesis BP: glycolysis BP: proteasomal ubiquitin-dependent protein catabolic process BP: proteasome core complex assembly BP: response to misfolded protein BP: response to salt stress CC: chloroplast CC: plasma membrane CC: vacuolar membrane

Xh_DSC (desiccation response in *X. humilis* leaves); At_SEED (*A. thaliana* seed development); At_OSM (osmotic stress response in *A. thaliana* shoots); NDE (non-differentially expressed). BP: Biological process; MF: Molecular function; CC: Cellular component. Additional annotation information was obtained from the *A. thaliana* GO Slim annotation file (version 2013/08/20) accessible at: ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/ATH_GO_GOSLIM.txt

The up-regulation of the reported seed specific 1-cysteine peroxiredoxin 1 observed in the dehydrating leaves, as well as its maintained transcript abundance detected in the desiccated leaves of *X. humilis*, have once again provided a preliminary support to the hypothesis that vegetative desiccation tolerance may be achieved by the activation of seed mechanism in the vegetative tissues of resurrection plants.

4.3.11.3 Repressors of genes required during seed development were down-regulated in X. humilis leaves during desiccation but were constantly expressed in A. thaliana shoots during osmotic stress

In an attempt to dissect the similarities or differences during the regulation of vegetative desiccation, seed maturation and abiotic stress responses, only the 19 orthologues whose corresponding contigs were annotated with embryogenesis or abiotic stresses related GO terms were analyzed. The two transcription repressors PICKLE and SWINGER were shown to be down-regulated in leaves of *X. humilis* during desiccation and in *A. thaliana* seeds towards the end of maturation phase, but were shown to be constantly expressed in *A. thaliana* shoots in response to osmotic stress (Fig. 4.17). The down-regulation of known embryogenesis repressors observed in *X. humilis* leaves during desiccation again illustrated the involvement of seed specific mechanisms in the vegetative desiccation tolerance in resurrection plants.

Seed desiccation tolerance is normally lost upon germination. Several studies have shown that desiccation tolerance in germinated seeds can be re-established with application of mild osmotic stress in desiccation sensitive plants such as *M. truncatula* and *A. thaliana* (Buitink *et al.*, 2003; Maia *et al.*, 2011). However, the tolerance can only be restored within a limited time frame, usually during the very early phase of germination depending on the protrusion of the radicle length. The ability to activate the seed specific mechanisms by down-regulating the expression of their repressors observed in *X. humilis* suggested that desiccation tolerance in *X. humilis* leaves may have evolved by an extension of, or the maintenance of the early germination programme, where seedlings are initially desiccation tolerant by the reversible epigenetic modification of seed maturation genes mediated by PRC2.

Table 4.17. Expression of *X. humilis* transcription factor othologues in *A. thaliana* during seed development and osmotic stress.

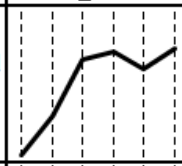
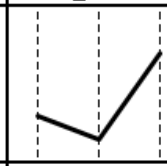
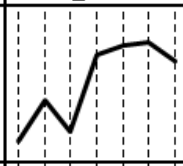
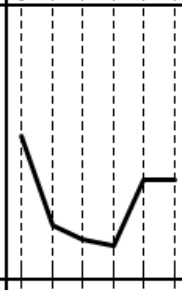
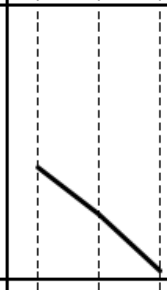
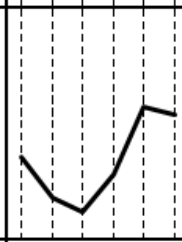
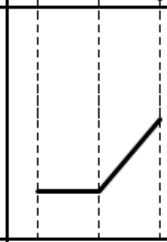
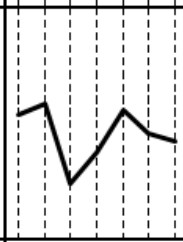
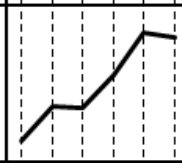
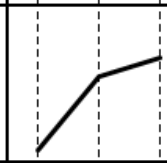
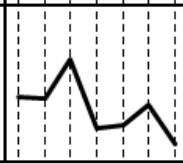
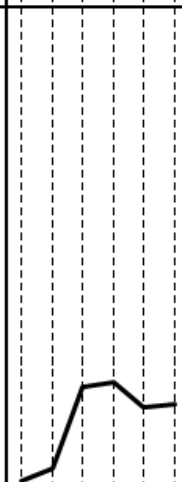
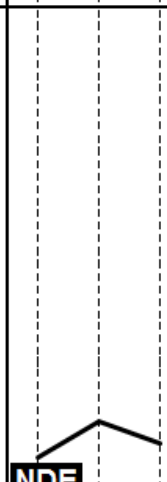
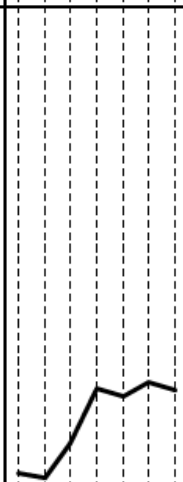
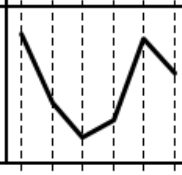
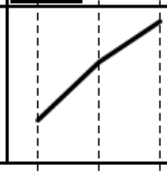
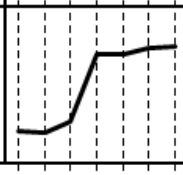
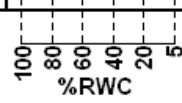
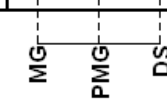
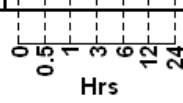
Ortholog ID	Ortholog description	Expression Xh_DSC	Expression At_SEED	Expression At_OSM	Additional annotation information
At2g19810	**Oxidation-related Zinc Finger 1**				BP: chlorophyll catabolic process BP: response to oxidative stress
At1g06040	**B-Box domian protein 24**				BP: glycolysis BP: Golgi organization BP: hyperosmotic response BP: response to abscisic acid stimulus BP: response to karrikin BP: response to salt stress BP: response to temperature stimulus BP: water transport
At4g37260	**myb domain protein 73**				BP: response to abscisic acid stimulus BP: response to chitin BP: response to ethylene stimulus BP: response to jasmonic acid stimulus BP: response to salicylic acid stimulus
At4g23750	**cytokinin response factor 2**				BP: cotyledon development BP: root development BP: transcription factor import into nucleus
At2g46270	**G-box binding factor 3**				BP: abscisic acid mediated signaling pathway BP: ethylene mediated signaling pathway BP: hyperosmotic salinity response BP: response to auxin stimulus BP: response to carbohydrate stimulus BP: response to ethylene stimulus BP: response to jasmonic acid stimulus BP: response to salt stress BP: response to superoxide BP: response to water deprivation BP: response to wounding
At3g14180	**6B-interacting protein 1-like 2**				BP: embryo development ending in seed dormancy BP: post-translational protein modification BP: seed maturation
					

Table 4.17. (continued)

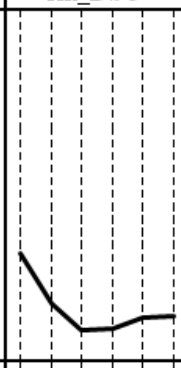
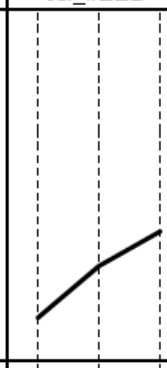
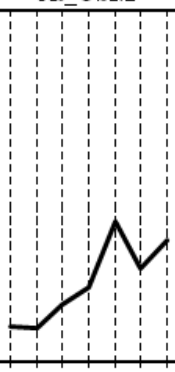
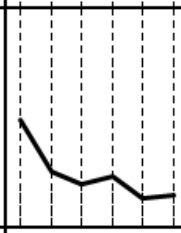
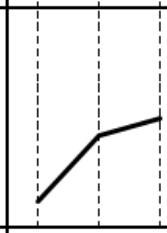
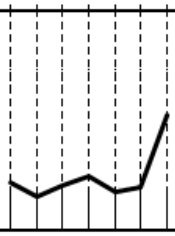
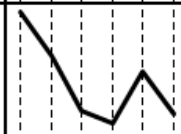
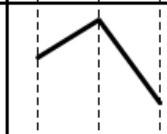
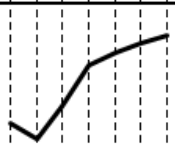
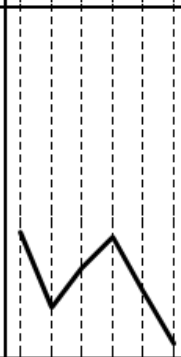
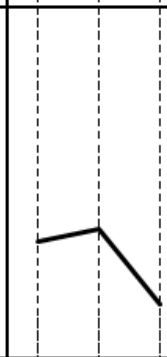
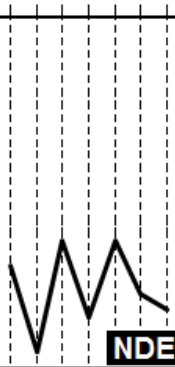
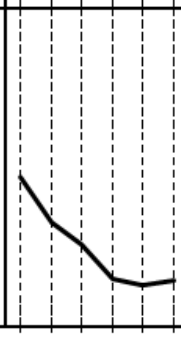
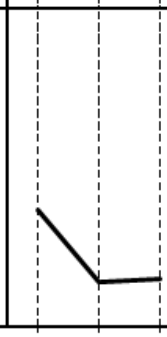
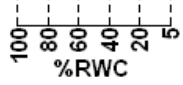
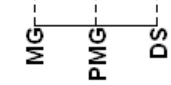
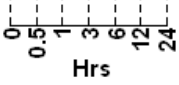
Ortholog ID	Ortholog description	Expression Xh_DSC	Expression At_SEED	Expression At_OSM	Additional annotation information
At3g20770	ETHYLENE-INSENSITIVE3				BP: cell death BP: cotyledon development BP: defense response to bacterium BP: response to abscisic acid stimulus BP: response to auxin stimulus BP: response to salt stress BP: response to superoxide BP: salicylic acid mediated signaling pathway BP: sugar mediated signaling pathway
At1g55920	serine acetyltransferase 2;1				BP: cysteine biosynthetic process from serine BP: nitrate transport BP: response to cold CC: chloroplast MF: serine O-acetyltransferase activity MF: transferase activity
At3g01470	homeobox 1				BP: chlorophyll catabolic process BP: leaf morphogenesis BP: response to salt stress
At2g25170	PICKLE				BP: cell proliferation BP: chromatin modification BP: methylation-dependent chromatin silencing BP: negative regulation of abscisic acid mediated signaling pathway BP: negative regulation of transcription, DNA-dependent BP: root development MF: DNA helicase activity
At4g02020	SWINGER				BP: chromatin silencing BP: DNA methylation BP: endosperm development BP: histone methylation BP: post-translational protein modification BP: production of miRNAs involved in gene silencing by miRNA BP: vegetative phase change BP: vernalization response
					

Table 4.17. (continued)

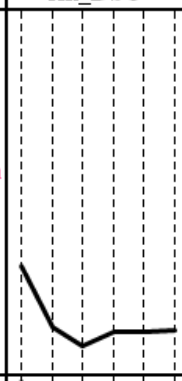
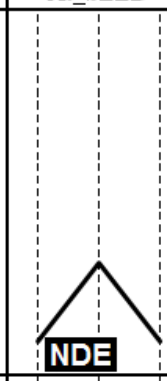
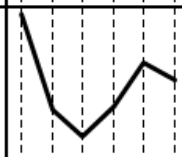
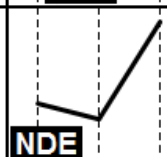
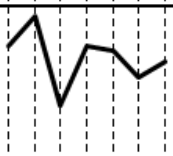
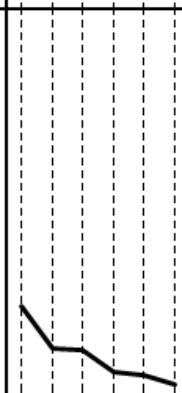
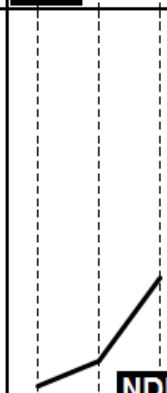
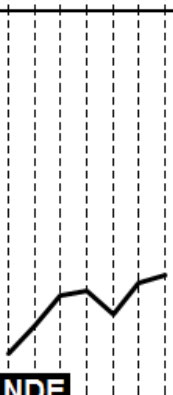
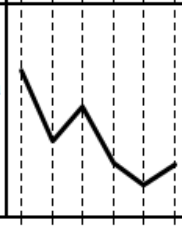
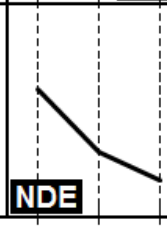
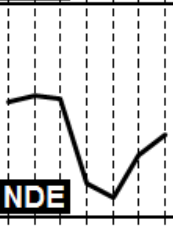
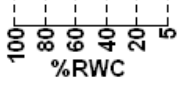
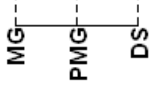
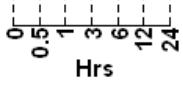
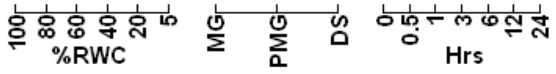
Ortholog ID	Ortholog description	Expression Xh_DSC	Expression At_SEED	Expression At_OSM	Additional annotation information
At3g19580	zinc-finger protein 2				BP: abscisic acid mediated signaling pathway BP: embryo development ending in seed dormancy BP: negative regulation of transcription, DNA-dependent BP: response to auxin stimulus BP: response to chitin BP: response to water deprivation BP: response to wounding BP: signal transduction
At5g47390	myb-like transcription factor family protein				BP: response to abscisic acid stimulus BP: response to gibberellin stimulus BP: response to salt stress
At1g28370	ERF domain protein 11				BP: hyperosmotic salinity response BP: MAPK cascade BP: negative regulation of defense response BP: negative regulation of programmed cell death BP: protein targeting to membrane BP: response to abscisic acid stimulus BP: response to cold BP: response to water deprivation
At2g47900	tubby like protein 3				BP: cellular response to osmotic stress BP: post-translational protein modification BP: response to fungus BP: response to hydrogen peroxide BP: response to salt stress
					

Table 4.17. (continued)

Ortholog ID	Ortholog description	Expression Xh_DSC	Expression At_SEED	Expression At_OSM	Additional annotation information
At3g56400	WRKY DNA-binding protein 70		NDE	NDE	BP: detection of biotic stimulus BP: MAPK cascade BP: negative regulation of defense response BP: negative regulation of leaf senescence BP: negative regulation of programmed cell death BP: negative regulation of transcription, DNA-dependent BP: photosynthesis, light reaction BP: protein targeting to membrane BP: regulation of protein dephosphorylation BP: response to cold BP: response to hypoxia BP: response to other organism
At2g38470	WRKY DNA-binding protein 33	NDE		NDE	BP: detection of biotic stimulus BP: MAPK cascade BP: negative regulation of defense response BP: negative regulation of programmed cell death BP: positive regulation of autophagy BP: protein targeting to membrane BP: response to abiotic stress
At3g62420	basic region/leucine zipper motif 53	NDE			BP: positive regulation of seed maturation BP: protein targeting to vacuole BP: vesicle-mediated transport MF: protein heterodimerization activity
At4g12620	origin of replication complex 1B	NDE	NDE	NDE	BP: DNA replication BP: double fertilization forming a zygote and endosperm MF: double-stranded methylated DNA binding



Xh_DSC (desiccation response in *X. humilis* leaves); At_SEED (*A. thaliana* seed development); At_OSM (osmotic stress response in *A. thaliana* shoots); NDE (non-differentially expressed). BP: Biological process; MF: Molecular function; CC: Cellular component. Additional annotation information was obtained from the *A. thaliana* GO Slim annotation file (version 2013/08/20) accessible at: ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene_Ontology/ATH_GO_GOSLIM.txt

Chapter 5
Conclusion

The current study was set out to analyze to a set of 3105 *X. humilis* cDNAs that were printed on a 'boutique' microarray slide. The aims of this study are (1) to investigate the changes in mRNA transcript abundance in leaf tissue at six different stages of water loss (100%, 80%, 60%, 40%, 20% and 5% RWC) during a desiccation treatment; (2) to identify different temporal classes of genes, as well as functional classes that are activated or repressed during desiccation; (3) to compare the results to the publicly available results obtained from other microarray studies that measured the transitional changes in gene expression in the vegetative tissues of desiccation sensitive plants such as *A. thaliana* during abiotic stresses, as well as the transitional expression changes in seed during seed development, in order to test the hypotheses that vegetative desiccation tolerance observed in *X. humilis* may have derived from the activation of seed specific traits.

An analysis pipeline involved sequence clustering and sequence annotation was set up to identify and to cluster any cDNA clones that were derived from a same gene, followed by the annotation of the resulted contig sequences. 7312 reads were generated from the sequencing of 3123 cDNA clones, with 3105 being analyzed by the current microarray study and 18 being sequenced previously. These sequences were first clustered by d2 cluster and PHRAP algorithms, then carefully checked and manually curated to result in 1775 contigs that represent unique genes in *X humilis*. The peptide sequences of the 1775 contigs were predicted by prot4EST algorithm, and 1331 of which were significantly functionally annotated in Blast2GO.

A microarray dataset of mRNA transcript abundance values for 1680 unique *X. humilis* genes (with 1268 significantly annotated by Blast2GO) under 6 different % RWC conditions, ranging from 100% to <5% RWC was successfully captured and normalized. A large number of the genes (81%) arrayed on the microarray slide were differentially expressed in desiccating *X. humilis* leaves. Clustering analysis on the 6 RWC stages of desiccation treatment was performed by using Sammon mapping, PCA and DIANA. Three distinct clades are defined: a hydrated clade, an early desiccation clade and a late desiccation clade. The results suggested that a rapid major switch from the hydrated to the desiccated program of gene expression occurs around 80% RWC. The second transition occurs around 40% RWC,

and the transcriptional activities associated with desiccation tolerance in *X. humilis* leaves may have been fully established by the time the leaves reached the stage of 20% RWC, and that no further substantial changes occur in the final stages of desiccation.

Cluster analysis on the 1361 differentially expressed contigs identified 7 clusters of cohorts, which then can be divided into 3 different groups based on their expression patterns: the desiccation down-regulated, the early desiccation up-regulated and the late desiccation up-regulated genes. Functional enrichment analysis revealed that among the earliest down-regulated cluster of genes, genes involved in photosynthesis and transcription regulator activity were over-represented. Whereas in the second cluster of down-regulated genes, genes associated with cellular developmental processes and DNA repair were found enriched. The results suggested that the desiccation response in *X. humilis* is accompanied with significant down-regulation of transcription factors. These include LIL3:1 that plays an essential role in chlorophyll and tocopherol biosynthesis, as well as many transcriptional repressors that negatively regulate ABA and ethylene signaling pathways (i.e. ERF4 and FERONIA). Over-represented histone proteins and unbiquitin conjugating enzyme involved in the ubiquitination of histone protein, suggested that chromatin modification may play as an essential mechanism in the regulation of gene activating or silencing in *X. humilis* desiccation response. Furthermore, a large proportion (68%) of the differentially expressed transcription factors (including 9 repressors) was identified down-regulated during desiccation, and no repressors where found in the up-regulated transcription factors in *X. humilis*. The results implicated that the genes conferring desiccation tolerance in *X. humilis* may be predominantly activated via de-repression than being activated by newly up-regulated transcription activators. Nevertheless, these conclusions were made based on the results of the 93 transcription factors represented in the *X. humilis* boutique array, which are a small portion out of approximately 2000 genes encoding transcription factors in plants (Pérez-Rodríguez *et al.*, 2010). With the use of newer technologies such as RNA-Seq would circumvent this problem, and facilitate a comprehensive analysis of the expression levels of all transcription factors during desiccation in *X. humilis*. However the deep sequencing technology which RNA-Seq is based on was not available when this study was initiated. There are many advantages in using RNA-Seq to analyze changes in a transcriptome, including a more precise measurement of transcript levels of all expressed genes, as well as a greater sensitivity in the detection of lowly expressed transcripts (Wang *et al.*, 2009b). The identification and analysis of a larger number of transcription factors in *X. humilis* by RNA-

Seq would aid in the confirmation of the hypothesis developed in this thesis, that vegetative desiccation tolerance is induced via the release from the repression of desiccation associated genes.

Functional enrichment analysis among the earliest up-regulated cluster of genes revealed that genes encoding ribosomal proteins were over-represented. In addition, genes associated with oil body synthesis and stabilization (i.e. caleosins) and storage protein transport (i.e. vacuolar protein sorting protein) in seed, were identified to be enriched in the second early up-regulated cluster of genes. The results suggested that during early phase of desiccation, activation of the desiccation response is accompanied by a switch in the proteome of *X. humilis*, involving the turnover of proteins, and the simultaneous synthesis of proteins required for protection or storage. The synthesis of lipid bodies may also take place, which may be transported to and stored in the vacuoles with the storage proteins synthesized. They fill the vacuole and serve as a mechanical support during loss of cellular water, and may also serve as one of the stored energy sources needed for the restoration of desiccated leaves upon rehydration, or the stored intermediate products readily to be utilized to resume photosynthesis in the rehydrating leaves of *X. humilis*.

Genes encoding LEA proteins and chlorophyll synthesis (i.e. PORA and GUN4) were identified over-represented in the first late up-regulated cluster of genes. The mRNA transcripts of these genes were found abundantly stored in the desiccated leaves, and may be immediately translated in absence of *de novo* transcription upon rehydration. The results suggested that the restoration in processes related to chlorophyll biosynthesis and photosynthesis may be an immediate primary focus when water becomes available again in *X. humilis* leaves. The resumption of processes related to cellular development may be accompanied by protection from these late up-regulated LEA proteins.

High throughput transcript analysis has been carried out in the resurrection plants *C. plantagineum* (Rodriguez *et al.*, 2010) and H. rhodopensis (Gechev *et al.*, 2013). Although functional classification of genes up- or down-regulated identified in the hydrated, dehydrated (42% to 80% RWC), desiccated and rehydrated leaves, as well as links to seed transcriptome were discussed, very little was addressed on the difference in the regulation of vegetative desiccation, seed maturation and abiotic stress response. An analysis comparing the expression patterns of a set of 772 orthologues during desiccation in *X. humilis*, and in

late seed development and osmotic stress in *A. thaliana*, was carried out. The analysis revealed significant overlaps on the up- or down-regulated orthologues identified in the 3 response profiles. Expression and induction of seed specific LEAs and antioxidants were found evident. Furthermore, the two transcription repressors known to negative regulate the expression of seed genes (PICKLE and SWINGER (PRC2)) were shown to be down-regulated in leaves of *X. humilis* during desiccation and in *A. thaliana* seeds towards the end of maturation phase, but were shown to be constantly expressed in *A. thaliana* shoots in response to osmotic stress.

In conclusion, this study has successfully identified different temporal classes of genes, as well as functional classes that are activated or repressed during desiccation in the leaves of *X. humilis*. In addition, the activation of seed specific traits in the vegetative desiccation tolerance observed in *X. humilis* was shown to be evident, and such activation may arise from the ability to down-regulation of repressors of seed specific genes in the leaves of *X. humilis* upon water loss. Such ability appeared to be absent in *A. thaliana* shoot during osmotic stress response, as no significant changes in expression of PICKLE and SWINGER were observed. Desiccation tolerance is normally lost upon germination when ABA level decreases and gibberellin (GA) level increases. In which the genes involved in seed maturation and desiccation tolerance are repressed by the up-regulated PICKLE and PRC2. However, the tolerance can only be restored within a limited time frame, usually during the very early phase of germination depending on the protrusion of the radicle length. The ability to reactivate the seed specific mechanisms by down-regulating the expression of their repressors observed in *X. humilis* suggested that *X. humilis* may have evolved to extend, or to maintain that desiccation restoration time frame observed during early seed germination programme, throughout its life cycle. Furthermore, the different temporal classes of contigs identified in *X. humilis* during desiccation, may provide as an important resource for promoter analysis. Identification of existing or novel CREs within these cohorts of genes may aid in the elucidation of regulatory pathways involved in vegetative desiccation tolerance. Lastly, a more detailed analysis of all expressed mRNA transcripts in *X. humilis* leaves, roots and seeds during the desiccation and rehydration process could be carried out in the future with the more sensitive and less time consuming RNA-Seq technology. Together, these results may provide a more complete and comprehensive insights to the regulations of mechanisms involved in the desiccation tolerance in *X. humilis*.

References

Aalen RB (1999) Peroxiredoxin antioxidants in seed physiology. Seed Science Research 9: 285-295.

Aaronson JS, Eckman B, Blevins RA, Borkowski JA, Myerson J, Imran S and Elliston KO (1996) Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data. Genome Research 6: 829-845.

Abd-El Baki GK, Siefritz F, Man H-M, Weiner H, Kaldenodd R and Kaiser WM (2001) Nitrate reductase in *Zea mays* L. under salinity. Plant, Cell and Environment 23: 515-521.

Abe H, Yamaguchi-Shinozaki K, Urao T, Iwasaki T, Hosokawa D and Shinozaki K (1997) Role of Arabidopsis MYC and MYB homologs in drought- and abscisic acid-regulated gene expression. The Plant Cell 9: 1859-1868.

Abe S, Kurashima A, Yokohama Y and Tanaka J (2001) The cellular ability of desiccation tolerance in Japanese intertidal seaweeds. Botanica Marina 44: 125-131.

Abe H, Urao T, Ito T, Seki M, Shinozaki K and Yamaguchi-Shinozaki K (2003) Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling.

Adams RP, Kendall E and Kartha KK (1990) Comparison of free sugars in growing and desiccated plants of *Selaginella lepidophylla*. Biochemical Systematics and Ecology 18: 107–110.

Adhikari ND, Froehlich JE, Strand DD, Buck SM, Kramer DM and Larkin RM (2011) GUN4-porphyrin complexes bind the ChlH/GUN5 subunit of Mg-Chelatase and promote chlorophyll biosynthesis in Arabidopsis. The Plant Cell 23: 1449-1467.

Ahmad P, Jaleel CA, Salem MA, Nabi G and Sharma S (2010) Roles of enzymatic and nonenzymatic antioxidants in plants during abiotic stress. Critical Reviews in Biotechnology 30: 161-175.

Aichinger E, Villar CB, Farrona S, Reyes JC, Hennig L and Köhler C (2009) CHD3 proteins and polycomb group proteins antagonistically determine cell identity in Arabidopsis. PLoS Genetics 5: e1000605.

Alamillo JM and Bartels D (1996) Light and stage of development influence the expression of desiccation-induced genes in the resurrection plant *Craterostigma plantagineum*. Plant, Cell and Environment 19: 300–310.

Alamillo J, Roncarati R, Heino P, Velasco R, Nelson D, Elster R, Brenacchia G, Furini A, Schwall G, Salamini F and Bartels D (1995) Molecular analysis of desiccation tolerance in barley embryos and in the resurrection plant *Craterostigma plantagineum*. Agronomie 2: 161-167.

Albini FM, Murelli C, Patritti G, Rovati M, Zienna P and Finzi PV (1994) Low-molecular weight substances from the resurrection plant *Sporobolus stapfianus*. The International Journal of Plant Biochemistry 37: 137-142.

Allen RD (1995) Dissection of Oxidative Stress Tolerance Using Transgenic Plants. Plant Physiology 107: 1049–1054.

Alpert P (2005) The limits and frontiers of desiccation-tolerant life. Integrative and Comparative Biology 45: 685-695.

Alpert P (2006) Constraints of tolerance: why are desiccation-tolerant organisms so small or rare? The Journal of Experimental Biology 209: 1575-1584.

Alpert P and Oliver MJ (2002) Drying without dying. In: Desiccation and survival in plants: Drying without dying, edited by Black M and Prichard HW, pp. 1-43. CAB International, Wallingford, UK.

Al-Shahrour F, Díaz-Uriarte R and Dopazo J (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 20: 578-580.

Apel K and Hirt H (2004) Reactive oxygen species: metabolism, oxidative stress, and signal transduction. Annual Review of Plant Biology 55: 373-399.

Asada K and Takahashi M (1987) Production and scavenging of active oxygen in chloroplasts. In: Photoinhibition, edited by Kyle DJ, Osmond CB and Arntzen CJ, pp. 227-287. Elsevier, Amsterdam.

Asada K (1994) Production and action of active oxygen species in photosynthetic tissues. In: Causes of Photooxidative Stress and Amelioration of Defense Systems in Plants, edited by Foyer CH and Mullineaux PM, pp. 77-104. CRC Press, Boca Raton, FL.

Augusti A, Scartazza A, Navari-Izzo F, Sgherri CLM, Stevanovic B and Brugnoli E (2001) Photosystem II photochemical efficiency, zeaxanthin and antioxidant contents in the poikilohydric Ramonda serbica during dehydration and rehydration. Photosynthesis Research 67: 79-88.

Babu M (2004) Introduction to microarray data analysis. In: Computational Genomics: Theory and Application, edited by Grant RP, pp. 225-249. Horizon Bioscience.

Bailly C (2004) Active oxygen species and antioxidants in seed biology. Seed Science Research 14: 93-107.

Baker J, Steele C and Dure L (1988) Sequence and characterization of 6 LEA proteins and their genes from cotton. Plant Molecular Biology 11: 277-291.

Baker SS, Wilhelm KS and Thomashow MF (1994) The 5'-region of *Arabidopsis thaliana* cor15a has cis-acting elements that confer cold-, drought- and ABA-regulated gene expression. Plant Mol Biol 24: 701-713.

Balsamo RA, Vander Willigen C, and Farrant JM (2006) Relating leaf tensile properties to drought tolerance for selected species of *Eragrostis*. Ann Bot 97: 985-991.

Barder GD and Enright AJ (2005) Intermolecular interactions and biological pathways. In Bioinformatics: A practical guide to the analysis of genes and proteins, Third edition, edited by Baxevanis AD and Ouellette BF, pp 253-291. John Wiley & Sons, Inc. Hoboken, New Jersey.

Barkla BJ and Pantoja O (2011) Plasma membrane and abiotic stress. In: The plant plasma membrane. Edited by Murphy AS, Peer W and Schulz B, pp. 457-470. Springer Berlin Heidelberg.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A and Soboleva A (2011) NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic Acids Research 39: D1005-D1010.

Bartels D and Hussain SS (2011) Resurrection plants: Physiology and molecular biology. In: Plant desiccation tolerance, Ecological studies 215, edited by Lüttge U, Beck E and Bartels D, pp. 339-364. Springer-Verlag, Berlin Heidelberg.

Bartels D and Salamini F (2001) Desiccation tolerance in the resurrection plant *Craterostigma plantagineum*: A contribution to the study of drought tolerance at the molecular level. Plant Physiology 127: 1346-1353.

Bartels D and Sunkar R (2005) Drought and salt tolerance in plants. Critical Reviews in Plant Sciences 24: 23-58.

Bartels D, Hanke C, Schneider K, Michel D and Salamini F (1992) A desiccation-related Elip-like gene from the resurrection plant *Craterostigma plantagineum* is regulated by light and ABA. The EMBO Journal 11: 2771-2778.

Bartels D, Phillips J and Chandler J (2007) Desiccation tolerance: Gene expression, pathways, and regulation of gene expression. In Plant desiccation tolerance, edited by Jenks MA and Wood AJ, pp. 115-148. Blackwell Publishing, Iowa.

Baud S, Boutin J-P, Miquel M, Lepiniec L and Rochat C (2002) An integrated overview of seed development in *Arabidopsis thaliana* ecotype WS. Plant Physiology and Biochemistry 40: 151-160.

Beckett RP, Csintalan Z and Tuba Z (2000) ABA treatment increases both the desiccation tolerance of photosynthesis, and nonphotochemical quenching in the moss *Atrichum undulatum*. Plant Ecology 151: 65-71.

Beissbarth T and Speed TP (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 20: 1464-1465.

Belin C, and Lopez-Molina L (2008) Arabidopsis seed germination responses to osmotic stress involve the chromatin modifier PICKLE. Plant Signaling and Behaviour 3: 478-479.

Benjamini Y and Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society 57: 289-300.

Berjak P (2006) Unifying perspectives of some mechanisms basic to desiccation tolerance across life forms. Seed Science Research 16: 1-15.

Berjak P, Farrant JM and Pammenter NW (2007) Seed desiccation tolerance mechanisms. In: Plant desiccation tolerance, edited by Jenks MA and Wood AJ, pp. 51-90. Blackwell Publishing, Ames, Iowa.

Bernacchia G and Furini A (2004) Biochemical and molecular responses to water stress in resurrection plants. Physiologia Plantarum 121: 175-181.

Bewley JD and Black M (1994) Seed. Physiology of Development and Germination, 2nd editon. Plenum Press. New York

Bewley JD and Krochko JE (1982) Desiccation tolerance. In Encyclopedia of Plant Physiologyogy 12B, Physiological Ecology II, edited by Lange OL, Nobel PS, Osmond CB and Ziegler H, pp. 325-378. Springer-Verlag, Berlin.

Bewley JD (1979) Physiological aspects of desiccation tolerance. Annual Review of Plant Physiology 30: 195-238.

Bianchi G, Gamba A, Murelli C, Salamini F and Bartels D (1991) Novel carbohydrate metabolism in the resurrection plant *Craterostigma plantagineum*. The Plant Journal 1: 355-359.

Bianchi G, Gamba A, Limiroli R, Pozzi N, Elster R, Salamini F and Bartels D (1993) The unusual sugar composition in leaves of the resurrection plant *Myrothamnus flabellifolia*. Physiologia Plantarun 87: 223-226.

Bies-Ethève N, Gaubier-Comella P, Debures A, Lasserre E, Jobet E, Raynal M, Cooke R and Delseny M (2008) Inventory, evolution and expression profiling diversity of the LEA (late embryogenesis abundant) protein gene family in *Arabidopsis thaliana*. Plant Molecular Biology 67: 107-124.

Billi D and Potts M (2002) Life and death of dried prokaryotes. Research in Microbiology 153: 7-12.

Blomstedt CK, Gianello RD, Gaff DF, Hamill JD and Neale AD (1998) Differential gene expression in desiccation-tolerant and desiccation-sensitive tissue of the resurrection grass, *Sporobolus stapfianus*. Australian Journal of Plant Physiology 25: 937-946.

Bockel C, Salamini F and Bartels D (1998) Isolation and characterization of genes expressed during early events of the dehydration process in the resurrection plant *Craterostigma plantagineum*. Journal of Plant Physiology 152: 158-166.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S and Schneider M (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research 31: 365-370.

Bolstad BM, Irizarray RA, Astrand M and Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185-193.

Borrell A, Cruz Cutanda M, Lumbreras V, Pujal J, Goday A, Culianez-Macia FA and Pages M (2002) *Arabidopsis thaliana* Atrab28: a nuclear targeted protein related to germination and toxic cation tolerance. Plant Molecular Biology 50:249-259.

Boucher V, Buitink J, Lin X, Boudet J, Hoekstra FA, Hundertmark M, Renard D and Leprince O (2010) MtPM25 is an atypical hydrophobic late embryogenesis-abundant protein that dissociates cold and desiccation-aggregated proteins. Plant, Cell and Environment 33: 418-430.

Bouyer D, Roudier F, Heese M, Andersen ED, Gey D, Nowack MK, Goodrich J, Renou JP, Grini PE, Colot V and Schnittger A (2011) Polycomb repressive complex 2 controls the embryo-to-seedling phase transition. PLoS Genetics 7: e1002014.

Boyer JS (1982) Plant productivity and the environment. Science 218: 443-448.

Bray EA (1993) Molecular responses to water deficit. Plant Physiology 103: 1035-1040.

Bray EA (1994) Alterations in gene expression in response to water deficit. In: Stress-induced gene expression in plants, edited by Basra AS, pp. 1-23. Harwood Academic, Newark, NJ.

Bray EA (1997) Plant responses to water deficit. Trends in Plant Science 2: 48-54.

Braybrook SA, Stone SL, Park S, Bui AQ, Le BH, Fischer RL, Goldberg RB and Harada JJ (2006) Genes directly regulated by LEAFY COTYLEDON2 provide insight into the control of embryo maturation and somatic embryogenesis. Proceedings of the Nationl Academy of Sciences USA 103: 3468-3473.

Broin M, Cuiné S, Eymery F and Rey P (2002) The plastidic 2-cysteine peroxiredoxin is a target for a thioredoxin involved in the protection of the photosynthetic apparatus against oxidative damage. The Plant Cell 14: 1417-1432.

Buitink J and Leprince O (2004) Glass formation in plant anhydrobiotes: survival in the dry state. Cryobiology 48: 215-228.

Buitink J, Hoekstra FA and Leprince O (2002) Biochemistry and biophysics of tolerance systems. In Desiccation and survival in plants: Drying without dying, edited by Black M and Pritchard HW, pp. 293-318. CABI publishing, New York.

Burke J, Davison D and Hide W (1999) d2_cluster: a validated method for clustering EST and full-length cDNAsequences. Genome Research 9: 1135-1142.

Busk PK and Pages M (1998) Regulation of abscisic acid-induced transcription. Plant Molecular Biology 37: 425-435.

Butte A (2002) The use and analysis of microarray data. Nature Reviews Drug Discovery 1: 951-960

Cadman CS, Toorop PE, Hilhorst HW and Finch-Savage WE (2006) Gene expression profiles of Arabidopsis Cvi seeds during dormancy cycling indicate a common underlying dormancy control mechanism. The Plant Journal 46: 805-822.

Calonje M (2014) PRC1 Marks the Difference in Plant PcG Repression. Molecular Plant 7: 459-471.

Calza S and Pawitan Y (2010) Normalization of gene-expression microarray data. Methods in Molecular Biology 673: 37-52.

Carles C, Bies-Etheve N, Aspart L, Leon-Kloosterziel KM, Koornneef M, Echeverria M and Delseny M (2002) Regulation of *Arabidopsis thaliana* Em genes: role of ABI5. The Plant Journal 30: 373-383.

Catalá R, Medina J and Salinas J (2011) Integration of low temperature and light signaling during cold acclimation response in Arabidopsis. Proceedings of the National Academy of Sciences USA 108: 16475-16480.

Causier B, Ashworth M, Guo W and Davies B (2012) The TOPLESS interactome: a framework for gene repression in Arabidopsis. Plant Physiology 158: 423-438.

Causton HC, Quakenbush J and Brazma A (2003) A Beginner's Guide Microarray Gene Expression Data Analysis. Blackwell Publishing. Oxford.

Chen JC and Tzen JT (2001) An in vitro system to examine the effective phospholipids and structural domain for protein targeting to seed oil bodies. Plant and Cell Physiology 42: 1245-1252.

Chen W, Provart NJ, Glazebrook J, Katagiri F, Chang H, Eulgem T, Mauch F, Luan S, Zou G, Whitham SA, Budworth PR, Tao Y, Xie Z, Chen X, Lam S, Kreps JA, Harper JF, Si-Ammour A, Mauch-Mani B, Heinlein M, Kobayashi K, Hohn T, Dangl JL (2002) Expression profile matrix of Arabidopsis transcription factor genes suggests their putative functions in response to environmental stresses. The Plant Cell 14: 559-574.

Chen D, Molitor A, Liu C, Shen WH (2010) The Arabidopsis PRC1-like ring-finger proteins are necessary for repression of embryonic traits during vegetative growth. Cell Research 20: 1332-1344.

Choi H, Hong JH, Ha J, Kang JY and Kim SY (2000) ABFs, a family of ABA-responsive elements binding factors. Journal of Biological Chemistry 275: 1723-1730.

Christoffels A, van Gelder A, Greyling G, Miller R, Hide T and Hide W (2001) STACK: Sequence Tag Alignment and Consensus Knowledgebase. Nucleic Acids Research 29: 234-238.

Ciarmiello LF, Woodrow P, Fuggi A, Pontecorvo G and Carillo P (2011) Plant Genes for Abiotic Stress. In: Abiotic stress in plants - Mechanisms and adaptations, edited by Shanker A and Venkateswarlu. InTech. Available from: http://www.intechopen.com/books/abiotic-stress-in-plants-mechanisms-and-adaptations/plant-genes-for-abiotic-stress.

Close TJ (1996) Dehydrins: Emergence of a biochemical role of a family of plant dehydration proteins. Physiologia Plantarum 97: 795-803.

Close TJ (1997) Dehydrins: a commonality in the response of plants to dehydration and low temperature. Physiologia Plantarum 100: 291-296.

Collada C, Gomez L, Casado R and Aragoncillo C (1997) Purification and in vitro chaperone activity of a class I small heat-shock protein abundant in recalcitrant chestnut seeds. Plant Physiology 115: 71-77.

Collett H, Butowt R, Smith J, Farrant J and Illing N (2003) Photosynthetic genes are differentially transcribed during the dehydration-rehydration cycle in the resurrection plant, *Xerophyta humilis*. Journal of Experimental Botany 54: 2593-2595.

Collett H, Shen A, Gardner M, Farrant JM, Denby KJ and Illing N (2004) Towards transcript profiling of desiccation tolerance in *Xerophyta humilis*: Construction of a normalized 11k *X. humilis* cDNA set and microarray expression analysis of 424 cDNAs in response to dehyration. Physiologia Plantarum 122: 39–53.

Collins CH.and Clegg JS (2004) A small heat-shock protein, p26, from the crustacean *Artemia* protects mammalian cells (Cos-1) from oxidative damage. Cell Biology International 28: 449-455.

Conesa A and Götz S (2008) Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. International Journal of Plant Genomics 2008: 1-13.

Conesa A, Götz S, Garcia-Gomez JM, Terol J, Talon M and Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676.

Cox GW, Beaudet MP, Agnew JY and Ruth JL (2004) Possible sources of dye-related signal correlation bias in two-color DNA microarray assays. Analytical Biochemistry 331: 243-254.

Crowe LM, Womersley C, Crowe JH, Reid D, Appel L and Rudolph A (1986). Prevention of fusion and leakage in freeze-dried liposomes by carbohydrates. Interactions of sugars with membranes. Biochimica *et* Biophysica Acta 861: 131-140.

Crowe JH, Crowe LM, Carpenter JF and Aurell Wistrom C (1987) Stabilization of dry phospholipid bilayers and proteins by sugars. The Biochemical Journal 242: 1-10.

Crowe JH, Crowe LM, Carpenter JF, Rudolph AS, Wistrom CA, Spargo BJ and Anchordoguy TJ (1988) Interactions of sugars with membranes. Biochimica et Biophysica Acta 974: 367-84.

Crowe JH, Hoekstra FA and Crowe LM (1992) Anhydrobiosis. Annual Review of Physiology 54: 579-599.

Crowe JH, Crowe LM, Wolkers WF, Oliver AD, Ma X, Auh J-H, Tang M, Zhu S, Norris J and Tablin F (2005) Stabilization of dry mammalian cells: lessons from nature. Integrative and Comparative Biology 45: 810 -820.

Cui X and Churchill GA (2003) Statistical tests for differential expression in cDNA microarray experiments. Genome Biology 4: 210.1-210.9.

Cuming AC (1999) LEA proteins. In: Seed Proteins, edited by Casey R and Shewry RR, pp. 753-780. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Cushman JC and Oliver MJ (2011) Understanding vegetative desiccation tolerance using integrated functional genomics approaches within a comparative evolutionary framework. In: Plant desiccation tolerance, Ecological studies 215, edited by Lüttge U, Beck E and Bartels D, pp. 307-338. Springer-Verlag, Berlin Heidelberg.

Dace H, Sherwin HW, Illing N and Farrant JM (1998) Use of metabolic inhibitors to elucidate mechanisms of recovery from desiccation stress in the resurrection plant *Xerophyta humilis*. Plant Growth Regulation 24: 171-177.

Datta S and Datta S (2005) Empirical bayes screening of many p-values with applications to microarray studies. Bioinformatics 21: 1987-1994

de Haan JR, Wehrens R, Bauerschmidt S, Piek E, van Schaik RC and Buydens LM (2007) Interpretation of ANOVA models for microarray data using PCA. Bioinformatics 23: 184-190.

Dean RT, Gieseg S and Davies MJ (1993) Reactive species and their accumulation on radical-damaged proteins. Trends in Biochemical Sciences 18: 437-441.

Delmer DP (2005) Agriculture in the developing world: connecting innovations in plant research to downstream applications. Proceedings of the National Academy of Sciences USA 102: 15739-15746.

Deng X, Phillips J, Meijer AH, Salamini F and Bartels D (2002) Characterization of five novel dehydration-responsive homeodomain leucine zipper genes from the resurrection plant *Craterostigma plantagineum*. Plant Molecular Biology 49: 601-610.

Deng X, Hu ZA, Wang HX, Wen XG and Kuang TY (2003) A comparison of photosynthetic apparatus of the detached leaves of the resurrection plant *Boea hygrometrica* with its non-tolerant relative *Chirita heterotrichia* in response to dehydration and rehydration. Plant Science 165: 851-861.

Deshaies RJ and Joazeiro CA (2009) RING domain E3 ubiquitin ligases. Annual Review of Biochemistry 78: 399-434.

Dickie JB and Prichard HW (2002) Systematic and evolutionary aspects of desiccation tolerance in seeds. In Desiccation and survival in plants: Drying without dying, edited by Black M and Prichard HW, pp. 239-259. CAB International, Wallingford, UK.

Dickinson CD, Evans RP and Nielsen NC (1988) RY repeats are conserved in the 5'-flanking regions of legume seed-protein genes. Nucleic Acids Research 16: 371.

Dietz KJ, Horling F, König J and Baier M (2002) The function of the chloroplast 2-cysteine peroxiredoxin in peroxide detoxification and its regulation. Journal of Experimental Botany 53: 1321-1329.

Dimmer E, Berardini TZ, Barrell D and Camon E (2007) Methods for gene ontology annotation. Methods in Molecular Biology 406: 495-520.

Dinakar C, Djilianov D and Bartels D (2012) Photosynthesis in desiccation tolerant plants: Energy metabolism and antioxidative stress defense. Plant Science 182: 29-41.

Ditzer A and Bartels D (2006) Identification of a dehydration and ABA-responsive promoter regulon and isolation of corresponding DNA binding proteins for the group 4 LEA gene CpC2 from *C. plantagineum*. Plant Molecular Biology 61: 643-663.

Dizdaroglu M (1994) Chemical determination of oxidative DNA damage by gas chromatography-mass spectrometry. Methods in Enzymology 234: 3-16.

Do JH and Choi DK (2007) Clustering approaches to identifying gene expression patterns from microarray data. Molecules and Cells 25: 1.

Domany E (2003) Cluster analysis of gene expression data. Journal of Statistical Physics 110: 1117-1139.

Dopazo J (2009) Formulating and testing hypotheses in functional genomics. Artificial Intelligence in Medicine 45: 97-107.

Drãghici S (2003) Data analysis tools for microarrays. Chapman & Hall / CRC Press, London.

Drennan PM, Smith MT, Goldsworth D, Van Staden J (1993) The occurrence of trehalose in the leaves of the desiccation tolerant angiosperm *Myrothamnus flabellifolia* Welw. Journal of Plant Physiology 142: 493-496.

Dure L III, Crouch M, Harada J, Ho T.-HD, Mundy J, Quatrano R, Thomas T and Sung ZR (1989) Common amino acid sequence domains among the LEA proteins of higher plants. Plant Molecular Biology 12: 475-486.

Dure L (1993) Structural motifs in LEA proteins. In: Plant Responses to Cellular Dehydration during Environmental Stress, edited by Close TJ and Bray EA, pp.91-103. American Society of Plant Physiologists, Rockville, MD.

Edgar R, Domrachev M and Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research 30: 207-210.

Edward H Ip (2007) General linear models. In Methods in molecular biology 404, Topics in biostatistics, edited by Ambrosius WT, pp. 189-211. Humana Press Inc., Totowa, NJ.

Efron B, Tibshirani R, Storey JD and Tusher V (2001) Empirical Bayes analysis of a microarray experiment. Journal of American Statistical Association 96: 1151-1169.

Eisen MB, Spellamn PT, Brown PO and Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences USA95: 14863-14868.

Elo LL, Filén S, Lahesmaa R and Aittokallio T (2008) Reproducibility-optimized test statistic for ranking genes in microarray studies. IEEE/ACM Trans Comput Biol Bioinform 5: 423-431.

Elo LL, Hiissa J, Tuimala J, Kallio A, Korpelainen E and Aittokallio T (2009) Optimized detection of differential expression in global profiling experiments: case studies in clinical transcriptomic and quantitative proteomic datasets. Brief Bioinform 10: 547-555.

Espindola AD, Gomes DS, Panek AD and Eleutherio ECA (2003) The role of glutathione in yeast dehydration tolerance. Cryobiology 47: 236-241.

Ewing B and Green P (1998) Base-calling of automated sequencer traces using PHRED. II. Error probabilities. Genome Research 8: 186-194.

Ewing B, Hillier L, Wendl MC and Green P (1998) Base-calling of automated sequencer traces using PHRED. I. Accuracy assessment. Genome Research 8: 175-185.

Ezcurra I, Ellerstrom M, Wycliffe P, Stalberg K and Rask L (1999) Interaction between composite elements in the napA promoter: both the B-box ABA-responsive complex and the RY/G complex are necessary for seed-specific expression. Plant Molecular Biology 40: 699-709.

Ezcurra I, Wycliffe P, Nehlin L, Ellerstrom M, and Rask L (2000) Transactivation of the *Brassica napus* napin promoter by ABI3 requires interaction of the conserved B2 and B3 domains of ABI3 with different cis-elements: B2 mediates activation through an ABRE, whereas B3 interacts with an RY/G-box. The Plant Journal 24: 57-66.

Fait A, Angelovici R, Less H, Ohad I, Urbanczyk-Wochniak E, Fernie AR and Galili G (2006) Arabidopsis seed development and germination is associated with temporally distinct metabolic switches. Plant Physiology 142: 839-854.

Farnsworth E (2004) Hormones and shifting ecology throughout plant development. Ecology 85: 5-15.

Farrant JM, Cooper K, Kruger LA and Sherwin HW (1999) The effect of rate of drying on three different resurrection plants. Annals of Botany 84: 371-379.

Farrant JM, Van der Willigen C, Loffell DA, Bartsch S and Whittaker A (2003) An investigation into the role of light during desiccation of three angiosperm resurrection plants. Plant, Cell and Environment 26: 1275-1286.

Farrant JM (2000) A comparison of patterns of desiccation tolerance among three angiosperm resurrection plant species. Plant Ecology 151: 29-39.

Farrant JM (2007) Mechanisms of desiccation tolerance in angiosperm resurrection plants. In Plant desiccation tolerance, edited by Jenks MA and Wood AJ, pp. 51-90. Blackwell Publishing, Iowa.

Feldman AJ, Costouros NG, Wang E, Qian M, Marincola FM, Alexander HR and Libutti SK (2002) Advantages of mRNA amplification for microarray analysis. Biothechniques 33: 906-914.

Finkelstein RR and Lynch TJ (2000) The Arabidopsis abscisic acid response gene ABI5 encodes a basic leucine zipper transcription factor. The Plant Cell 12: 599-609.

Fisher KM (2008) Bayesian reconstruction of ancestral expression of the LEA gene families reveals propagule-derived desiccation tolerance in resurrection plants. American Journal of Botany 95: 506-515.

Focks N and Benning C (1998) wrinkled1: A novel, low-seed-oil mutant of Arabidopsis with a deficiency in the seed-specific regulation of carbohydrate metabolism. Plant Physiology 118: 91-101.

Fowler SG and Thomashow MF (2002) Arabidopsis transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway. The Plant Cell 14: 1675-1690.

Frank W, Phillips J, Salamini F and Bartels D (1998) Two dehydration-inducible transcripts from the resurrection plant *Craterostigma plantagineum* encode interacting homeodomain-leucine zipper proteins. The Plant Journal 15: 413-421.

Frank W, Munnik T, Kerkmann K, Salamini F and Bartels D (2000) Water deficit triggers phospholipase D activity in the resurrection plant *Craterostigma plantagineum*. The Plant Cell 12: 111-124.

Fresno C, Llera AS, Girotti MR, Valacco MP, López JA, Podhajcer OL, Balzarini MG, Prada F and Fernández EA (2012). The multi-reference contrast method: Facilitating set enrichment analysis. Computers in Biology and Medicine 42: 188-194.

Fridlyand J and Dudoit S (2001) Application of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method. Technical Report 600, Statistics Department, University of California: 217

Fujita M, Fujita Y, Maruyama K, Seki M, Hiratsu K, Ohme-Takagi M, Tran LSP, Yamaguchi-Shinozaki K and Shinozaki K (2004) A dehydration-induced NAC protein, RD26, is involved in a novel ABA-dependent stress-signaling pathway. The Journal 39: 863-876.

Fukunishi Y, Hayashizaki Y (2001) Amino acid translation program for full-length cDNA sequences with frameshift errors. Physiological Genomics 5: 81-87.

Furini A, Koncz C, Salamini F and Bartels D (1997) High level transcription of a member of a repeated gene family confers dehydration tolerance to callus tissue of *Craterostigma plantagineum*. The EMBO Journal 16: 3599-3608.

Gaff DF and Oliver M (2013) The evolution of desiccation tolerance in angiosperm plants: a rare yet common phenomenon. Functional Plant Biology 40: 315–328.

Gaff DF (1971) Desiccation tolerant plants in Southern Africa. Science174: 1033-1034.

Gaff DF (1989) Responses of desiccation tolerant "resurrection" plants to water stress. In Structural and functional responses to environmental stresses, edited by Kreeb KH, Richter H and Hinckley TM, pp. 264-311. SPB Academic Publishing, The Hague.

Gaff DF (1997) Mechanisms of desiccation tolerance in resurrection vascular plants. In Mechanisms of environmental stress resistance in plants, edited by Basra AS and Basra RK, pp. 43-58. Harwood Academic, London.

Galau GA, Hughes DW and Dure L 3rd (1986) Abscisic acid induction of cloned cotton late embryogenesis-abundant (Lea) mRNAs. Plant Molecular Biology 7: 155-170.

Galau GA, Wang HY and Hughes DW (1993) Cotton Lea5 and Lea74 encode atypical Late Embryogenesis-Abundant proteins. Plant Physiology 101: 695-696.

Gallardo K, Thompson R and Burstin J (2008) Reserve accumulation in legume seeds. Comptes Rendus Biologies 331: 755-762.

Gechev T and Hille J (2012) Molecular basis of plant stress. Cellular and Molecular Life Sciences 69: 3161 -3163.

Gecheve TS, Dinakar C, Benina M, Toneva V and Bartels D (2012) Molecular mechanisms of desiccation tolerance in resurrection plants. Cellular and Molecular Life Sciences 69: 3175-3186.

Gechev TS, Benina M, Obata T, Tohge T, Sujeeth N, Minkov I, Hille J, Temanni MR, Marriott AS, Bergström E, Thomas-Oates J, Antonio C, Mueller-Roeber B, Schippers JH, Fernie AR and Toneva V

(2013) Molecular mechanisms of desiccation tolerance in the resurrection glacial relic *Haberlea rhodopensis*. Cellular and Molecular Life Sciences 70: 689-709.

Gilmour SJ, Sebolt AM, Salazar MP, Everard JD and Thomashow MF (2000) Over expression of the Arabidopsis CBF3 transcriptional activator minics multiple biochemical changes associated with cold acclimation. Plant Physiology 124: 1854-1865.

Giraudat J, Hauge BM, Valon C, Smalle J, Parcy F and Goodman HM (1992) Isolation of the Arabidopsis ABI3 gene by positional cloning. The Plant Cell 4: 1251-1261.

Girke T, Todd J, Ruuska S, White J, Benning C and Ohlrogge J (2000) Microarray analysis of developing Arabidopsis seeds. Plant Physiology 124: 1570–1581.

Goldstein DR, Ghosh D and Conlon EM (2002) Statistical issues in the clustering of gene expression data. Statistica Sinica 12: 219-240.

González FJ and Vizcaíno JA (2011) EST analysis pipeline: Use of distributed computing resources. Methods in Molecular Biology 722: 103-120.

Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J and Conesa1 A (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Research 36: 3420-3435.

Götz S, Arnold R, Sebastián-León P, Martín-Rodríguez S, Tischler P, Jehl MA, Dopazo J, Rattei T and Conesa A (2011) B2G-FAR, a species-centered GO annotation repository.Bioinformatics 27: 919-924.

Goyal K, Walton LJ, Browne JA, Burnell AM and Tunnacliffe A (2005) Molecular anhydrobiology: identifying molecules implicated in invertebrate anhydrobiosis. Integrative and Comparative Biology 45: 702-709.

Grene R, Vasuez-Robinet C and Bohnert HJ (2011) Molecular biology and physiological genomics of dehydration stress. In: Plant desiccation tolerance, Ecological studies 215, edited by Lüttge U, Beck E and Bartels D, pp. 255-287. Springer-Verlag, Berlin Heidelberg.

Groth D, Lehrach H, and Hennig S (2004) GOblet: a platform for gene ontology annotation of anonymous sequence data. Nucleic Acids Research 32: W313-W317.

Guiltinan MJ, Marcotte WR and Quatrano RS (1990) A plant leucine zipper protein recognizes an abscisic acid responsive element. Science 25: 267-271.

Guo J, Wu J, Ji Q, Wang C, Luo L, Yuan Y, Wang Y and Wang J (2008) Genome-wide analysis of heat shock transcription factor families in rice and Arabidopsis. Journal of Genetics and Genomics 35: 105-118.

Gutierrez L, Van Wuytswinkel O, Castelain M and Bellini C (2007) Combined networks regulating seed maturation. Trends in Plant Science 12: 294-300.

Hansen NF, Thomas PJ and Bouffard GG (2005) Sequence assembly and finishing methods. In Bioinformatics: A practical guide to the analysis of genes and proteins, Third edition, edited by Baxevanis AD and Ouellette BFF. John Wiley & Sons, Inc., Hoboken, New Jersey.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG,

Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T and White R (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research 32(Database issue): D258-D261.

Hartung W, Schiller P, and Dietz KJ (1998) Physiology of poikilohydric plants. In: Progress in Botany, edited by Behnke HD, Esser K, Kadereit JW, Lüttge U and Runge M. pp 299-327. Springer-Verlag Berlin Heidelberg

Haslekås C, Stacy RA, Nygaard V, Culiáñez-Macià FA and Aalen RB (1998) The expression of a peroxiredoxin antioxidant gene, AtPer1, in *Arabidopsis thaliana* is seed-specific and related to dormancy. Plant Molecular Biology 36: 833-845.

Haslekås C, Grini PE, Nordgard SH, Thorstensen T, Viken MK, Nygaard V and Aalen RB. ABI3 mediates expression of the peroxiredoxin antioxidant AtPER1 gene and induction by oxidative stress. Plant Molecular Biology 53: 313-326.

Helm KW, LaFayette PR, Nagao RT, Key JL and Vierling E (1993) Localization of small heat shock proteins to the higher plant endomembrane system. Molecular and Cell Biology 13: 238-247.

Helm KW, Schmeits J and Vierling E (1995) An endomembrane-localized small heat-shock protein from *Arabidopsis thaliana*. Plant Physiology 107: 287-288.

Henderson JT, Li HC, Rider SD, Mordhorst AP, Romero-Severson J, Cheng JC, Robey J, Sung ZR, de Vries SC and Ogas J (2004) PICKLE acts throughout the plant to repress expression of embryonic traits and may play a role in gibberellin-dependent responses. Plant Physiology 134: 995-1005.

Hendry GAF (1993) Oxygen, free radical processes and seed longevity. Seed Science Research 3: 141-153.

Heyer LJ, Kruglyak S and Yooseph S (1999) Exploring expression data: Identification and analysis of coexpressed genes. Genome Research 9: 1106-1115.

Heyndrickx KS and Vandepoele K (2012) Systematic identification of functional plant modules through the integration of complementary data sources. Plant Physiology 159: 884-901.

Hinch DK and Thalhammer A (2012) LEA proteins: IDPs with versatile functions in cellular dehydration tolerance. Biochemical Society Transactions 45: 1000-1003.

Hirai M, Yamakawa R, Nishio J, Yamaji T, Kashino Y, Koike H and Satoh K (2004) Deactivation of photosynthetic activities is triggered by loss of a small amount of water in a desiccation-tolerant cyanobacterium, *Nostoc commune*. Plant Cell Physiol. 45: 872-878.

Hoekstra FA and Golovina EA (1999) Membrane behavior during dehydration: implications for desiccation tolerance. Russian Journal of Plant Physiology 46: 295-306.

Hoekstra FA and Golovina EA (2002) The role of amphiphiles. Comparative Biochemistry and Physiology 131A: 527-533.

Hoekstra FA, Golovian EA and Buitink J (2001) Mechanisms of plant desiccation tolerance. Trends in Plant Science 6: 431-438.

Hoekstra FA (2005) Differential longevities in desiccated anhydrobiotic plant systems. Integrative and Comparative Biology 45: 725-733.

Holdsworth MJ, Bentsink L and Soppe WJJ (2008) Molecular networks regulating Arabidopsis seed maturation, after-ripening, dormancy and germination. New Phytologist 179: 33-54.

Hua J, Sakai H, Nourizadeh S, Chen QG, Bleecker AB, Ecker JR and Meyerowitz EM (1998) EIN4 and ERS2 are members of the putative ethylene receptor gene family in Arabidopsis. The Plant Cell 10: 1321-1332.

Huang X and Madan A (1999) CAP3: a DNA sequence assembly program. Genome Research 9: 868-877.

Huang DW, Sherman BT and Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research 37: 1-13.

Hughes DW and Galau GA (1989) Temporally modular gene expression during cotyledon development. Genes and Development 3: 358-369.

Hundertmark M and Hincha DK (2008) LEA (late embryogenesis abundant) proteins and their encoding genes in *Arabidopsis thaliana*. BMC Genomics 9: 118.

Hutin C, Nussaume L, Moise N, Moya I, Kloppstech K and Havaux M (2003) Early light-induced proteins protect Arabidopsis from photooxidative stress. Proceedings of the National Academy of Sciences USA 100: 4921-4926.

Iljin WS (1957) Drought resistance in plants and physiological processes. Annual Review of Plant Physiology 3: 341-363.

Illing N, Denby K, Collett H, Shen A and Farrant JM (2005) The signature of seeds in resurrection plants: a molecular and physiological comparison of desiccation tolerance in seeds and vegetative tissues. Integrative and Comparative Biology 45: 771–787

Imelfort M (2009) Sequence Comparison Tools. In: Applied Bioinformatics. Edited by Edwards D, Hansen D and Stajich J. Springer.

Ingle RA, Schmidt UG, Farrant JM, Thomson JA and Mundree SG (2007) Proteomic analysis of leaf proteins during dehydration of the resurrection plant *Xerophyta viscosa*. Plant, Cell and Environment 30: 435-446.

Ingle RA, Collett H, Cooper K, Takahashi Y, Farrant JM and Illing N (2008) Chloroplast biogenesis during rehydration of the resurrection plant *Xerophyta humilis*: parallels to the etioplast-chloroplast transition. Plant, Cell and Environment 31: 1813-1824.

Ingram J and Bartels D (1996) The molecular basis of dehydration tolerance in plants. Ann Rev Plant Physiology Plant Molecular Biology 47: 377-403.

Ingram J, Chandler JW, Gallagher L, Salamini F and Bartels D (1997) Analysis of cDNA clones encoding sucrose-phosphate synthase in relation to sugar interconversions associated with dehydration in the resurrection plant *Craterostigma plantagineum* Hochst. Plant Physiology 115: 113-121.

Iseli C, Jongeneel CV and Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proceedings International Conference on Intelligent Systems for Molecular Biology 1999: 138-148.

Iturriaga G, Leyns L, Villegas A, Gharaibeh R, Salamini F and Bartels D (1996) A family of novel myb-related genes from the resurrection plant *Craterostigma plantagineum* are specifically expressed in callus and roots in response to ABA or desiccation. Plant Molecular Biology 32: 707-716.

Iturriaga G, Cushman M and Cushman J (2006). An EST catalogue from the resurrection plant *Selaginella lepidophylla* reveals abiotic stress-adaptive genes. Plant Biology 170: 1173–1184.

Iwasaki T, Yamaguchi-Shinozaki K and Shinozaki K (1995) Identification of a cis-regulatory region of a gene in *Arabidopsis thaliana* whose induction by dehydration is mediated by abscisic acid and requires protein synthesis. Molecular Genetics and Genomics 247: 391-398.

Jensen M, Chakir S and Feige GB (1999) Osmotic and atmospheric dehydration effects in the lichens *Hypogymnia physodes*, *Lobaria pulmonaria*, and *Peltigera aphthosa*: an in vivo study of the chlorophyll fluorescence induction. Photosynthetica 37: 393-404.

Jiang C, Iu B and Singh J (1996) Requirement of a CCGAC cis-acting element for cold induction of the BN115 gene from winter *Brassica napus*. Plant Molecular Biology 30: 679-684.

Jones L and McQueen-Mason S (2004) A role for expansins in dehydration and rehydration of the resurrection plant *Craterostigma plantagineum*. FEBS Letter. 559, 61-65.

Jones SI, Gonzalez DO, Lila O and Vodkin LO (2010) Flux of transcript patterns during soybean seed development. BMC Genomics 11: 136.

Kagaya Y, Okuda R, Ban A, Toyoshima R, Tsutsumida K, Usui H, Yamamoto A and Hattori T (2005) Indirect ABA-dependent regulation of seed storage protein genes by FUSCA3 transcription factor in Arabidopsis. Plant and Cell Physiology 46: 300-311.

Kaufman L and Rousseeuw PJ (1990) Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons Ltd., New York.

Kaushik JK and Bhat R (2003) Why is trehalose an exceptional protein stabilizer? An analysis of the thermal stability of proteins in the presence of the compatible osmolyte trehalose. Journal of Biological Chemistry 278: 26458-26465.

Kermode AR and Finch-Savage BE (2002) Desiccation sensitivity in orthodox and recalcitrant seeds in relation to development. In Desiccation and survival in plants: Drying without dying, edited by Black M and Pritchard HW, pp. 149-184. CABI publishing, New York.

Kerr MK, Martin M and Churchill GA (2000) Analysis of variance for gene expression microarray data. Journal of Computational Biology 7: 819-837.

Khan S, Situ G, Decker K and Schmidt CJ (2003) GoFigure: automated gene ontology annotation. Bioinformatics 19: 2484-2485.

Khatri P and Drăghici S (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 21: 3587-3895.

Khojasteh M, Lam WL, Ward RK and MacAulay C (2005) A stepwise framework for the normalization of array CGH data. BMC Bioinformatics 6: 274.

Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, D'Angelo C, Bornberg-Bauer E, Kudla J and Harter K (2007) The AtGenExpress global stress expression data set: protocols,

evaluation and model data analysis of UV-B light, drought and cold stress responses. The Plant Journal 50: 347-363.

Kim H, Chen J and Yu X (2007) Ubiquitin-Binding Protein RAP80 Mediates BRCA1-Dependent DNA Damage Response. Science 316: 1202-1205.

Kirch HH, Nair A and Bartels D. (2001) Novel ABA- and dehydration-inducible aldehyde dehydrogenase genes isolated from the resurrection plant *Craterostigma plantagineum* and *Arabidopsis thaliana*. The Plant Journal 28: 555-567.

Kirkpatrick DS, Hathaway NA, Hanna J, Elsasser S, Rush J, Finley D, King RW, and Gygi SP (2006) Quantitative analysis of in vitro ubiquitinated cyclin B1 reveals complex chain topology. Nature Cell Biology 8: 700–710.

Kodaira KS, Qin F, Tran LS, Maruyama K, Kidokoro S, Fujita Y, Shinozaki K and Yamaguchi-Shinozaki K (2011) Arabidopsis Cys2/His2 zinc-finger proteins AZF1 and AZF2 negatively regulate abscisic acid-repressive and auxin-inducible genes under abiotic stress conditions. Plant Physiology 157: 742-756.

Kohonen T (1995) Self organizing maps. Springer, Berlin.

Koornneef M, Leon-Kloosterzeil K, Schwartz SH and Zeevarrt JAD (1998) The genetic and molecular dissection of abscisic acid biosynthesis and signal transduction in Arabidopsis. Plant Physiology and Biochemistry 36: 83-89.

Kotchoni SO, Kuhns C, Ditzer A, Kirch HH and Bartels D (2006) Over-expression of different aldehyde dehydrogenase genes in *Arabidopsis thaliana* confers tolerance to abiotic stress and protects plants against lipid peroxidation and oxidative stress. Plant, Cell and Environment 29: 1033-1048.

Kovacs D, Kalmar E, Torok Z and Tompa P (2008) Chaperone activity of ERD10 and ERD14, two disordered stress-related plant proteins. Plant Physiology 147: 381-390.

Kranner I and Birtic S (2005) A modulating role for antioxidants in desiccation tolerance. Integrative and Comparative Biology 45: 734-740.

Kranner I and Lutzoni F (1999) Evolutionary consequences of transition to a lichen symbiotic state and physiological adaptation to oxidative damage associated with poikilohydry. In: Plant response to environmental stress: From phytohormones to genome reorganization, edited by Lerner HR, pp. 591-628. Dekker, New York.

Kranner I, Beckett RP, Wornik S, Zorn M and Pfeifhofer HW (2002) Revival of a resurrection plant correlates with its antioxidant status. The Plant Journal 31: 13-24.

LaFayette PR, Nagao RT, O'Grady K, Vierling E and Key JL (1996) Molecular characterization of cDNAs encoding low-molecular-weight heat shock proteins of soybean. Plant Molecular Biology 30: 159–169.

Larkin RM, Alonso JM, Ecker JR and Chory J (2003) GUN4, a regulator of chlorophyll synthesis and intracellular signaling. Science 299: 902-906.

Lata C, Yadav A and Prasad M (2011) Role of Plant transcription factors in abiotic stress tolerance. In: Abiotic stress response in plants: Physiological, biochemical and genetic perspectives, edited by Shanker A and Venkateswarlu. InTech. Available from: http://www.intechopen.com/books/abiotic-stress-response-in-plants-physiological-biochemical-and-geneticperspectives/role-of-plant-transcription-factors-in-abiotic-stress-tolerance.

Le BH, Cheng C, Bui AQ, Wagmaister JA, Henry KF, Pelletier J, Kwong L, Belmonte M, Kirkbride R, Horvath S, Drews GN, Fischer RL, Okamuro JK, Harada JJ and Goldberg RB (2010) Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. Proceedings of the National Academy of Sciences USA 107: 8063-8070.

Lee ML, Kuo FC, Whitmore GA and Sklar J (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proceedings of the National Academy of Sciences USA 97: 9834–9839.

Lehmann L, Ferrari R, Vashisht AA, Wohlschlegel JA, Kurdistani SK and Carey M (2012) Polycomb repressive complex 1 (PRC1) disassembles RNA polymerase II preinitiation complexes. The Journal of Biological Chemistry 287: 35784-35794.

Lenne C and Douce R (1994) A Low Molecular Mass Heat-Shock Protein Is Localized to Higher Plant Mitochondria. Plant Physiology 105: 1255-1261.

Lenne C, Block MA, Garin J and Douce R (1995) Sequence and expression of the mRNA encoding HSP22, the mitochondrial small heat-shock protein in pea leaves. The Biochemical Journal 311: 805-813.

Leprince O and Buitink J (2007) The glassy state in dry seeds and pollen. In Plant desiccation tolerance, edited by Jenks MA and Wood AJ, pp. 193-214. Blackwell Publishing, Iowa.

Leprince O and Buitink J (2010) Desiccation tolerance: From genomics to the field. Plant Science 179: 554-564.

Leprince O, Harren JM, Buitink J, Alberda M and Hoekstra FA (2000) Metabolic dysfunction and unabated respirartion precede the loss of membrane integrity during dehydration of germinating radicals. Plant Physiologyogy 122: 597-608.

Lersten NR, Czlapinski AR, Curtis JD, Freckmann R and Horner HT (2006) Oil bodies in leaf mesophyll cells of angiosperms: overview and a selected survey. American Journal of Botany 93: 1731-1739.

Li Y, Li T, Liu S, Qiu M, Han Z, Jiang Z, Li R, Ying K, Xie Y and Mao Y (2004) Systematic comparison of the fidelity of aRNA, mRNA and T-RNA on gene expression profiling using cDNA microarray. Journal of Biothechnology 107: 19-28.

Li HC, Chuang K, Henderson JT, Rider SD Jr, Bai Y, Zhang H, Fountain M, Gerber J and Ogas J (2005) PICKLE acts during germination to repress expression of embryonic traits. The Plant Journal 44: 1010-1022.

Lim YS, Cha MK, Kim HK, Uhm TB, Park JW, Kim K and Kim IH (1993) Removals of hydrogen peroxide and hydroxyl radical by thiol-specific antioxidant protein as a possible role in vivo. Biochemical and Biophysical Research Communications 192: 273-280.

Lin C and Thomashow MF (1992) DNA sequence analysis of a complementary DNA for cold-regulated Arabidopsis gene cor15 and characterization of the COR15 polypeptide. Plant Physiology 99:519-525.

Liu Y, Koornneef M and Soppe WJ (2007) The absence of histone H2B monoubiquitination in the Arabidopsis hub1 (rdo4) mutant reveals a role for chromatin remodeling in seed dormancy. The Plant Cell 19: 433-444.

Liu JH, Peng T and Dai W (2013) Critical cis-Acting Elements and Interacting Transcription Factors: Key Players Associated with Abiotic Stress Responses in Plants. Plant Molecular Biology Reporter: 1-15.

Loewe RP and Nelson PJ (2011) Microarray bioinformatics. Methods in Molecular Biology 671: 295-320.

Loiacono FV and De Tullio MC (2012) Why we should stop inferring simple correlations between antioxidants and plant stress resistance: towards the antioxidomic era. OMICS: A Journal of Integrative Biology 16: 160-167.

Long JA, Ohno C, Smith ZR and Meyerowitz EM (2006) TOPLESS regulates apical embryonic fate in Arabidopsis. Science 312: 1520-1523.

Lopez-Molina L and Chua NH (2000) A null mutation in a bZIP factor confers ABA-insensitivity in *Arabidopsis thaliana*. Plant and Cell Physiology 41: 541-547.

Lotan T, Ohto M, Yee KM, West MA, Lo R, Kwong RW, Yamagishi K, Fischer RL, Goldberg RB and Harada JJ (1998) Arabidopsis LEAFY COTYLEDON1 is sufficient to induce embryo development in vegetative cells. Cell 93: 1195-1205.

Luerssen H, Kirik V, Herrmann P and Misera S (1998) FUSCA3 encodes a protein with a conserved VP1/AB13-like B3 domain which is of functional importance for the regulation of seed maturation in *Arabidopsis thaliana*. The Plant Journal 15: 755-764.

Lyzenga WJ and Stone SL (2012) Abiotic stress tolerance mediated by protein ubiquitination. Journal of Experimental Botany 63: 599-616.

Maere S, Heymans K and Kuiper M (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21: 3448-3449.

Mandadi KK, Misra A, Ren S and McKnight TD (2009) BT2, a BTB protein, mediates multiple responses to nutrients, stresses, and hormones in Arabidopsis. Plant Physiology 150: 1930-1939.

Manfre AJ, Lanni LM and Marcotte WR Jr (2006) The Arabidopsis group 1 late embryogenesis abundant protein ATEM6 is required for nomal seed development. Plant Physiology 140: 140-149.

Marcotte WR Jr, Russell SH and Quatrano RS (1989) Abscisic acid-responsive sequences from the em gene of wheat. The Plant Cell 1: 969-976.

Martin D, Brun C, Remy E, Mouren P, Thieffry D and Jacq B (2004) GOToolBox: functional analysis of gene datasets based on Gene Ontology. Genome Biology 5: R101.

Maruyama K, Sakuma Y, Kasuga M, Ito Y, Seki M, Goda H, Shimada Y, Yoshida S, Shinozaki K and Yamaguchi-Shinozaki K (2004). Identification of cold-inducible downstream genes of the Arabidopsis DREB1A/CBF3 transcriptional factor using two microarray systems. The Plant Journal 38: 982-993.

McCubbin WD, Kay CM and Lane B (1985) Hydrodynamic and optical properties of the wheat germ Em protein. Canadian Journal of Biochemistry and Cell Biology 63: 803-811.

McLachlan GJ, Bean RW and Ng SK (2008) Clustering. Methods in Molecular Biology 453: 423-439.
Miller RT, Christoffels AG, Gopalakrishnan C, Burke J, Ptitsyn AA, Broveak TR and Hide WA (1999) A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. Genome Research 9: 1143-1155.

Mittler R (2002) Oxidative stress, antioxidants and stress tolerance. Trends in Plant Science 7: 405-410.

Mohapatra SK and Krishnan A (2011) Microarray data analysis. Methods in Molecular Biology 678: 27-43.

Mönke G, Altschmied L, Tewes A, Reidt W, Mock HP, Bäumlein H and Conrad U (2004) Seed-specific transcription factors ABI3 and FUS3: molecular interaction with DNA. Planta 219: 158-166.

Moore JP, Hearshaw M, Ravencroft N, Lindsey GG, Farrant JM and Brandt WF (2007) Desiccation-induced ultrastructural and biochemical changes in the leaves of the resurrection plants *Myrothamnus flabellifolia*. Australian Journal of Botany 55: 482-491.

Moore JP, Vicré-Gibouin M, Farrant JM and Driouich A (2008) Adaptations of higher plant cell walls to water loss: drought *vs* desiccation. Physiologia Plantarum 134: 237-245.

Morey JS, Ryan JC and Van Dolah FM (2006) Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. Biological Procedure Online 8: 175-193.

Morgan PW and Drew MC (1997) Ethylene and plant responses to stress. Physiologia Plantarum 100: 620-630.

Mowla SB, Thomson JA, Farrant JM and Mundree SG (2002) A novel stress-inducible antioxidant enzyme identified from the resurrection plant *Xerophyta viscosa* Baker. Planta 215: 716-726.

Mtwisha L, Farrant J, Brandt W and Lindsey G (2006) Protection mechanisms against water deficit stress: Desiccation tolerance in seeds as a study case. In Drought adaptations in cereals, edited by Ribout J-M, pp. 531-549. Haworth Press, New York.

Mukhopadhyay D and Riezman H (2004) Proteasome-Independent Functions of Ubiquitin in Endocytosis and Signaling. Science 315: 201-205.

Mundree SG and Farrant JM (2002) Some Physiological and Molecular Insights into the Mechanisms of Desiccation Tolerance in the Resurrection Plant *Xerophyta viscosa* Baker. In: Plant Tolerance to Abiotic Stresses in Agriculture: Role of Genetic Engineering, edited by Cherry JH, Locy RD and Rychter A, pp. 201-222. Springer Netherlands.

Mundree SG, Whittaker A, Thomson JA and Farrant JM (2000) An Aldose Reductase Homologue from the resurrection plant *Xerophyta vicosa* Baker. Planta 211: 693-700.

Mundy J, Yamaguchi-Shinozaki K and Chua NH (1990) Nuclear proteins bind conserved elements in the abscisic acid-responsive promoter of a rice rab gene. Proceedings of the National Academy of Sciences USA 87: 1406-1410.

Murphy DJ, Hernendez-Pinzon I, Patel K, Hope RG and McLauchlan J (2000) New insights into the mechanisms of lipid-body biogenesis in plants and other organisms. Biochemical Society Transactions 28: 710-711.

Murphy DJ, Hernandez-Pinzon I and Patel K (2001) Role of lipid bodies and lipid-body proteins in seeds and other tissues. Journal of Plant Physiology 158: 471-478.

Myhre S, Tveit H, Mollestad T and Laegreid A (2006). Additional gene ontology structure for improved biological reasoning. Bioinformatics 22: 2020–2027.

Nagaraj SH, Gasser RB and Ranganathan (2007) A hitchhiker's guide to expressed sequence tag (EST) analysis. Briefings in Bioinformatics 8: 6- 21.

Nakabayashi K, Okamoto M, Koshiba T, Kamiya Y and Nambara E (2005) Genome-wide profiling of stored mRNA in *Arabidopsis thaliana* seed germination: epigenetic and genetic regulation of transcription in seed. The Plant Journal 41: 697-709.

Nakashima K and Yamaguchi-Shinozaki K (2013) ABA signaling in stress-response and seed development. Plant Cell Reports 32: 959-970.

Ndima TB, Farrant JM, Thomson JA and Mundree SG (2001) Molecular characterization of XVT8, a stress-responsive gene from the resurrection plant *Xerophyta viscosa* Baker. Plant Growth Regulation 35: 137-145.

Neumann PM (2008) Coping mechanisms for crop plants in drought-prone environments. Annals of Botany 101: 901-907.

Norwood M, Truesdale MR, Richter A and Scott P (2000) Photosynthetic carbohydrate metabolism in the resurrection plant *Craterostigma plantagineum*. Journal of Experimental Botany 51: 159-165.

Norwood M, Toldi O, Richter A and Scott P (2003) Investigation into the ability of roots of the poikilohydric plant *Craterostigma plantagineum* to survive dehydration stress. Journal of Experimental Botany 54: 2313-2321.

Nugent R and Meila M (2010) An overview of clustering applied to molecular biology. Methods in Molecular Biology 620: 369-404.

Nunes C, O' Hara LE, Primavesi LF, Delatte TL, Schluepmann H, Somsen GW, Silva AB, Fevereiro PS, Wingler A and Paul MJ (2013) The trehalose 6-phosphate/SnRK1 signaling pathway primes growth recovery following relief of sink limitation. Plant Physiology 162: 1720-1732.

Oliver MJ and Bewley JD (1997) Desiccation-tolerance of plant tissues: a mechanistic overview. Horticultural Reviews 18: 171-213.

Oliver MJ, Wood AJ and O'Mahony P (1998) "To dryness and beyond"- preparation for the dried state and rehydration in vegetative desiccation-tolerant plants. Plant Growth Regulation 24: 193-201.

Oliver MJ, Tuba Z and Mishler BD (2000) The evolution of vegetative desiccation tolerance in land plants. Plant Ecology 151: 85-100.

Oliver AE, Hincha DK and Crowe JH (2002) Looking beyond sugars: the role of amphiphilic solutes in preventing adventitious reactions in anhydrobiotes at low water contents. Comparative Biochemistry and Physiology 131A: 515-525.

Oliver MJ, Velten J and Mishler BD (2005) Desiccation tolerance in bryophytes: a reflection the primitive strategy for plant survival in dehydrating habitats? Integrative and Comparative Biology 45: 788-799.

Oliver MJ (1996) Desiccation tolerance in vegetative plant cells. Physiologia Plantarum 97: 779-787.

Oliver MJ (2007) Lessons on dehydration tolerance from desiccation-tolerant plants. In Plant desiccation tolerance, edited by Jenks MA and Wood AJ, pp. 151-192. Blackwell Publishing, Iowa.

Olvera-Carrillo Y, Campos F, Reyes JL, Garciarrubio A and Covarrubias AA (2010) Functional analysis of the group 4 late embryogenesis abundant proteins reveals their relevance in the adaptive response during water deficit in Arabidopsis. Plant Physiology 154: 373-390.

Pabon C, Modrusan Z, Ruvolo MV, Coleman IM, Daniel S, Yue H and Arnold LJ Jr (2001) Optimized T7 amplification system for microarray analysis. Biotechniques 31: 874-879

Paddock T, Lima D, Mason ME, Apel K, Armstrong GA (2012) Arabidopsis light-dependent protochlorophyllide oxidoreductase A (PORA) is essential for normal plant growth and development. Plant Molecular Biology 78: 447-460.

Page GP, Zakharkin SO, Kim K, Mehta T, Chen L and Zhang K (2007) Microarray analysis. In Methods in molecular biology 404, Topics in biostatistics. edited by Ambrosius WT, pp. 409-430. Humana Press Inc., Totowa, NJ.

Pampurova S and Van Dijck P (2014) The desiccation tolerant secrets of *Selaginella lepidophylla*: What we have learned so far? Plant Physiology and Biochemistry 80: 285-290.

Parcy F, Valon C, Raynal M, Gaubier-Comella P, Delseny M and Giraudat J (1994) Regulation of gene expression programs during Arabidopsis seed development: roles of the ABI3 locus and of endogenous abscisic acid. The Plant Cell 6: 1567–1582.

Park S and Harada JJ (2008) Arabidopsis embryogenesis. In Methods in Molecular Biology 427: Plant Embryogenesis, edited by Suárez MF and Bozhkov PV, pp. 3-16. Humana Press, Totowa, NJ.

Parkinson J, Anthony A, Wasmuth J, Schmid R, Hedley A and Blaxter M (2004) PartiGene--constructing partial genomes. Bioinformatics 20: 1398-1404.

Peeva V and Maslenkova L (2004) Thermoluminescence study of photosystem II activity in *Haberlea rhodopensis* and spinach leaves during desiccation. Plant Biology. 6: 319-324.

Pérez-Rodríguez P, Riaño-Pachón DM, Corrêa LG, Rensing SA, Kersten B and Mueller-Roeber B (2010) PlnTFDB: updated content and new features of the plant transcription factor database. Nucleic Acids Research 38: 822-827.

Phillips JR, Oliver MJ and Bartels D (2002) Molecular genetics of desiccation and tolerant systems. In Desiccation and survival in plants: Drying without dying, edited by Black M and Pritchard HW, pp. 319-341. CABI publishing, New York.

Phillips JR, Fisher E, Baron M, van den Dries N, Facchinelli F, Kutzer M, Rahmanzadeh R, Remus D and Bartels D (2008) *Lindernia brevidens*: a novel desiccation-tolerant vascular plants, endemic to ancient tropical rainforests. The Plant Journal 54: 938-948.

Piatkowski D, Schneider K, Salamini F and Bartels D (1990) Characterization of Five Abscisic Acid-Responsive cDNA Clones Isolated from the Desiccation-Tolerant Plant *Craterostigma plantagineum* and Their Relationship to Other Water-Stress Genes. Plant Physiolgy 94: 1682-1688.

Pickart CM and Fushman D (2004) Polyubiquitin chains: polymeric protein signals. Current Opinion in Chem Biology 8: 610-616.

Polacek DC, Passerini AG, Shi C, Francesco NM, Manduchi E, Grant GR, Powell S, Bischof H, Winkler H, Stoeckert CJ Jr and Davies PF (2003) Fidelity of enhanced sensitivity of differential transcription profiles following linear amplification of nanogram amounts of endothelial mRNA. Physiological Genomics 13: 147-156

Pollard KS and van der Laan MJ (2005) Cluster analysis of genomic data. In Bioinformatics and Computational Biology Solutions using R and Bioconductor, edited by Gentleman R, Carey VJ, Huber W, Irizarry RA and Dudoit S, pp. 209-228. Springer, NewYork.

Porembski S and Barthlott W (2000) Granitic and gneissic outcrops (inselbergs) as centres of diversity for desiccation-tolerant vascular plants. Plant Ecology 151: 19-28.

Porembski S (2011) Evolution, diversity, and habitats of poikilohydrous vascular plant. In: Plant desiccation tolerance, Ecological studies 215, edited by Lüttge U, Beck E and Bartels D, pp. 139-156. Springer-Verlag, Berlin Heidelberg.

Potts M, Slaughter SM, Hunneke F-U, Garst JF and Helm RF (2005) Desiccation tolerance of prokaryotes: application of principles to human cells. Integrative and Comparative Biology 45: 800-809.

Potts M (1996) The anhydrobiotic cyanobacterial cell. Physiologia Plantarum 97: 788-794.

Price AH, Taylor A, Ripley SJ, Griffiths A, Trewavas AJ and Knight MR (1994) Oxidative Signals in Tobacco Increase Cytosolic Calcium. The Plant Cell 6: 1301-1310.

Prieto-Dapena P, Castaño R, Almoguera C and Jordano J (2008) The ectopic overexpression of a seed-specific transcription factor, HaHSFA9, confers tolerance to severe dehydration in vegetative organs. The Plant Journal 54: 1004-1014.

Proctor MCF and Pence VC (2002) Vegetative tissues: Bryophytes, vascular resurrection plants and vegetative propagules. In Desiccation and survival in plants: Drying without dying, edited by Black M and Pritchard HW, pp. 207-237. CABI publishing, New York.

Proctor MC, Ligrone R and Duckett JG (2007) Desiccation tolerance in the moss *Polytrichum formosum*: physiological and fine-structural changes during desiccation and recovery. Annals of Botany 99: 75-93.

Quackenbush J (2001) Computational analysis of microarray data. Nature Reviews Genetics 2: 418-427.

Quackenbush J (2005) Using DNA microarrays to assay gene expression. In Bioinformatics: A practical guide to the analysis of gene and proteins, edited by Baxevanis AD and Ouellette BFF. John Wiley & Sons, Inc. Hoboken, New Jersey.

Quartacci MF, Glisic O, Stevanovic B and Navari-Izzo F (2002) Plasma membrane lipids in the resurrection plant *Ramonda serbica* following dehydration and rehydration. Journal of Experimental Botany. 53: 2159-2166.

Quick P, Siegl G, Neuhaus E, Feil R, Stitt M (1989) Short-term water stress leads to a stimulation of sucrose synthesis by activating sucrose-phosphate synthase. Planta 177: 535-546.

Ramanjulu S and Bartels D (2002) Drought- and desiccation-induced modulation of gene expression in plants. Plant, Cell and Environment 25: 141-151.

Ramanjulu S, Veeranjaneyulu K and Sudhakar C (1994) Short-term shifts in nitrogen-metabolism in mulberry morus-albaunder salt shock. Phytochemistry 37: 991-995.

Raychaudhuri S, Stuart JM and Altman RB (2000) Principal components analysis to summarize microarray experiments: Application to sporulation time series. Pacific Symposium on Bioconducting 5: 452-463.

Raz V, Bergervoet J and Koornneef M (2001) Sequential steps for developmental arrest in Arabidopsis seeds. Development 128: 243-252.

Reidt W, Wohlfarth T, Ellerström M, Czihal A, Tewes A, Ezcurra I, Rask L and Bäumlein H (2000) Gene regulation during late embryogenesis: the RY motif of maturation-specific gene promoters is a direct target of the FUS3 gene product. The Plant Journal 21: 401-408.

Reilly C (2009) Statistical genomics. In Statistics in human genetics and moleculr biology. Chapman & Hall/CRC, Boca Raton, Florida. pp 179-191.

Reimers M and Carey VJ (2006) Bioconductor: an open source framework for bioinformatics and computational biology. Methods in Enzymology 411: 119-134.

Reymond P (2001) DNA microarray and plant defence. Plant Physiology and Biocheistry 39: 313-321.
Rivals I, Personnaz L, Taing L and Potier MC (2007) Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 23: 401-407.

Robert HS, Quint A, Brand D, Vivian-Smith A and Offringa R (2009) BTB and TAZ domain scaffold proteins perform a crucial function in Arabidopsis development. The Plant Journal 58: 109-121.

Roberts JK, DeSimone NA, Lingle WL and Dure L 3rd (1993) Cellular concentrations and uniformity of cell-type accumulation of two LEA proteins in cotton embryos. The Plant Cell 5: 769-780.

Rodrigo-Brenni MC, Foster SA and Morgan DO (2010) Catalysis of lysine 48-specific ubiquitin chain assembly by residues in E2 and ubiquitin. Molecular Cell 39: 548-559.

Rodriguez MC, Edsgärd D, Hussain SS, Alquezar D, Rasmussen M, Gilbert T, Nielsen BH, Bartels D and Mundy J (2010) Transcriptomes of the desiccation-tolerant resurrection plant *Craterostigma plantagineum*. The Plant Journal 63: 212-228.

Rohila JS, Jain RK and Wu R (2002) Genetic improvement of Basmati rice for salt and drought tolerance by regulated expression of a barley Hva1 cDNA. Plant Science 163: 525-532.

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

Roxas VP, Smith RK Jr, Allen ER and Allen RD (1997) Overexpression of glutathione S-transferase/glutathione peroxidase enhances the growth of transgenic tobacco seedlings during stress. Nature Biotechnology 15: 988-991.

Rozen S and Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Bioinformatics Methods and Protocols: Methods in Molecular Biology, edited by Krawetz S and Misener S. Humana Press, Totowa, NJ, pp 365-386.

Rubin DB (1991) Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. Biometrics 47: 1213-1234.

Ruuska SA, Girke T, Benning C and Ohlrogge JB (2002) Contrapuntal networks of gene expression during Arabidopsis seed filling. The Plant Cell 14: 1191-1206.

Sadowski M, Suryadinata R, Tan AR, Roesley SN and Sarcevic B (2012) Protein monoubiquitination and polyubiquitination generate structural diversity to control distinct biological processes. IUBMB Life 64: 136-142.

Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, Pico AR, Bader GD and Ideker T (2012) A travel guide to Cytoscape plugins. Nature Methods 9: 1069-1076.

Sakai H, Hua J, Chen QG, Chang C, Medrano LJ, Bleecker AB and Meyerowitz EM (1998) ETR2 is an ETR1-like gene involved in ethylene signaling in Arabidopsis. Proceedings of the National Academy of Sciences USA 95: 5812-5817.

Sakamoto H, Maruyama K, Sakuma Y, Meshi T, Iwabuchi M, Shinozaki K and Yamaguchi-Shinozaki K (2004) Arabidopsis Cys2/His2-type zinc-finger proteins function as transcription repressors under drought, cold, and high-salinity stress conditions. Plant Physiology 136: 2734-2746.

Sammon JW (1969) A non-linear mapping for data structure analysis. IEEE Transactions on Computers 18: 401-409.

Santos-Mendoza M, Dubreucq B, Baud S, Parcy F, Caboche M and Lepiniec L (2008) Deciphering gene regulatory networks that control seed development and maturation in Arabidopsis. The Plant Journal 54: 608-620.

Scharf K-D, Siddique M and Vierling E (2001) The expanding family of *Arabidopsis thaliana* small heat stress proteins and a new family of proteins containing α-crystallin domains (Acd proteins). Cell Stress Chaperones 6: 225-237.

Schena M, Shalon D, Davis RW and Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467-470.

Schmid R and Blaxter M (2009) EST processing: from trace to sequence. Methods in Molecular Biology 533: 189-220.

Schmidt MT, Handschuh L, Zyprych J, Szabelska A, Olejnik-Schmidt AK, Siatkowski I and Figlerowicz (2011) impact of DNA microarray data transformation on gene expression analysis – comparison of two normalization methods. Acta Biochimica Polonica 58: 573-580.

Schuettengruber B, Chourrout D, Vervoort M, Leblanc B and Cavalli G (2007) Genome regulation by polycomb and trithorax proteins. Cell 128: 735-745.

Schwartz YB and Pirrotta V (2008) Polycomb complexes and epigenetic states. Current Opinion in Cell Biology 20: 266-273.

Scott JG and and Berger JO (2010) Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Annals of Statistics 38: 2587-2619.

Scott P (2000) Resurrection plants and the secrets of eternal leaf. Annals of Botany 85: 159-166.

Seki M, Narusaka M, Abe H, Kasuga M, Yamaguchi-Shinozaki K, Carninci P, Hayashizaki Y and Shinozaki K (2001) Monitoring the expression pattern of 1300 Arabidopsis genes under drought and cold stresses by using a full-length cDNA microarray. The Plant Cell 13: 61-72.

Seki M, Narusaka M, Ishida J, Nanjo T, Fujita M, Oono Y, Kamiya A, Nakajima M, Enju A, Sakurai T, Satou M, Akiyama K, Taji T, Yamaguchi-Shinozaki K, Carninci P, Kawai J, Hayashizaki Y, Shinozaki K (2002) Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. The Plant Journal 31: 279-292.

Seki M, Kamei Ayako, Yamaguchi-Shinozaki K and Shinozaki K (2003) Molecular responses to drought, salinity and frost: common and different paths for plant protection. Current Opinion in Biothechnology 14: 194-199.

Seki M, Satou M, Sakurai T, Akiyama K, Iida K, Ishida J, Nakajima M, Enju A, Narusaka M, Fujita M, Oono Y, Kamei A, Yamaguchi-Shinozaki K and Shinozaki K (2004) RIKEN Arabidopsis full-length (RAFL) cDNA and its applications for expression profiling under abiotic stress conditions. Journal of Experimental Botany 55: 213-223.

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

Shao H-B, Chu L-Y, Jaleel CA and Zhao C-X (2008) Water-deficit stress-induced anatomical changes in higher plants. C R Biologies 331: 215-225.

Shen Q, Uknes SJ and Ho TH (1993) Hormone response complex in a novel abscisic acid and cycloheximide-inducible barley gene. Journal of Biological Chemistry 268: 23652-13660.

Sherwin HW and Farrant JM (1996) Differences in rehydration of three desiccation-tolerant angiosperm species. Annals of Botany 78: 703-710.

Sherwin HW, Pammenter NW, February ED, Van der Willigen C and Farrant JM (1998) Xylem hydraulic characteristics, water relations and wood anatomy of the resurrection plant *Myrothamnus flabellifolius* Welw. Annals of Botany 81: 567-575.

Shimada T, Koumoto Y, Li L, Yamazaki M, Kondo M, Nishimura M and Hara-Nishimura I (2006) AtVPS29, a putative component of a retromer complex, is required for the efficieent sorting of seed storage proteins. Plant and Cell Physiology 47: 1187-1194.

Shinozaki K and Yamaguchi-Shinozaki K (2000) Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signaling pathways. Current Opinion in Plant Biology 3:217–223.

Shinozaki K and Yamaguchi-Shinozaki K (2007) Gene networks involved in drought stress response and tolerance. Journal of Experimental Botany 58: 221-227.

Shirkey B, Kovarcik DP, Wright DJ, Wilmoth G, Prickett, TF, Helm RF, Gregory EM and Potts M (2000) Active Fe-containing superoxide dismutase and abundant sodF mRNA in *Nostoc commune* (Cyanobacteria) after years of desiccation. Journal of Bacteriology 182: 189-197.

Simon JA and Kingston RE (2009) Mechanisms of polycomb gene silencing: knowns and unknowns. Nature Reviews. Molecular Cell Biology 10: 697-708.

Simpson SD, Nakashima K, Narusaka Y, Seki M, Shinozaki K and Yamaguhci-Shinozaki K (2003) Two different novel cis-acting elements of erd1, a clpA homologous Arabidopsis gene function in induction by dehydration stress and dark-induced senescence. The Plant Journal 33: 259-270.

Sivamani E, Bahieldin1 A, Wraith JM, Al-Niemi T, Dyer WE, Ho TD and Qu R (2000) Improved biomass productivity and water use efficiency under water deficit conditions in transgenic wheat constitutively expressing the barley HVA1 gene. Plant Science 155: 1-9.

Smirnoff N (1993) The role of active oxygen in the response of plants to water deficit and desiccation. New Phytologist 125: 27-58.

Smirnoff N (1998) Plant resistance to environmental stress. Current Opinion in Biotechnology 9: 214-219.

Smith ZR and Long JA (2010) Control of Arabidopsis apical-basal embryo polarity by antagonistic transcription factors. Nature 464: 423-426.

Smith-Espinoza CJ, Richter A, Salamini F and Bartels D (2003) Dissecting the response to dehydration and salt (NaCl) in the resurrection plant *Craterostigma plantagineum*. Plant, Cell and Environment 26: 1307-1315.

Smith-Espinoza CJ, Phillips JR, Salamini F and Bartels D (2005) Identification of further *Craterostigma plantagineum* cdt mutants affected in abscisic acid mediated desiccation tolerance. Molecular Genetics and Genomics 274: 364-372.

Smyth GK and Speed TP (2003) Normalization of cDNA microarray data. Methods 31: 265-273.

Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 3: Article 3.

Smyth GK (2005) Limma: linear models for microarray data. In Bioinformatics and Computational Biology Solutions using R and Bioconductor, edited by Gentleman R, Carey VJ, Huber W, Irizarry RA and Dudoit S, pp. 397-420. Springer, NewYork.

Soeda Y, Konings MCJM, Vorst O, van Houwelingen AMML, Stoopen GM, Maliepaard CA, Kodde J, Bino RJ, Groot SPC and van der Geest AHM (2005) Gene expression programs during *Brassica oleracea* seed maturation, osmopriming, and germination are indicators of progression of the germination process and the stress tolerance level. Plant Physiology 137: 354-368.

Spellman P, Sherlock G, Zhang M, Iyer V, Anders K, Eisen M, Brown P, Botstein D and Futcher B (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Molecular Biology of the Cell 9: 3273–3297.

Spitzer C, Reyes FC, Buono R, Sliwinski MK, Haas TJ and Otegui MS (2009) The ESCRT-related CHMP1A and B proteins mediate multivesicular body sorting of auxin carriers in Arabidopsis and are required for plant development. The Plant Cell 21: 749-766.

Sreenivasulu N, Sopory SK and Kavi Kishor PB (2007) Deciphering the regulatory mechanisms of abiotic stress tolerance in plants by genomic approaches. Gene 388: 1-13.

Stark LR, Greenwood JL, Brinda JC and Oliver MJ (2013) The desert moss *Pterygoneurum lamellatum* (Pottiaceae) exhibits an inducible ecological strategy of desiccation tolerance: effects of rate of drying on shoot damage and regeneration. American Journal of Botany 100: 1522-1531.

Stekel D (2003) Microarray bioinformatics. Cambridge University Press. Cambridge, UK.

Stockinger EJ, Gilmour SJ and Thomashow MF (1997) *Arabidopsis thaliana* CBF1 encodes an AP2 domain-containing transcriptional activator that binds to the C-repeat/DRE, a cis-acting DNA regulatory element that stimulates transcription in response to low temperature and water deficit. Proceedings of the National Academy of Sciences USA 94: 1035-1040.

Stone SL, Kwong LW, Yee KM, Pelletier J, Lepiniec L, Fischer RL, Goldberg RB and Harada JJ (2001) LEAFY COTYLEDON2 encodes a B3 domain transcription factor that induces embryo development. Proceedings of the National Academy of Sciences USA 98: 11806-11811.

Streeter JG, Lohnes DG and Fioritto RJ (2001) Patterns of pinitol accumulation in soybean plants and relationships to drought tolerance. Plant, Cell and Environment 24: 429-438.

Subramanian AR, Kaufmann M and Morgenstern B (2008) DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. Algorithms for Molecular Biology 3: 6.

Sun W, Bernard C, van de Cotte B, Van Montagu M, and Verbruggen N (2001) At-HSP17.6A, encoding a small heat-shock protein in Arabidopsis, can enhance osmotolerance upon overexpression. The Plant Journal 27: 407-415.

Sun W, van Montagu M and Verbruggen N (2002) Small heat shock proteins and stress tolerance in plants. Biochimica *et* Biophysica Acta 1557: 1-9.

Sunkar R, Bartels D and Kirch HH (2003) Overexpression of a stress-inducible aldehyde dehydrogenase gene from *Arabidopsis thaliana* in transgenic plants improves stress tolerance. The Plant Journal 35: 452-464.

Suzuki M, Kao CY and McCarty DR (1997) The conserved B3 domain of VIVIPAROUS1 has a cooperative DNA binding activity. The Plant Cell 9: 799-807.

Taji T, Ohsumi C, Iuchi S, Seki M, Kasuga M, Kobayashi M, Yamaguchi-Shinozaki K and Shinozaki K (2002) Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance in *Arabidopsis thaliana*. The Plant Journal 29: 417-426.

Takahashi K, Takabayashi A, Tanaka A and Tanaka R (2014) Functional analysis of light-harvesting-like protein 3 (LIL3) and its light-harvesting chlorophyll-binding motif in Arabidopsis. The Journal of Biological Chemistry 289: 987-999.

Takahashi S, Katagiri T, Yamaguchi-Shinozaki K and Shinozaki K (2000) An Arabidopsis Gene Encoding a Ca2+-Binding Protein is Induced by Abscisic Acid during Dehydration. Plant and Cell Physiology 41: 898-903.

Tanaka R, Rothbart M, Oka S, Takabayashi A, Takahashi K, Shibata M, Myouga F, Motohashi R, Shinozaki K, Grimm B and Tanaka A (2010) LIL3, a light-harvesting-like protein, plays an essential role in chlorophyll and tocopherol biosynthesis. Proceedings of National Academy of Sciences of USA 107: 16721-16725.

Tárraga J, Medina I, Carbonell J, Huerta-Cepas J, Minguez P, Alloza E, Al-Shahrour F, Vegas-Azcárate S, Goetz S, Escobar P, Garcia-Garcia F, Conesa A, Montaner D and Dopazo J (2008) GEPAS, a web-based tool for microarray data analysis and interpretation. Nucleic Acids Research 36: w308-w314.

Tavazoie S, Hughes D, Campbell MJ, Cho RJ and Church GM (1999) Systematic determination of genetic network architecture. Nature Genetics 22: 281-285.

Thomashow MF (1999) Plant cold acclimation: freezing tolerance genes and regulatory. Annual Review of Plant Physiology and Plant Molecular Biology 50: 571-599.

Tibshirani R, Walther G and Hastie T (2000) Estimating the number of clusters in a dataset via the gap statistic. Technical report, Department of Statistics, Standford University: 217.

Toldi O, Tuba Z and Scott P (2009) Vegetative desiccation tolerance: is it a goldmine for bioengineering crops? Plant Science 176: 187-199.

Trainor FR and Gladych R (1995) Survival of algae in a desiccated soil: a 35-year study. Phycologia 34: 191-192.

Tran LSP, Nakashima K, Sakuma Y, Simpson SD, Fujita Y, Maruyama K, Fujita M, Seki M, Shinozaki K and Yamaguchi-Shinozaki K (2004) Isolation and functional analysis of Arabidopsis stress-inducible NAC transcription factors that bind to a drought-responsive cis-element in the early responsive to dehydration stress 1 promoter. The Plant Cell 16: 2481-2498.

Tuba Z and Lichtenthaler HK (2011) Ecophysiology of homoiochlorophyllous and poikilochlorophyllous desiccation tolerant plants and vegetations. In: Plant desiccation tolerance, Ecological studies 215, edited by Lüttge U, Beck E and Bartels D, pp. 157-183. Springer-Verlag, Berlin Heidelberg.

Tuba Z, Lichtenthaler HK, Csintalan Z, Nagy Z and Szente K (1996) Loss of chlorophylls, cessation of photosynthetic CO2 assimilation and respirartion in the poikilochlorophyllous plant *Xerophyta scabrida* during desiccation. Physiologia Plantarum 96: 383-388.

Tunnacliffe A and Wise MJ (2007) The continuing conundrum of the LEA proteins. Die Naturwissenschaften 94: 791-812.

Tusher VG, Tibshirani R and Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences USA 98: 5116-5121.

Tweddle JC, Dickie JB, Baskin CC and Baskin JM (2003) Ecological aspects of seed desiccation sensitivity. Journal of Ecology 91: 294-304.

Tzen JT and Huang AH (1992) Surface structure and properties of plant seed oil bodies. The Journal of Cell Biology 117: 327-335.

Uno Y, Furihata T, Abe H, Yoshida R, Shinozaki K and Yamaguchi-Shinozaki K (2000) Arabidopsis basic leucine zipper transcriptional transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions. Proceedings of the National Academy of Sciences USA 97: 11632-11637.

Van der Willigen C, Mundree SG, Pammenter NW and Farrant JM (2003) An ultrastructural study using anhydrous fixation of *Eragrostis nindensis*, a resurrection grass with both desiccation-tolerant and -sensitive tissues. Functional Plant Biology 30: 281-290.

Van der Willigen C, Pammenter NW, Mundre SG and Farrant JM (2004) Mechanical stabilization of desiccated vegetative tissues of resurrection grass *Eragrostis nindensis*: does a TIP 3:1 and/or compartmentalization of subcellular components and metabolites play a role? Journal of Experimental Botany 55, 651-661.

Van Gelder RN, von Xastrow ME, Yool A, Dement DC, Barchas JD and Eberwine JH (1990) Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proceedings of the National Academy of Sciences USA87: 1663-1667

Van Heerden J, Walford S-A, Shen A and Illing N (2007) A framework for the informed normalization of printed microarrays. South African Journal of Science 103: 381-390

Van Zanten M, Tessadori F, Peeters AJM and Fransz P (2013) Environment-Induced Chromatin Reorganisation and Plant Acclimation. In: Epigenetic Memory and Control in Plants, edited by Grafi G and Ohad N, pp. 21-40. Springer-Verlag Berlin Heidelberg.

Vicré M, Farrant JM and Driouich A (2004a) Insights into the cellular mechanisms of desiccation tolerance among angiosperm resurrection plant species. Plant, Cell and Environment 27: 1329–1340.

Vicré M, Lerouxel O, Farrant J, Lerouge P, and Driouich A (2004b) Composition and desiccation-induced alterations of the cell wall in the resurrection plant *Craterostigma wilmsii*. Physiologia Plantarum 120: 229 -239.

Vierling E (1991) The roles of heat shock proteins in plants. Annuual Review of Plant Molecular Biology 42: 579-620.

Villalobos MA, Bartels D and Iturriaga G (2004) Stress tolerance and glucose insensitive phenotypes in Arabidopsis overexpressing the CpMYB10 transcription factor gene. Plant Physiology 135: 309-324.

Vogel JT, Zarka DG, Van Buskirk HA, Fowler SG and Thomashow MF (2005) Roles of the CBF2 and ZAT12 transcription factors in configuring the low temperature transcriptome of Arabidopsis. The Plant Journal 412: 195-211.

Walters C, Farrant JM, Pammenter NW and Berjak P (2002) Desiccation and damage. In: Desiccation and survival in plants: Drying without dying, edited by Black M and Pritchard HW, pp. 263-291. CABI publishing, New York.

Wan X, Mo A, Liu S, Yang L and Li L (2010) Constitutive expression of a peanut ubiquitin-conjugating enzyme gene in Arabidopsis confers improved water-stress tolerance through regulation of stress- responsive gene expression. Journal of Bioscience and Bioengineering 111: 478–484.

Wang JP, Lindsay BG, Leebens-Mack J, Cui L, Wall K, Miller WC and dePamphilis CW (2004a) EST clustering error evaluation and correction. Bioinformatics 17: 2973-2984.

Wang H, Wang L, Erdjument-Bromage H, Vidal M, Tempst P, Jones RS and Zhang Y (2004b) Role of histone H2A ubiquitination in Polycomb silencing. Nature 431: 873-878.

Wang D, Tyson MD, Jackson SS and Yadegari R (2006) Partially redundant functions of two SET-domain polycomb-group proteins in controlling initiation of seed development in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America 103: 13244-13249.

Wang Z, Zhu Y, Wang L, Liu X, Liu Y, Phillips J and Deng X (2009a) A WRKY transcription factor participates in dehydration tolerance in *Boea hygrometrica* by binding to the W-box elements of the galactinol synthase (BhGolS1) promoter. Planta 230: 1155-1166.

Wang Z, Gerstein M and Snyder M (2009b) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews: Genetics 10: 57-63.

Wasmuth JD and Blaxter ML (2004) prot4EST: translating expressed sequence tags from neglected genomes. BMC Bioinformatics 5: 187.

Wasmuth J and Blaxter M (2009) Obtaining accurate translations from expressed sequence tags. Methods in Molecular Biology 533: 221-239.

Wasmuth J and Blaxter M (2009) Obtaining accurate translations from expressed sequence tags. Methods in Molecular Biology 533: 221-239.

Waters ER, Lee GJ and Vierling E (1996) Evolution structure and function of the small heat shock proteins in plants. Journal of Experimental Botany 47: 325-338.

Waters ER (2013) The evolution, function, structure, and expression of the plant sHSPs. Journal of Experimental Botany 64: 391-403.

Weake VM and Workman JL (2008) Histone ubiquitination: triggering gene activity. Molecular Cell 29: 653-663.

Weber H, Borisjuk L and Wobus U (2005) Molecular physiology of legume seed development. Annual Review of Plant Biology 56: 253-279.

Weeraratna AT and Taub DD (2007) Microarray data analysis: An overview of design, methodology, and analysis. In Methods in molecular biology 377 Microarray data analysis: Methods and applications, edited by Korenberg MJ, pp. 1-16. Humana Press Inc., Totowa, NJ.

Wen R, Torres-Acosta JA, Pastushok L, Lai X, Pelzer L, Wang H and Xiao W (2008) Arabidopsis UEV1D promotes Lysine-63-linked polyubiquitination and is involved in DNA damage response. The Plant Cell 20: 213-227.

Wernisch L, Kendall SL, Soneji S, Wietzorrek A, Parish T, Hinds J, Butcher PD, and Stoker NG (2003) Analysis of whole-genome microarray replicates using mixed models. Bioinformatics 19: 53-61.

West MAL and Harada JJ (1993) Embryogenesis in higher plants-an overview. The Plant Cell 5: 1361-1369.

Whittaker A, Bochicchio A, Vazzana C, Lindsey G and Farrant JM (2001) Changes in leaf hexokinase activity and metabolite levels in response to drying in the desiccation-tolerant species Sporobolus stapfianus and *Xerophyta viscosa*. Journal of Experimental Botany 352: 961-969.

Whittaker A, Martinelli T, Bochicchio A, Vazzana C and Farrant JM (2004) Comparison of sucrose metabolism during the rehydration of desiccation-tolerant and desiccation-sensitive leaf material of *Sporobolus stapfianus*. Physiologia Plantarum 122: 11-20.

Wilson C, Caton TM, Buchheim JA, Buchheim MA, Schneegurt MA and Miller RV (2004) DNA-repair potential of *Halomonas* spp. from the Salt Plains Microbial Observatory of Oklahoma. Microbial Ecology. 48: 541-549.

Wingler A (2002) The function of trehalose biosynthesis in plants. Phytochemistry 60: 437-440.

Wise MJ and Tunnacliffe A (2004) POPP the question: what do LEA proteins do? Trends in Plant Science 9: 13-17.

Wise MJ (2003) LEAping to conclusions: A computational reanalysis of late embryogenesis abundant proteins and their possible roles. BMC Bioinformatics 4: 52.

Wit E and McClure J (2004) Statistics for microarrays: Design, analysis and inference. John Wiley & Sons, Ltd, Chichester, West Sussex, UK.

Wohlbach DJ, Quirino BF and Sussman MR (2008) Analysis of the Arabidopsis histidine kinase ATHK1 reveals a connection between vegetative osmotic stress sensing and seed maturation. The Plant Cell 20: 1101-1117.

Wood AJ and Jenks MA (2007) Plant desiccation tolerance: diversity, distribution, and real-world applications. In: Plant desiccation tolerance, edited by Jenks MA and Wood AJ, pp. 3-10. Blackwell Publishing, Iowa.

Wood AJ (2005) Eco-physiological adaptations to limited water environments. In: Plant abiotic stress, edited by Jenks MA and Hasegawa PM, pp. 1-13. Blackwell publishing, Iowa.

Wood AJ (2007) Frontiers in bryological and lichenological research. The nature and distribution of vegetative desiccation tolerance in hornworts, liverworts and mosses. Bryologist 110: 163-177.

Wormit A, Trentmann O, Feifer I, Lohr C, Tjaden J, Meyer S, Schmidt U, Martinoia E and Neuhaus HE (2006) Molecular identification and physiological characterization of a novel monosaccharide transporter from Arabidopsis involved in vacuolar sugar transport. The Plant Cell 18: 3476-3490.

Wu S-H, Ramonell K, Gollub J and Somerville S (2001) Plant gene expression profiling with DNA microarrays. Plant Physiology and Biochemistry 39: 917-926.

Xiang L, Etxeberria E and Van den Ende W (2013) Vacuolar protein sorting mechanisms in plants. The FEBS Journal 280: 979-993.

Xu D, Duan X, Wang B, Hong B, Ho T and Wu R (1996) Expression of a late embryogenesis abundant protein gene, HVA1, from barley confers tolerance to water deficit and salt stress in transgenic rice. Plant Physiology 110: 249-257.

Yamada K, Osakabe Y, Mizoi J, Nakashima K, Fujita Y, Shinozaki K and Yamaguchi-Shinozaki K (2010) Functional analysis of an *Arabidopsis thaliana* abiotic stress-inducible facilitated diffusion transporter for monosaccharides. The Journal of Biological Chemistry 285: 1138-1146.

Yamaguchi S, Kamiya Y and Nambara E (2007) Regulation of ABA and GA levels during seed development and germination in Arabidopsis. In: Seed development, dormancy and germination, Annual plant reviews 27, edited by Bradford K and Nonogaki H, pp. 224-247. Blackwell Publishing Ltd, Oxford, UK.

Yamaguchi-Shinozaki K and Shinozaki K (1992) A novel Arabidopsis DNA binding protein contains the conserved motif of HMG-box proteins. Nucleic Acids Research 20: 6737.

Yamaguchi-Shinozaki K and Shinozaki K (1994) A novel cis-acting element in an Arabidopsis gene is involved in responsiveness to drought, low temperature, or high-salt stress. The Plant Cell 6: 251-264.

Yamaguchi-Shinozaki K and Shinozaki K (2005) Organization of cis-acting regulatory elements in osmotic- and cold-stress-responsive promoters. Trends in Plant Science 10: 88-94.

Yamaguchi-Shinozaki K and Shinozaki K (2006) Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. Annual Review of Plant Biology57: 781-803.

Yang YH, Dudoit S, Luu P and Speed TP (2001) Normalization for cDNA microarray data. Proceedings of SPIE, BIOS 2001, Micorarrays: Optical Technologies and Informatics 4266: 141-152.

Yang YH and Speed T (2002) Design issues for cDNA microarray experiments. Nature Reviews Genetics 3: 579-588.

Yang Z, Tian L, Latoszek-Green M, Brown D and Wu K (2005) Arabidopsis ERF4 is a transcriptional repressor capable of modulating ethylene and abscisic acid responses. Plant Molecular Biology 58: 585-596.

Yang C, Bratzel F, Hohmann N, Koch M, Turck F and Calonje M (2013) VAL- and AtBMI1-mediated H2Aub initiate the switch from embryonic to postgerminative growth in Arabidopsis. Current Biology 23: 1324-1329.

Yona G, Dirks W and Rahman S (2009) Comparing algorithems for clustering of expression data: How to access gene clusters. Computational Systems Biology 541: 479-509.

Yu F, Qian L, Nibau C, Duan Q, Kita D, Levasseur K, Li X, Lu C, Li H, Hou C, Li L, Buchanan BB, Chen L, Cheung AY, Li D and Luan S (2012) FERONIA receptor kinase pathway suppresses abscisic acid signaling in Arabidopsis by activating ABI2 phosphatase. Proceedings of the National Academy of Sciences USA 109: 14693-14698.

Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC and Weinstein JN (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biology 4: R28.

Zhang B, Schmoyer D, Kirov S and Snoddy J (2004) GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. BMC Bioinformatics 5: 16.

Zhang H, Rider SD Jr, Henderson JT, Fountain M, Chuang K, Kandachar V, Simons A, Edenberg HJ, Romero-Severson J, Muir WM and Ogas J (2008) The CHD3 remodeler PICKLE promotes trimethylation of histone H3 lysine 27. The Journal of Biological Chemistry 283: 22637-22648.

Zhang Y, Szustakowski J and Schinke M (2009) Bioinformatics analysis of microarray data. Methods in Molecular Biology 573: 259-284.

Zhou GA, Chang RZ and Qiu LJ (2010) Overexpression of soybean ubiquitin-conjugating enzyme gene GmUBC2 confers enhanced drought and salt tolerance through modulating abiotic stress-responsive gene expression in Arabidopsis. Plant Molecular Biology 72: 357-367.

Zrenner R and Stitt M (1991) Comparison of the effect of rapidly and gradually developing water stress on carbohydrate metabolism in spinach leaves. Plant, Cell and Environment 14: 939-946.

Appendix

A.3.1. R script: Quality assessment of expression data of 1680 contigs identified during desiccation in *X. humilis* leaves

```
setwd("C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\marray")
library(marray)
#reading target information from file TargetsMarray.txt into MyTargets
MyTargets<-read.marrayInfo("TargetsMarray.txt")
#reading in the raw fluorescent intensities data from all 30 slides into mraw
mraw<-read.GenePix(targets=MyTargets)

#spatial plots of slide 1 green background
image(mraw[,1],xvar="maGb")
#spatial plots of slide 1 red background
image(mraw[,1],xvar="maRb")
#spatial plots of slide 1 log2 ratios
image(mraw[,1],xvar="maM")

#box plots
boxplot(mraw[,1],xvar="maPrintTip",yvar="maM",main="Print-tip box plots formslide 1: pre-normalization")
boxplot(mraw,yvar="maM",main="Array boxplots:prenormalization")
mraw.norm<-maNorm(mraw,norm="p")
boxplot(mraw.norm[,1],xvar="maPrintTip",yvar="maM",main="Print-tip boxplots for slide 1:post-normalization")
boxplot(mraw.norm,yvar="maM",col="green",main="Array boxplots: pre-normalization")
plot(mraw[,1])
plot(mraw.norm[,1],lengend.func=NULL)

#spatial plots in arrayQuality
library(arrayQuality)
maQualityPlots(mraw,norm="median")
```

A.3.2. R script: Analysis of expression data of 1680 contigs identified during desiccation in *X. humilis* leaves

```
#rename all clone IDs with their mapped contig IDs in all 30 GenePix result text files
setwd("C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\XHP Microarray data analysis FEB 2013")
#reading expression data from 30 slides in
library(limma)
targets<-readTargets("Targetsnew.txt")
RG<-read.maimages(targets$FileName,source="genepix",wt.fun=wtflags(0.1))
RG$genes<-readGAL("GALedit.gal")
RG$printer<-getLayout(RG$genes)
spottypes<-readSpotTypes("SpotTypes.txt")
RG$genes$Status<-controlStatus(spottypes,RG)
write.table(RG$genes$Status,"status",sep="\t")

# spatial correction (within slide normalization)
library(marray)
library(convert)
#convert RG from limma RG object to marrayRaw object without correcting background (KJD states that if overall background is low, rather leave the background correction out since which further introduces variables into the data.)
RG.marray<-as(RG,"marrayRaw")
```

```
#set background to zero (because marray does its own background correction, therefore it is better to reset the background values in data to zero, to prevent unnecessary removal of important data information.)
RG.marray.nbg<-RG.marray
RG.marray.nbg@maGb<-RG.marray.nbg@maRb<-0*RG.marray.nbg@maRb
library(marray)
#normalizing data using median method(spatial correction)
RG.marray.med<-maNorm(RG.marray.nbg,norm="median")
#boxplot spatial corrected data
boxplot(RG.marray.med,yvar="maM",main="after spatial normalization(with no back ground correction)")

RG.marray.med.limma<-as(RG.marray.med,"MAList")
normdata<-RG.marray.med.limma
#Rquantile normalization (across slide normalization)
library(limma)
quantnorm_r<-normalizeBetweenArrays(normdata,method="Rquantile")
#checking box plot rquantile
boxplot(quantnorm_r$M[,1],quantnorm_r$M[,2],quantnorm_r$M[,3],quantnorm_r$M[,4],quantnorm_r$M[,5],q
uantnorm_r$M[,6],quantnorm_r$M[,7],quantnorm_r$M[,8],quantnorm_r$M[,9],quantnorm_r$M[,10],quantnor
m_r$M[,11],quantnorm_r$M[,12],quantnorm_r$M[,13],quantnorm_r$M[,14],quantnorm_r$M[,15],quantnorm_
r$M[,16],quantnorm_r$M[,17],quantnorm_r$M[,18],quantnorm_r$M[,19],quantnorm_r$M[,20],quantnorm_r$
M[,21],quantnorm_r$M[,22],quantnorm_r$M[,23],quantnorm_r$M[,24],quantnorm_r$M[,25],quantnorm_r$M[,
26],quantnorm_r$M[,27],quantnorm_r$M[,28],quantnorm_r$M[,29],quantnorm_r$M[,30],main="Rquantile")
#checking print tip groups in slide1
boxplot(quantnorm_r$M[,1][quantnorm_r$genes$Block==1],
quantnorm_r$M[,1][quantnorm_r$genes$Block==2],
quantnorm_r$M[,1][quantnorm_r$genes$Block==3],
quantnorm_r$M[,1][quantnorm_r$genes$Block==4],
quantnorm_r$M[,1][quantnorm_r$genes$Block==5],
quantnorm_r$M[,1][quantnorm_r$genes$Block==6],
quantnorm_r$M[,1][quantnorm_r$genes$Block==7],
quantnorm_r$M[,1][quantnorm_r$genes$Block==8],
quantnorm_r$M[,1][quantnorm_r$genes$Block==9],
quantnorm_r$M[,1][quantnorm_r$genes$Block==10],
quantnorm_r$M[,1][quantnorm_r$genes$Block==11],
quantnorm_r$M[,1][quantnorm_r$genes$Block==12],
quantnorm_r$M[,1][quantnorm_r$genes$Block==13],
quantnorm_r$M[,1][quantnorm_r$genes$Block==14],
quantnorm_r$M[,1][quantnorm_r$genes$Block==15],
quantnorm_r$M[,1][quantnorm_r$genes$Block==16],
quantnorm_r$M[,1][quantnorm_r$genes$Block==17],
quantnorm_r$M[,1][quantnorm_r$genes$Block==18],
quantnorm_r$M[,1][quantnorm_r$genes$Block==19],
quantnorm_r$M[,1][quantnorm_r$genes$Block==20],
quantnorm_r$M[,1][quantnorm_r$genes$Block==21],
quantnorm_r$M[,1][quantnorm_r$genes$Block==22],
quantnorm_r$M[,1][quantnorm_r$genes$Block==23],
quantnorm_r$M[,1][quantnorm_r$genes$Block==24],
quantnorm_r$M[,1][quantnorm_r$genes$Block==25],
quantnorm_r$M[,1][quantnorm_r$genes$Block==26],
quantnorm_r$M[,1][quantnorm_r$genes$Block==27],
quantnorm_r$M[,1][quantnorm_r$genes$Block==28],
quantnorm_r$M[,1][quantnorm_r$genes$Block==29],
quantnorm_r$M[,1][quantnorm_r$genes$Block==30],
quantnorm_r$M[,1][quantnorm_r$genes$Block==31],
quantnorm_r$M[,1][quantnorm_r$genes$Block==32],
main="slide 1 blocks(2dl + Rquantile)")
#checking box plot rquantile cc1
boxplot(quantnorm_r$M[,1][quantnorm_r$genes$ID=="LUSC_cc1"],quantnorm_r$M[,2][quantnorm_r$genes$
ID=="LUSC_cc1"],quantnorm_r$M[,3][quantnorm_r$genes$ID=="LUSC_cc1"],quantnorm_r$M[,4][quantnor
```

```
m_r$genes$ID=="LUSC_cc1"],quantnorm_r$M[,5][quantnorm_r$genes$ID=="LUSC_cc1"],main="Rquantile
cc1")

# filtering out all control genesquantnorm.filt<-quantnorm_r[quantnorm_r$genes$Status=="gene",]
filt.ID<-(as.character(quantnorm_r$genes$ID)[quantnorm_r$genes$Status=="gene"])
# reorder data (based on ID) so that replicate spots are grouped together
# averaging all technical replicate spots
i <- order(filt.ID)
filt.ID<-filt.ID[i]
quantnorm.filt<-quantnorm.filt[i,]
ave.quantnorm<-matrix(0,ncol=30,nrow=length(unlist(lapply(split(quantnorm.filt$M[,1],filt.ID),mean))))
for (j in 1:30)
{
ave.quantnorm[,j]<-unlist(lapply(split(quantnorm.filt$M[,j],filt.ID),mean))
}
ave.gene.IDs<-names(lapply(split(quantnorm.filt$M[,1],filt.ID),mean))
length(ave.gene.IDs)
write.table(ave.gene.IDs,"genenames after med rquant norm.txt",sep="\t")

# differential expression test (Linear Model Fits, common reference)
targets<-readTargets("Targetsnew.txt")
design<-modelMatrix(targets,ref="reference")
fit<-lmFit(ave.quantnorm,design)
contrast.matrix<-makeContrasts(RWC80-RWC100,RWC60-RWC100,RWC40-RWC100,RWC20-
RWC100,RWC5-RWC100,levels=design)
contrast.matrix
fit2<-contrasts.fit(fit,contrast.matrix)
fit2<-eBayes(fit2)
results<-decideTests(fit2)
diffex<-apply(abs(results),1,sum)
diffex<-diffex>0
cluster.data<-ave.quantnorm[diffex,]
# extracts the gene names for above data
cluster.names<-ave.gene.IDs[diffex]
rownames(cluster.data)<-cluster.names
#convert the default log2[Cy5/Cy3] ratios to log2[Cy3/Cy5] ratios
cluster.data.inv<-cluster.data*-1
write.table(cluster.data.inv,file="diffexp.txt",sep="\t")

# preparing differentially expressed data for clustering
# open diffexp.txt in Excel and average the biological replicates
# save file as clusterdatamean.txt
# remove all col names and row names and save file as clusterdatamean2.txt
# read in clusterdatamean2.txt
cluster.data.mean<-read.table(file="C:\\Documents and Settings\\Administrator\\My Documents\\My
Work\\XHP Microarray data analysis FEB 2013\\clusterdatamean2.txt",sep="\t")
#converting cluster.data.mean into matrix
cluster.data.mean<-as.matrix(cluster.data.mean)
#converting cluster.data.mean into expr
library(Biobase)
covdesc<- list("Condition")
names(covdesc) <-"Condition"
geneCov<-as.data.frame(c("RWC100","RWC80","RWC60","RWC40","RWC20","RWC5"))
pdata <- new("phenoData", pData=geneCov, varLabels=covdesc)
eset <- new("exprSet", exprs=cluster.data.mean, phenoData=pdata)
AveClusterData<-exprs(eset)
colnames(AveClusterData)<-pData(eset)[,1]
rownames(AveClusterData)<-cluster.names

# clustering of diffexp genes
```

```
#determining optimal k value using mclust
library(mclust)
mclust<-Mclust(AveClusterData)
mclust
plot(mclust)
#pamsam clustering
library(smida)
pamsam7<-pamsam(AveClusterData,k=7,metric="correlation")
#extracting contig IDs from each cluster
write.table(cluster.names[pamsam7$clustering==1],"pamsam7_cluster_1_genes.txt",sep="\t")
write.table(cluster.names[pamsam7$clustering==2],"pamsam7_cluster_2_genes.txt",sep="\t")
write.table(cluster.names[pamsam7$clustering==3],"pamsam7_cluster_3_genes.txt",sep="\t")
write.table(cluster.names[pamsam7$clustering==4],"pamsam7_cluster_4_genes.txt",sep="\t")
write.table(cluster.names[pamsam7$clustering==5],"pamsam7_cluster_5_genes.txt",sep="\t")
write.table(cluster.names[pamsam7$clustering==6],"pamsam7_cluster_6_genes.txt",sep="\t")
write.table(cluster.names[pamsam7$clustering==7],"pamsam7_cluster_7_genes.txt",sep="\t")

# clustering of RWC samples
library(smida)
#transposing dimension
sample.cluster<-t(AveClusterData)
sammon<-cluster.samples(sample.cluster,method="sammon",metric="correlation",cex.lab=0.8)
pca<-cluster.samples(sample.cluster,method="pca",metric="correlation",cex.lab=0.8)
library(cluster)
diana<-diana(sample.cluster,metric="correlation")
plot(diana,cex.lab=0.8)
```

A.3.3. R script: Analysis of expression data of 772 orthologues identified during desiccation in *X. humilis* leaves

```
#rename all contig IDs with their mapped Arabidopsis orthologue IDs in all 30 GenePix result text files
setwd("C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013")
#reading expression data from 30 slides in
library(limma)
targets<-readTargets("Targetsnew.txt")
RG<-read.maimages(targets$FileName,source="genepix",wt.fun=wtflags(0.1))
RG$genes<-readGAL("GALedit.gal")
RG$printer<-getLayout(RG$genes)
spottypes<-readSpotTypes("SpotTypes.txt")
RG$genes$Status<-controlStatus(spottypes,RG)
write.table(RG$genes$Status,"status",sep="\t")

# spatial correction (within slide normalization)
library(marray)
library(convert)
#convert RG from limma RG object to marrayRaw object without correcting background (KJD states that if
overall background is low, rather leave the background correction out since which further introduces variables
into the data.)
RG.marray<-as(RG,"marrayRaw")
#set background to zero (because marray does its own background correction, therefore it is better to reset the
background values in data to zero, to prevent unnecessary removal of important data information.)
RG.marray.nbg<-RG.marray
RG.marray.nbg@maGb<-RG.marray.nbg@maRb<-0*RG.marray.nbg@maRb

#normalizing data using median method(spatial correction)
RG.marray.med<-maNorm(RG.marray.nbg,norm="median")
RG.marray.med.limma<-as(RG.marray.med,"MAList")
normdata<-RG.marray.med.limma
#Rquantile normalization (across slide normalization)
```

```
library(limma)
quantnorm_r<-normalizeBetweenArrays(normdata,method="Rquantile")
# filtering out all control genesquantnorm.filt<-quantnorm_r[quantnorm_r$genes$Status=="gene",]
filt.ID<-(as.character(quantnorm_r$genes$ID)[quantnorm_r$genes$Status=="gene"])
# reorder data (based on ID) so that replicate spots are grouped together
# averaging all technical replicate spots
i <- order(filt.ID)
filt.ID<-filt.ID[i]
quantnorm.filt<-quantnorm.filt[i,]
ave.quantnorm<-matrix(0,ncol=30,nrow=length(unlist(lapply(split(quantnorm.filt$M[,1],filt.ID),mean))))
for (j in 1:30)
{
ave.quantnorm[,j]<-unlist(lapply(split(quantnorm.filt$M[,j],filt.ID),mean))
}
ave.gene.IDs<-names(lapply(split(quantnorm.filt$M[,1],filt.ID),mean))

#extracting expression data of 772 subset
# mutual772atgs.txt contains the list of 772 gene IDs
atg772<-read.table(file="C:\\Documents and Settings\\Administrator\\My Documents\\My
Work\\At_Xh_FEB2013\\mutual772atgs.txt",sep="\t")
atg772_matrix<-as.matrix(atg772)
ave.quantnorm772<-ave.quantnorm[ave.gene.IDs%in%atg772_matrix,]
ave.gene.IDs772<-ave.gene.IDs[ave.gene.IDs%in%atg772_matrix]

# differential expression test (Linear Model Fits, common reference)
targets<-readTargets("Targetsnew.txt")
design<-modelMatrix(targets,ref="reference")
fit<-lmFit(ave.quantnorm722,design)
contrast.matrix<-makeContrasts(RWC80-RWC100,RWC60-RWC100,RWC40-RWC100,RWC20-
RWC100,RWC5-RWC100,levels=design)
contrast.matrix
fit2<-contrasts.fit(fit,contrast.matrix)
fit2<-eBayes(fit2)
results<-decideTests(fit2)
diffex<-apply(abs(results),1,sum)
diffex<-diffex>0
cluster.data<-ave.quantnorm772[diffex,]
# extracts the gene names
cluster.names<-ave.gene.IDs772[diffex]
rownames(cluster.data)<-cluster.names
#convert the default log2[Cy5/Cy3] ratios to log2[Cy3/Cy5] ratios
cluster.data.inv<-cluster.data*-1

# preparing differentially expressed data for clustering
# open diffexp.txt in Excel and average the biological replicates
# save file as clusterdatamean.txt
# remove all col names and row names and save file as clusterdatamean2.txt
# read in clusterdatamean2.txt
cluster.data.mean<-read.table(file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\
At_Xh_FEB2013\\clusterdatamean2.txt",sep="\t")
#converting cluster.data.mean into matrix
cluster.data.mean<-as.matrix(cluster.data.mean)
#converting cluster.data.mean into expr
library(Biobase)
covdesc<- list("Condition")
names(covdesc) <-"Condition"
geneCov<-as.data.frame(c("RWC100","RWC80","RWC60","RWC40","RWC20","RWC5"))
pdata <- new("phenoData", pData=geneCov, varLabels=covdesc)
eset <- new("exprSet", exprs=cluster.data.mean, phenoData=pdata)
AveClusterData<-exprs(eset)
```

```
colnames(AveClusterData)<-pData(eset)[,1]
rownames(AveClusterData)<-cluster.names

# clustering of diffexp genes
#determining optimal k value using mclust
library(mclust)
mclust<-Mclust(AveClusterData)
mclust
plot(mclust)
#pamsam clustering
library(smida)
pamsam7xh<-pamsam(AveClusterData,k=7,metric="correlation")
#extracting orthologue IDs from each cluster
write.table(cluster.names[pamsam7xh$clustering==1],"pamsam7xh_cluster1_genes.txt",sep="\t")
write.table(cluster.names[pamsam7xh$clustering==2],"pamsam7xh_cluster2_genes.txt",sep="\t")
write.table(cluster.names[pamsam7xh$clustering==3],"pamsam7xh_cluster3_genes.txt",sep="\t")
write.table(cluster.names[pamsam7xh$clustering==4],"pamsam7xh_cluster4_genes.txt",sep="\t")
write.table(cluster.names[pamsam7xh$clustering==5],"pamsam7xh_cluster5_genes.txt",sep="\t")
write.table(cluster.names[pamsam7xh$clustering==6],"pamsam7xh_cluster6_genes.txt",sep="\t")
write.table(cluster.names[pamsam7xh$clustering==7],"pamsam7xh_cluster7_genes.txt",sep="\t")
```

A.3.4. R script: Analysis of expression data of 772 orthologues identified during seed maturation and osmotic stress in *A. thaliana*

```
setwd("C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013")
library(Biobase)
library(limma)
library(smida)
#reading expression values in from csv file containing data of 22414 genes
input<-read.table(file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\Le_Nakabayashi_Seed data_Osm_consolidated_ctrl removed_05OCT2010.csv",header = TRUE, sep = ",", quote="\"", dec=".")
genedata<-matrix(0,nrow(input),ncol=32)
#seed maturation data
genedata[,1]<-input[,"OV_1"]
genedata[,2]<-input[,"OV_2"]
genedata[,3]<-input[,"ZYG_1"]
genedata[,4]<-input[,"ZYG_2"]
genedata[,5]<-input[,"GLOB_1"]
genedata[,6]<-input[,"GLOB_2"]
genedata[,7]<-input[,"COT_1"]
genedata[,8]<-input[,"COT_2"]
genedata[,9]<-input[,"MG_1"]
genedata[,10]<-input[,"MG_2"]
genedata[,11]<-input[,"PMG_1"]
genedata[,12]<-input[,"PMG_2"]
genedata[,13]<-input[,"DS_1"]
genedata[,14]<-input[,"DS_2"]
genedata[,15]<-input[,"IMB24_1"]
genedata[,16]<-input[,"IMB24_2"]
genedata[,17]<-input[,"SDLG_1"]
genedata[,18]<-input[,"SDLG_2"]
#osmotic stress data
genedata[,19]<-input[,"L0_1"]
genedata[,20]<-input[,"L0_2"]
genedata[,21]<-input[,"LO0.5_1"]
genedata[,22]<-input[,"LO0.5_2"]
genedata[,23]<-input[,"LO1_1"]
genedata[,24]<-input[,"LO1_2"]
```

```
genedata[,25]<-input[,"LO3_1"]
genedata[,26]<-input[,"LO3_2"]
genedata[,27]<-input[,"LO6_1"]
genedata[,28]<-input[,"LO6_2"]
genedata[,29]<-input[,"LO12_1"]
genedata[,30]<-input[,"LO12_2"]
genedata[,31]<-input[,"LO24_1"]
genedata[,32]<-input[,"LO24_2"]
genedata_log<-log2(genedata)

# converting genedata into exprSet object
covdesc<- list("Condition")
names(covdesc) <-"Condition"
geneCov<-
as.data.frame(c("OV_1","OV_2","ZYG_1","ZYG_2","GLOB_1","GLOB_2","COT_1","COT_2","MG_1","MG
_2","PMG_1","PMG_2","DS_1","DS_2","IMB24_1","IMB24_2","SD_SDLG_1","SD_SDLG_2","L0_1","L0_
2","LO0.5_1","LO0.5_2","LO1_1","LO1_2","LO3_1","LO3_2","LO6_1","LO6_2","LO12_1","LO12_2","LO2
4_1","LO24_2"))
pdata <- new("phenoData", pData=geneCov, varLabels=covdesc)
eset <- new("exprSet", exprs=genedata_log, phenoData=pdata)
ExpData<-exprs(eset)
colnames(ExpData)<-pData(eset)[,1]
write.table(ExpData,file="C:\\Documents and Settings\\Administrator\\My Documents\\My
Work\\At_Xh_FEB2013\\ExpDat.txt",sep="\t",row.names=FALSE)

#extracting gene IDs and layout info from grp file
genes <- data.frame(ID=input$GENENAME,Name=input$GENENAME)
# set control status
spottypes<-readSpotTypes()
genes$Status<-controlStatus(spottypes,genes=genes)

#data normalization
dat<-exprs(eset)
dim(dat)
library(affy)
dat.cond<-normalize.quantiles(dat,copy=TRUE)
norm.eset<-new("exprSet", exprs=dat.cond , phenoData=pdata)
dat.cond<-exprs(norm.eset)
# filtering out all control genes
dat.cond.filt<-dat.cond[genes$Status=="gene",]
filt.ID<-(as.character(genes$ID)[genes$Status=="gene"])
# reordering data so that replicate spots are grouped together
# averaging of technical spots
i <- order(filt.ID)
filt.ID<-filt.ID[i]
dat.cond.filt<-dat.cond.filt[i,]
ave.dat.cond<-matrix(0,ncol=32,nrow=length(unlist(lapply(split(dat.cond.filt[,1],filt.ID),mean))))
for (j in 1:32)
{
ave.dat.cond[,j]<-unlist(lapply(split(dat.cond.filt[,j],filt.ID),mean))
}
ave.gene.IDs<-names(lapply(split(dat.cond.filt[,1],filt.ID),mean))
#creating Exprs Set of all filtered, normalised and averaged data
exprs(norm.eset)<-ave.dat.cond
norm.ExpData<-exprs(norm.eset)
colnames(norm.ExpData)<-pData(norm.eset)[,1]

# extracting normalized seed expression data
MG_1<-(norm.ExpData[,9])
MG_2<-(norm.ExpData[,10])
```

```
PMG_1<-(norm.ExpData[,11])
PMG_2<-(norm.ExpData[,12])
DS_1<-(norm.ExpData[,13])
DS_2<-(norm.ExpData[,14])
# arraying seed data into 22414 x 6 matrix
seed.subset.data<-matrix(0,nrow=22414,ncol=6)
seed.subset.data[,1]<-MG_1
seed.subset.data[,2]<-MG_2
seed.subset.data[,3]<-PMG_1
seed.subset.data[,4]<-PMG_2
seed.subset.data[,5]<-DS_1
seed.subset.data[,6]<-DS_2
#creating seed exprSet, assigning column names (experimental conditions)
seed.covdesc<- list("Condition")
seed.names(covdesc) <-"Condition"
seed.geneCov<-as.data.frame(c("MG_1","MG_2","PMG_1","PMG_2","DS_1","DS_2"))
seed.pdata <- new("phenoData", pData=seed.geneCov, varLabels=seed.covdesc)
seed.subset.eset <- new("exprSet", exprs=seed.subset.data, phenoData=seed.pdata)
seed.ExpData.subset<-exprs(seed.subset.eset)
colnames(seed.ExpData.subset)<-pData(seed.subset.eset)[,1]
# extracting normalized leaf osmotic expression data
L0_1<-(norm.ExpData[,19])
L0_2<-(norm.ExpData[,20])
LO0.5_1<-(norm.ExpData[,21])
LO0.5_2<-(norm.ExpData[,22])
LO1_1<-(norm.ExpData[,23])
LO1_2<-(norm.ExpData[,24])
LO3_1<-(norm.ExpData[,25])
LO3_2<-(norm.ExpData[,26])
LO6_1<-(norm.ExpData[,27])
LO6_2<-(norm.ExpData[,28])
LO12_1<-(norm.ExpData[,29])
LO12_2<-(norm.ExpData[,30])
LO24_1<-(norm.ExpData[,31])
LO24_2<-(norm.ExpData[,32])
# arraying osmotic data into 22414 x 14 matrix
subset.data<-matrix(0,nrow=22414,ncol=14)
subset.data[,1]<-L0_1
subset.data[,2]<-L0_2
subset.data[,3]<-LO0.5_1
subset.data[,4]<-LO0.5_2
subset.data[,5]<-LO1_1
subset.data[,6]<-LO1_2
subset.data[,7]<-LO3_1
subset.data[,8]<-LO3_2
subset.data[,9]<-LO6_1
subset.data[,10]<-LO6_2
subset.data[,11]<-LO12_1
subset.data[,12]<-LO12_2
subset.data[,13]<-LO24_1
subset.data[,14]<-LO24_2
#creating osm exprSet, assigning column names (experimental conditions)
osm.covdesc<- list("Condition")
osm.names(covdesc) <-"Condition"
osm.geneCov<-
as.data.frame(c("L0_1","L0_2","LO0.5_1","LO0.5_2","LO1_1","LO1_2","LO3_1","LO3_2","LO6_1","LO6_2
","LO12_1","LO12_2","LO24_1","LO24_2"))
osm.pdata <- new("phenoData", pData=osm.geneCov, varLabels=osm.covdesc)
osm.subset.eset <- new("exprSet", exprs=osm.subset.data, phenoData=osm.pdata)
osm.ExpData.subset<-exprs(osm.subset.eset)
```

```
colnames(osm.ExpData.subset)<-pData(osm.subset.eset)[,1]

#extracting expression data of 772 subset
# mutual772atgs.txt contains the list of 772 gene IDs
atg772<-read.table(file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\mutual772atgs.txt",sep="\t")
atg772_matrix<-as.matrix(atg772)
#for seed data
seed.ExpData.subset772<-seed.ExpData.subset[ave.gene.IDs%in%atg772_matrix,]
seed.ave.gene.IDs772<-ave.gene.IDs[ave.gene.IDs%in%atg772_matrix]
#for osm data
osm.ExpData.subset772<-osm.ExpData.subset[ave.gene.IDs%in%atg772_matrix,]
osm.ave.gene.IDs772<-ave.gene.IDs[ave.gene.IDs%in%atg772_matrix]

# differential expression test (Linear model fits)
#for seed data
seed.design<-model.matrix(~0+factor(c(1,1,2,2,3,3)))
seed.colnames(design)<-c("MG","PMG","DS")
seed.fit<-lmFit(seed.ExpData.subset772,seed.design)
seed.contrast.matrix<-makeContrasts("PMG-MG","DS-MG",levels=seed.design)
seed.fit2<-contrasts.fit(seed.fit,seed.contrast.matrix)
seed.fit2<-eBayes(seed.fit2)
seed.results<-decideTests(seed.fit2)
seed.diffex<-apply(abs(seed.results),1,sum)
seed.diffex<-seed.diffex>0
seed.cluster.data<-seed.ExpData.subset772[seed.diffex,]
#for osm data
osm.design<-model.matrix(~0+factor(c(1,1,2,2,3,3,4,4,5,5,6,6,7,7)))
osm.colnames(design)<-c("L0","LO0.5","LO1","LO3","LO6","LO12","LO24")
osm.fit<-lmFit(osm.ExpData.subset772,osm.design)
osm.contrast.matrix<-makeContrasts("LO0.5-L0","LO1-LO0","LO3-LO0","LO6-LO0","LO12-LO0","LO24-LO0",levels=osm.design)
osm.fit2<-contrasts.fit(osm.fit,osm.contrast.matrix)
osm.fit2<-eBayes(osm.fit2)
osm.results<-decideTests(osm.fit2)
osm.diffex<-apply(abs(osm.results),1,sum)
osm.diffex<-osm.diffex>0
osm.cluster.data<-osm.ExpData.subset772[osm.diffex,]

# preparing differentially expressed data for clustering
#for seed data
seed.cluster.names<-seed.ave.gene.IDs772[seed.diffex]
seed.rownames(cluster.data)<-seed.cluster.names
unique(seed.rownames(cluster.data))
dim(seed.cluster.data)
write.table(cluster.data,file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\seed.cluster_data.txt",sep="\t",row.names=FALSE)
#for osm data
osm.cluster.names<-osm.ave.gene.IDs772[osm.diffex]
osm.rownames(cluster.data)<-osm.cluster.names
unique(osm.rownames(cluster.data))
dim(osm.cluster.data)
write.table(cluster.data,file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\osm.cluster_data.txt",sep="\t",row.names=FALSE)

# open seed.cluster_data.txt/ osm.cluster_data.txt in Excel and average the biological replicates
# save the file as seed.cluster_data_mean.txt/ osm.cluster_data.txt
# read seed.cluster_data_mean.txt/ osm.cluster_data.txt in
seed.cluster.data.mean<-read.table(file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\seed.cluster_data_mean.txt",sep="\t",header=TRUE)
```

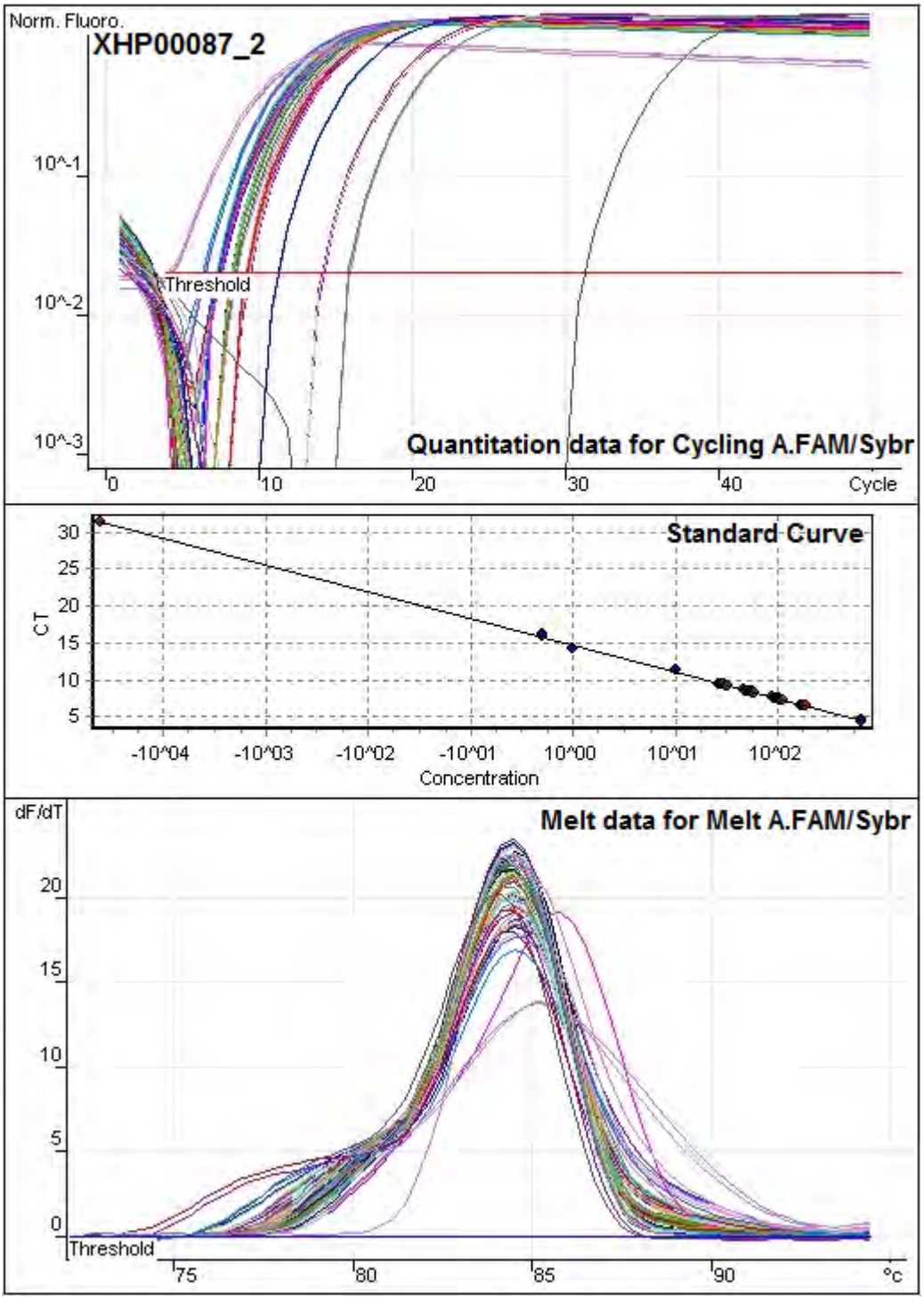
```
osm.cluster.data.mean<-read.table(file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\osm.cluster_data_mean.txt",sep="\t",header=TRUE)

#convert seed.cluster.data.mean into matrix
seed.cluster.data.mean<-as.matrix(seed.cluster.data.mean)
#convert seed.cluster.data.mean into expr
library(Biobase)
seed.covdesc<- list("Condition")
seed.names(covdesc) <-"Condition"
seed.geneCov<-as.data.frame(c("MG","PMG","DS"))
seed.pdata <- new("phenoData", pData=seed.geneCov, varLabels=seed.covdesc)
seed.eset <- new("exprSet", exprs=seed.cluster.data.mean, phenoData=seed.pdata)
seed.AveClusterData<-exprs(seed.eset)
colnames(seed.AveClusterData)<-pData(seed.eset)[,1]
rownames(seed.AveClusterData)<-seed.cluster.names
unique(rownames(seed.AveClusterData))
dim(seed.AveClusterData)
#convert osm.cluster.data.mean into matrix
osm.cluster.data.mean<-as.matrix(osm.cluster.data.mean)
#convert osm.cluster.data.mean into expr
library(Biobase)
osm.covdesc<- list("Condition")
osm.names(covdesc) <-"Condition"
osm.geneCov<-as.data.frame(c("L0","LO0.5","LO1","LO3","LO6","LO12","LO24"))
osm.pdata <- new("phenoData", pData=osm.geneCov, varLabels=osm.covdesc)
osm.eset <- new("exprSet", exprs=osm.cluster.data.mean, phenoData=osm.pdata)
osm.AveClusterData<-exprs(osm.eset)
colnames(osm.AveClusterData)<-pData(osm.eset)[,1]
rownames(osm.AveClusterData)<-osm.cluster.names
unique(rownames(osm.AveClusterData))
dim(osm.AveClusterData)

#clustering diffexp seed data into 7 clusters
pamsam7seed<-pamsam(seed.AveClusterData,k=7,metric="correlation")
#extracting orthologue IDs of each cluster
write.table(cluster.names[pamsam7seed$clustering==1],file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\pamsam7seed_cluster1_genes.txt",sep="\t",row.names=FALSE)
write.table(cluster.names[pamsam7seed$clustering==2],file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\pamsam7seed_cluster2_genes.txt",sep="\t",row.names=FALSE)
write.table(cluster.names[pamsam7seed$clustering==3],file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\pamsam7seed_cluster3_genes.txt",sep="\t",row.names=FALSE)
write.table(cluster.names[pamsam7seed$clustering==4],file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\pamsam7seed_cluster4_genes.txt",sep="\t",row.names=FALSE)
write.table(cluster.names[pamsam7seed$clustering==5],file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\pamsam7seed_cluster5_genes.txt",sep="\t",row.names=FALSE)
write.table(cluster.names[pamsam7seed$clustering==6],file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\pamsam7seed_cluster6_genes.txt",sep="\t",row.names=FALSE)
write.table(cluster.names[pamsam7seed$clustering==7],file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\pamsam7seed_cluster7_genes.txt",sep="\t",row.names=FALSE)

#clustering diffexp osm data into 7 clusters
pamsam7osm<-pamsam(osm.AveClusterData,k=7,metric="correlation")
#extracting orthologue IDs of each cluster
write.table(cluster.names[pamsam7osm$clustering==1],file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\pamsam7osm_cluster1_genes.txt",sep="\t",row.names=FALSE)
write.table(cluster.names[pamsam7osm$clustering==2],file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\pamsam7osm_cluster2_genes.txt",sep="\t",row.names=FALSE)
write.table(cluster.names[pamsam7osm$clustering==3],file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\pamsam7osm_cluster3_genes.txt",sep="\t",row.names=FALSE)
```

write.table(cluster.names[pamsam7osm$clustering==4],file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\pamsam7osm_cluster4_genes.txt",sep="\t",row.names=FALSE)
write.table(cluster.names[pamsam7osm$clustering==5],file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\pamsam7osm_cluster5_genes.txt",sep="\t",row.names=FALSE)
write.table(cluster.names[pamsam7osm$clustering==6],file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\pamsam7osm_cluster6_genes.txt",sep="\t",row.names=FALSE)
write.table(cluster.names[pamsam7osm$clustering==7],file="C:\\Documents and Settings\\Administrator\\My Documents\\My Work\\At_Xh_FEB2013\\pamsam7osm_cluster7_genes.txt",sep="\t",row.names=FALS

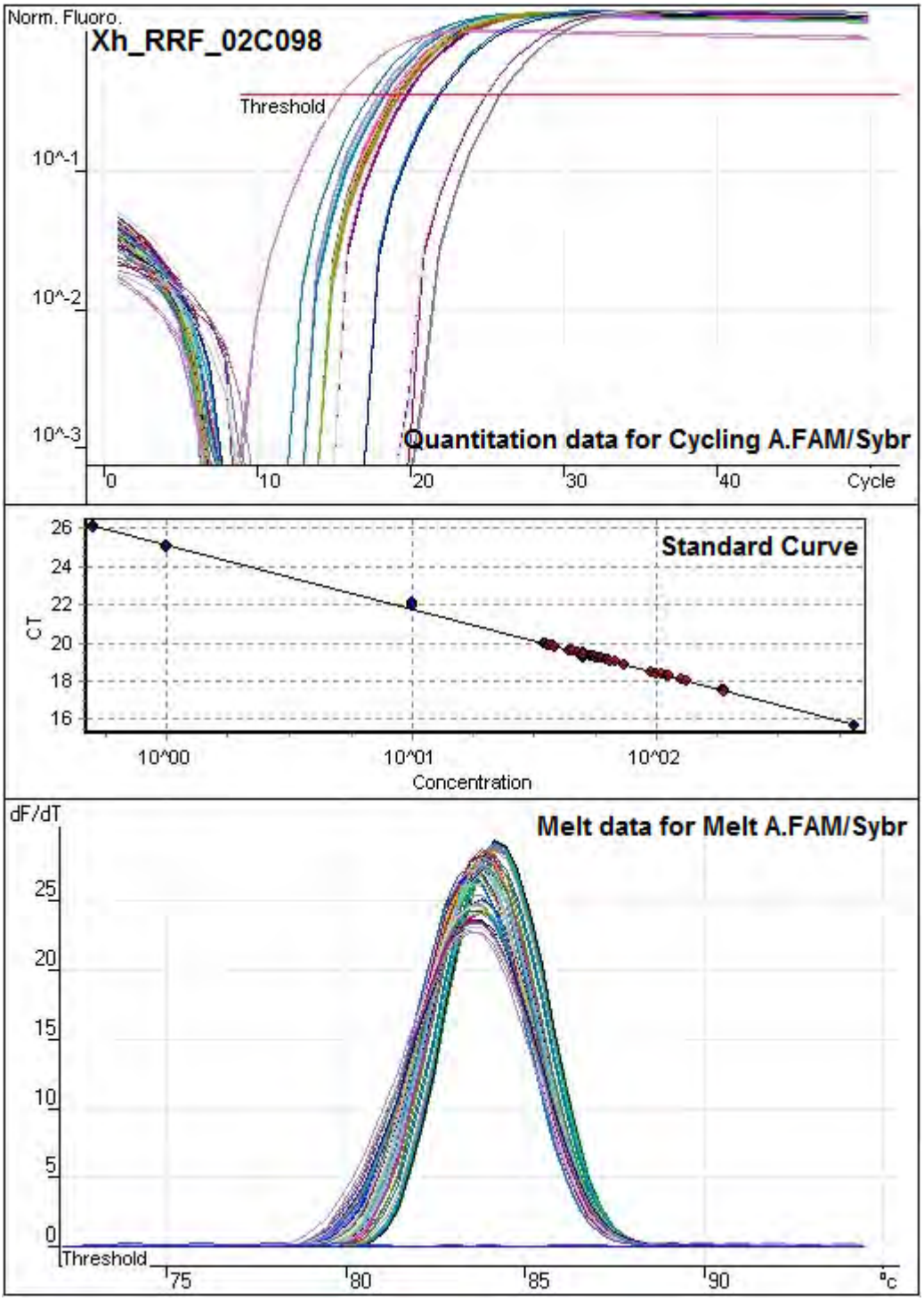
Table A.4.1. Primer sequences used in RT-qPCR analysis.

Contig ID	Forward primer	Reverse primer
XHP00087_2	GTG GAG GTG ATC GCT ATT CC	GCG GTC AAA AAC GAG AGA AC
Xh_RRF_02C098	ACG GAA GGA AGA GCT TGA CA	GGC GAC ATT CTT ACG CTG AT
XHP00531_2	GGT GGG TCT GTG CTT GAT TT	CAA CCC AGG AAA CGA GAC AT
XHP00196_1	ACA TCC AGC CTT TGT CCA AC	CAC GAG GCA AGA GAG AAA CC
XHP01603_1	AAT GTG GAT CTT GGC ACA GC	TTC ACT GGT CAC CTG GAA GG
XHP00030_10	CTT AGC AGC ACC GAA ACA CA	CCG GCA ACC AGT AAA CCT AA
XHP00474_1	AAG TAC GAG ACT CGG CCA AA	ACT CTT GAG GGC CGA CTT TT
XHP00230_2	CCT TCA GGC ACA ACA ACC TT	GTG TCA TGG ATT GTG CCA AG
XHP01261_1	CGT GCC ACT GCT TGT AAG TC	CAC AGA GGT TGC TGA TCG AA
XHP00052_3	GGA CAT CGA CTA CGA GTT CCA	GTT GGC AGC TTC AAC TCT CC
XHP00602_1	AAT GGC GAA CAA ATC TCT CG	GCA GCC AAC AGA AGA AAA GC
XHP01295_1	CAT GAA GTG GAT GGA TGC AG	AAA TGC AGG ATT GAC CCA AG
XHP01605_1	TAA AAT CGT TGG AGG GCT TG	ATG ATC GGT GCC TAA AAC CA
XHP01637_1	GAG AAC CAG TGG GCT TTT GA	CTG CAG CGA CGT TAA ATC CT
XHP00050_2	TCA AAA CCC TGA CTG GGA AG	CGA CCA TCC TCC AAC TGT TT
XHP00708_1	ACC TTC TCA GGT GAG CCA GA	GCC TGG ATG TGT CAA CAC TG
XHP00290_1	TTG AGG CTG GTT CGA CTC TT	GCT GCT TCA ACG ACC TTT TC
XHP00122_1	ATG CAT GGG CTT ATG CTA CC	GAA ACG TTG AAG GCA CGA TT
XHP01894_1	TGC GCA ATG AAT TCT AAC CA	TGA AGA TCC AGG GAA TCT GG
XHP00266_2	AGC TGG TTC TCA GTC GTC GT	GCG ATG CTT CAT TTC ACA GA
XHP00085_1	CGA CGA ATC ACT CGA CTT CA	TGC AGG GAT AAA GGA AAT GG
XHP01838_1	AAG CCC AGA ATG TGG ATT TG	GGA AAG AAC AAA TGG CGA AA
XHP00514_2	ATG CCA TCA AGA CTG GTT CC	GCT CTT GGA TCA GCG AAA TC
Xh_LDR_38B063	GCC AAT GGG TTA CTG GTG TT	AGT CTC GGA AGG GCT TGA TT

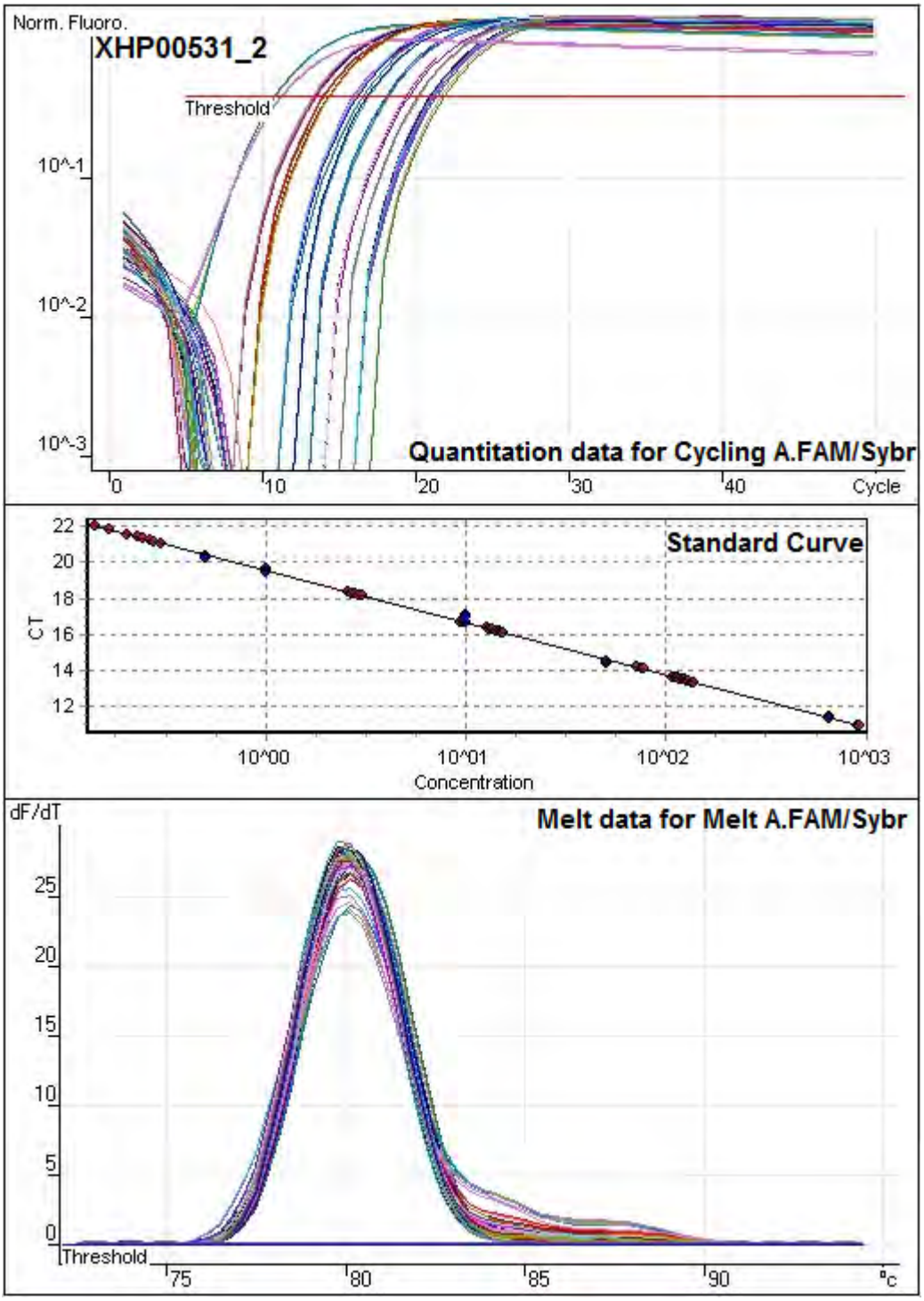
A.4.2. The amplification, melting curve and standard curve plots of RT-qPCR reactions.
A.4.2.1.



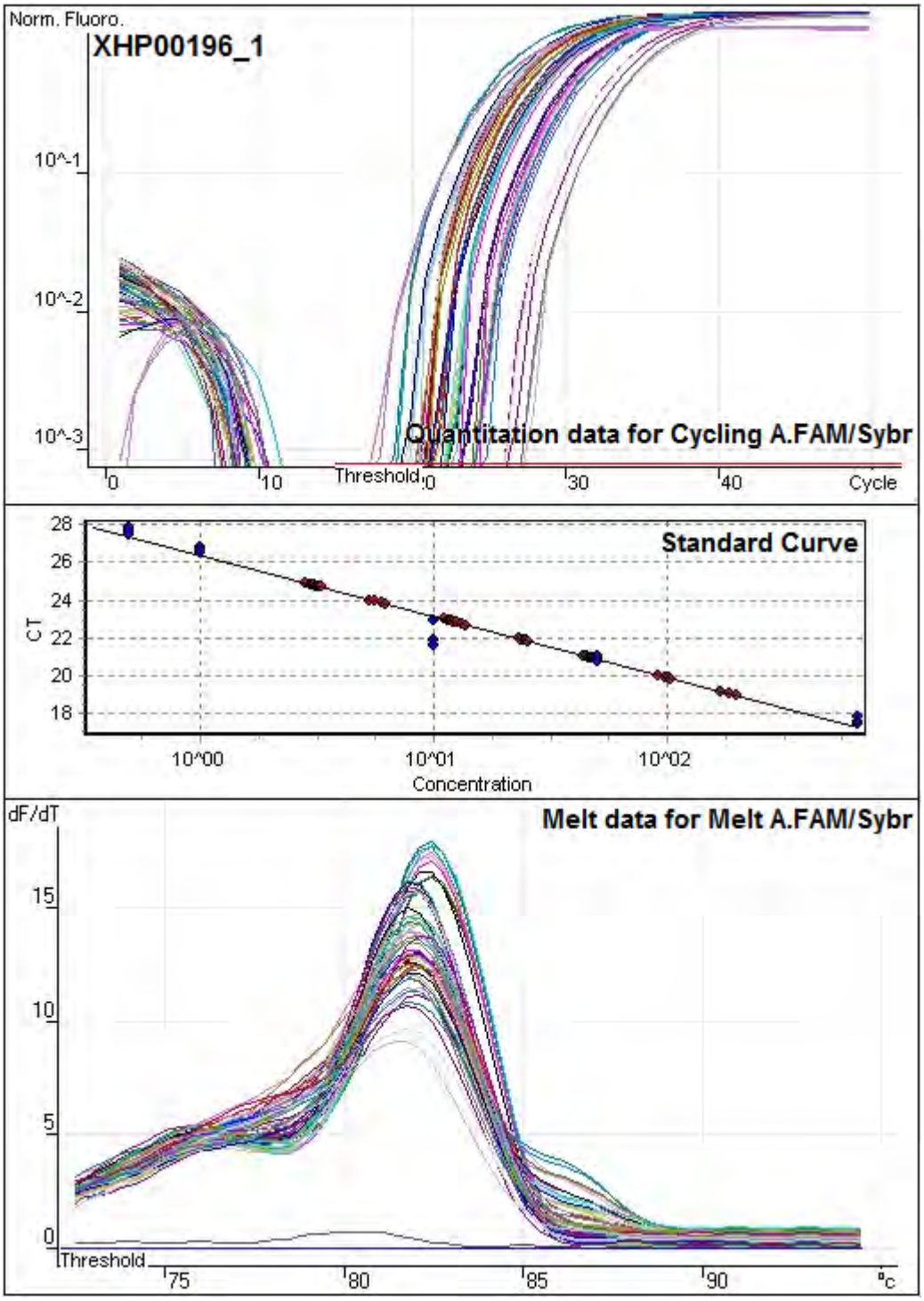
A.4.2.2.


Norm. Fluoro.
Xh_RRF_02C098
Threshold
10^-1
10^-2
10^-3
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
Standard Curve
CT
26
24
22
20
18
16
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
25
20
15
10
5
0
Threshold
75
80
85
90
°c

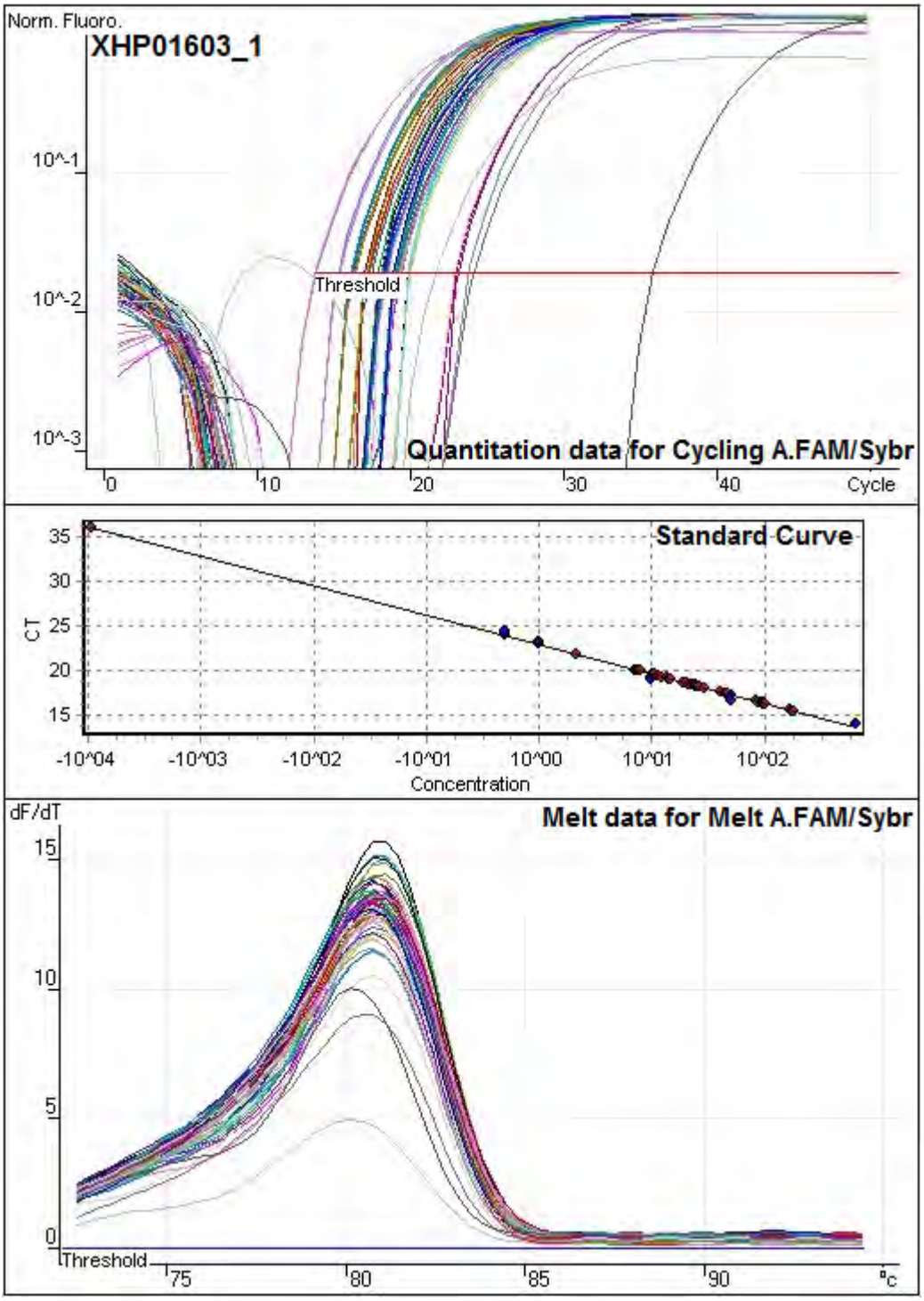
A.4.2.3.


Norm. Fluoro.
XHP00531_2
Threshold
10^-1
10^-2
10^-3
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
Standard Curve
CT
22
20
18
16
14
12
10^00
10^01
10^02
10^03
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
25
20
15
10
5
0
Threshold
75
80
85
90
°c

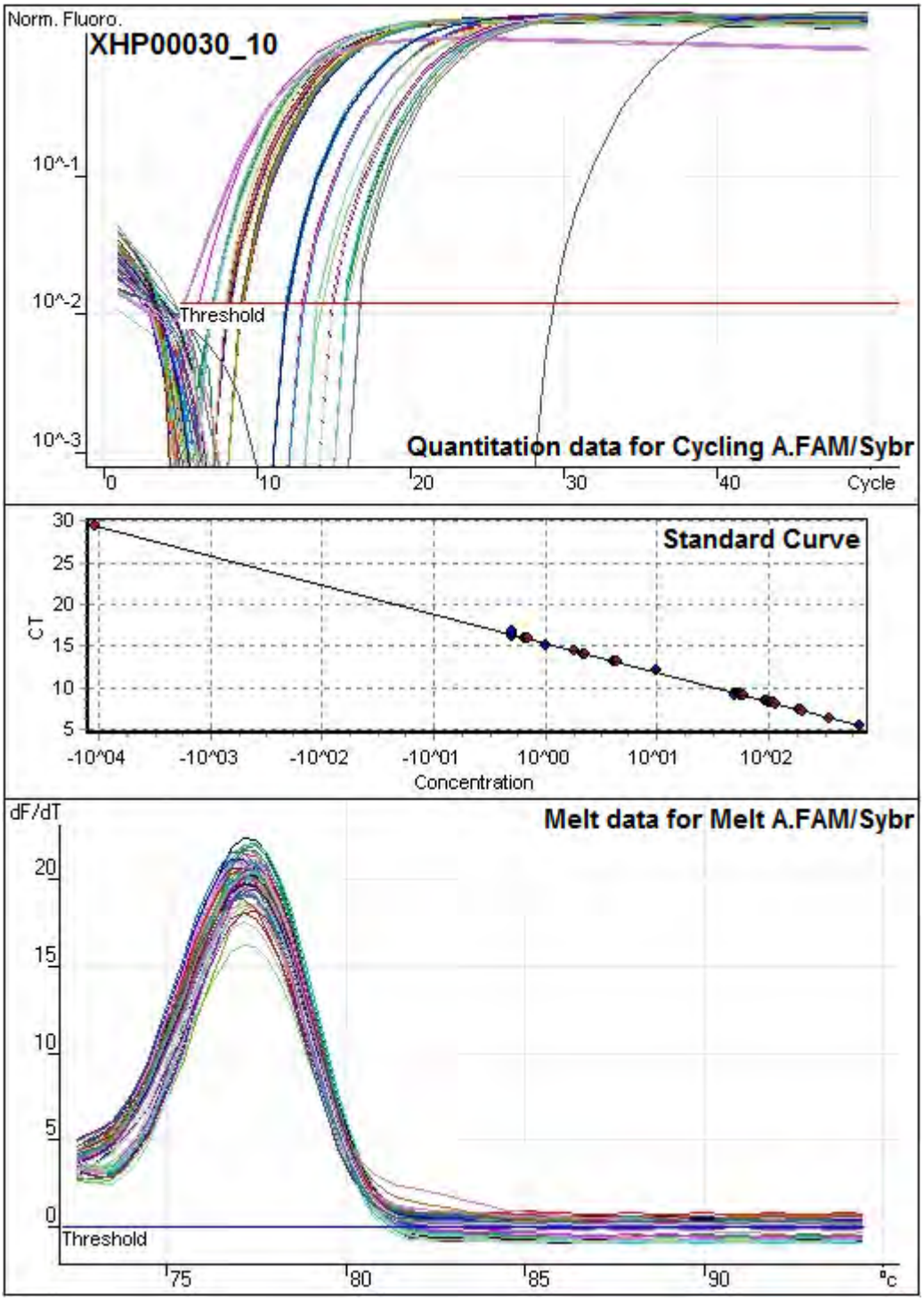
A.4.2.4.


Norm. Fluoro.
XHP00196_1
10^-1
10^-2
10^-3
Quantitation data for Cycling A.FAM/Sybr
0
10
Threshold
0
30
40
Cycle
Standard Curve
28
26
24
22
20
18
CT
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
15
10
5
0
Threshold
75
80
85
90
°c

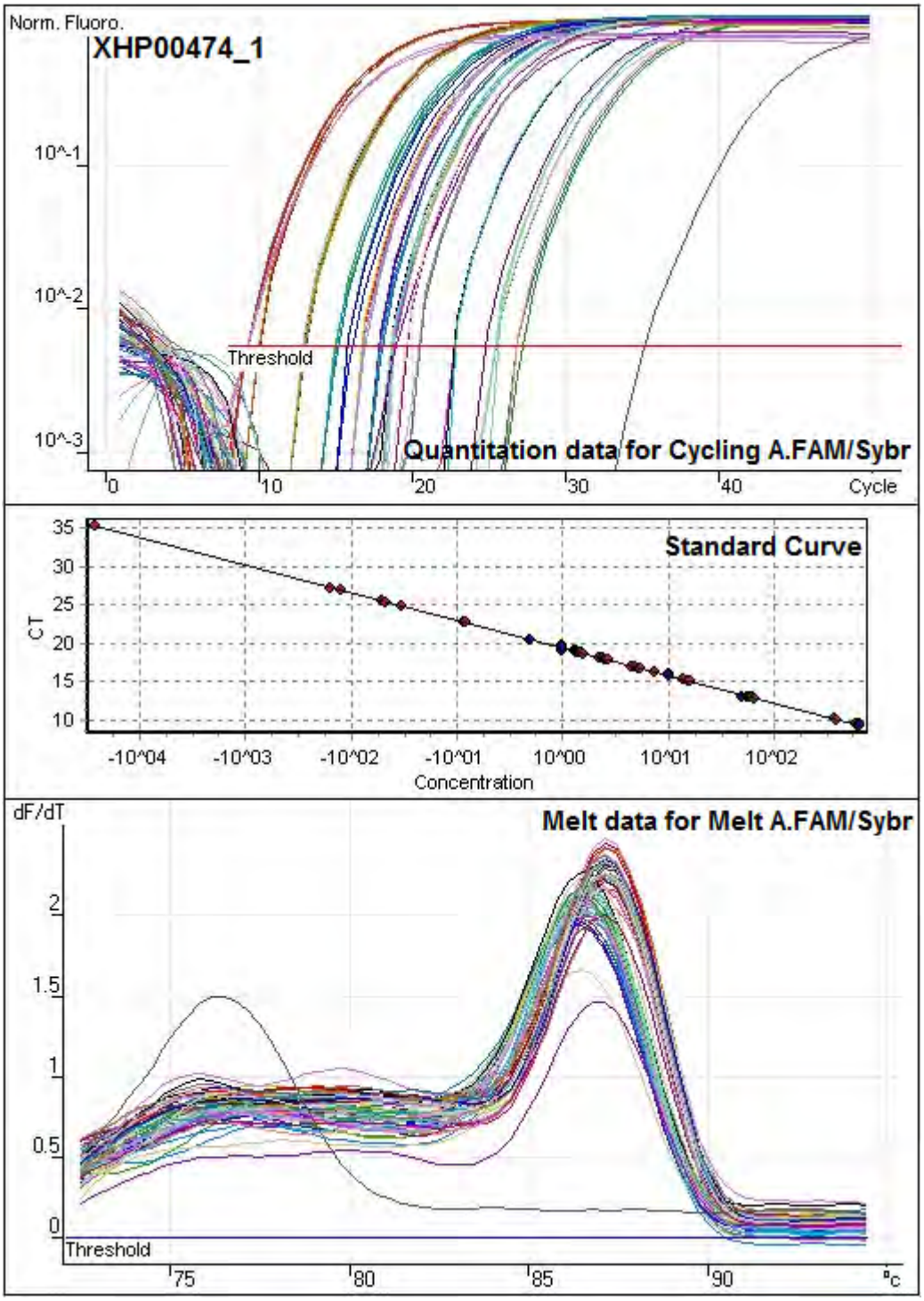
A.4.2.5.


Norm. Fluoro.
XHP01603_1
10^-1
10^-2
10^-3
Threshold
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
Standard Curve
CT
35
30
25
20
15
-10^04
-10^03
-10^02
-10^01
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
15
10
5
0
Threshold
75
80
85
90
°c

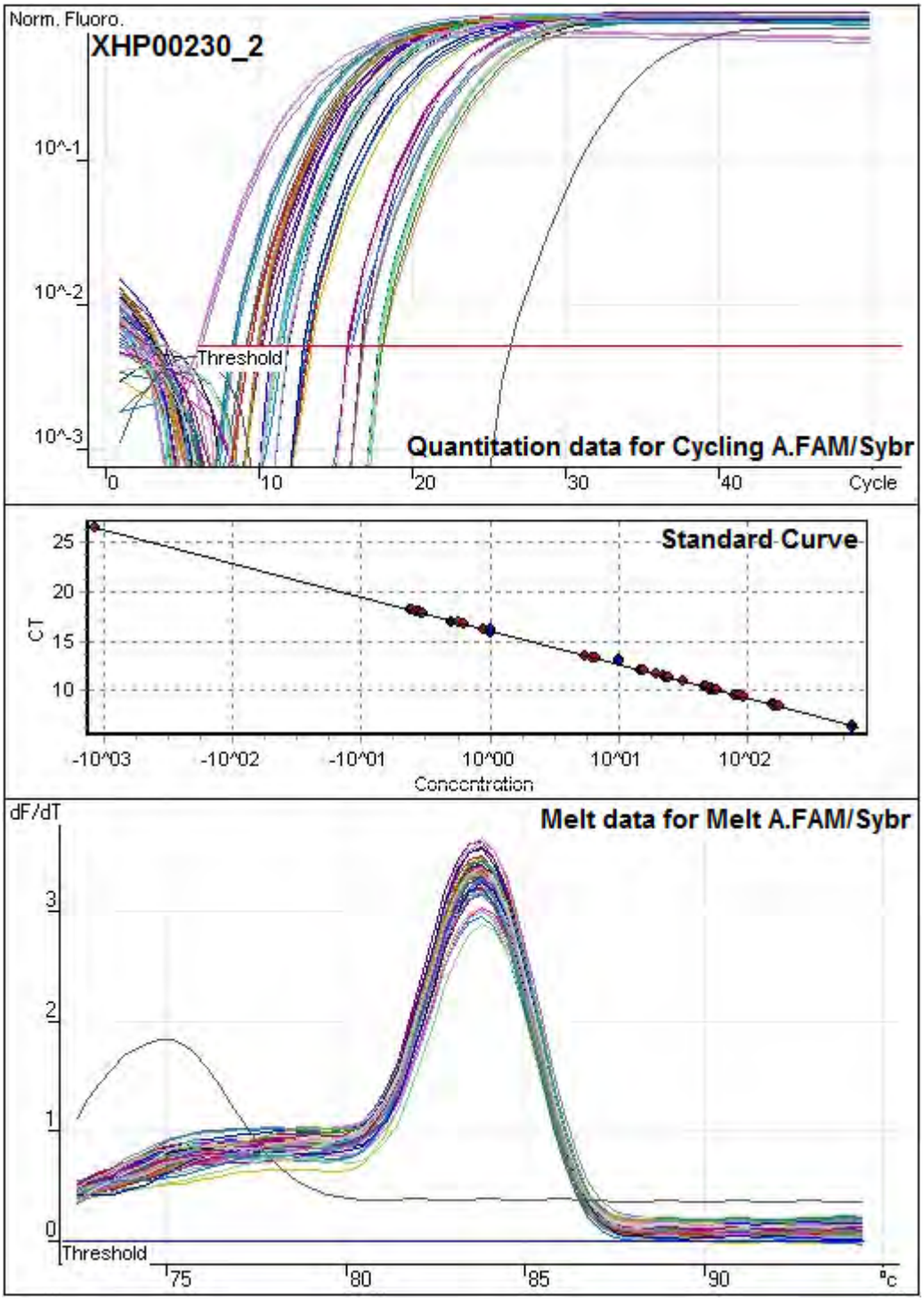
A.4.2.6.


Norm. Fluoro.
XHP00030_10
10^-1
10^-2
10^-3
Threshold
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
30
25
20
15
10
5
CT
Standard Curve
-10^04
-10^03
-10^02
-10^01
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
20
15
10
5
0
Threshold
75
80
85
90
°c

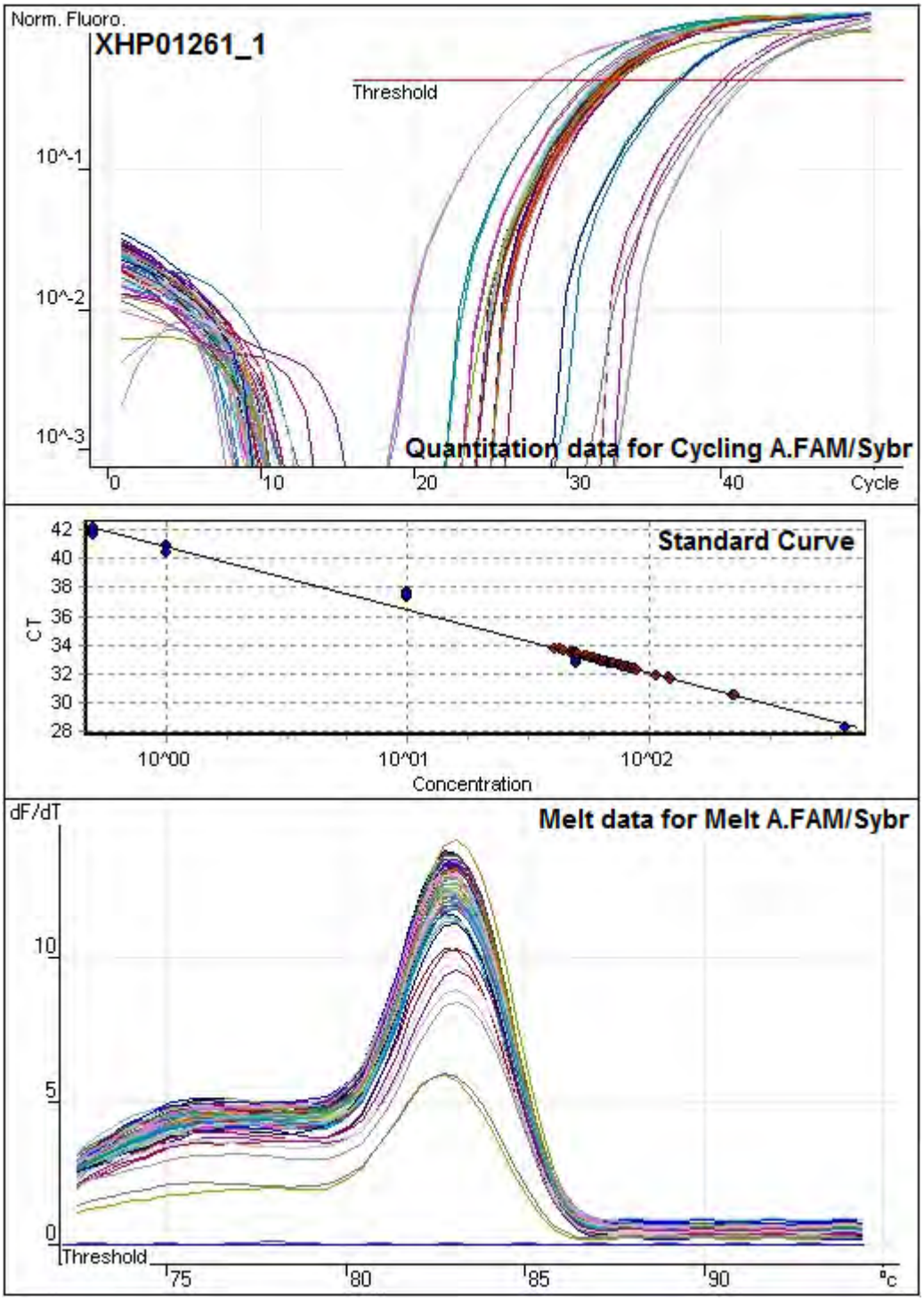
A.4.2.7.


Norm. Fluoro.
XHP00474_1
10^-1
10^-2
10^-3
Threshold
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
35
30
25
20
15
10
CT
Standard Curve
-10^04
-10^03
-10^02
-10^01
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
2
1.5
1
0.5
0
Threshold
75
80
85
90
°c

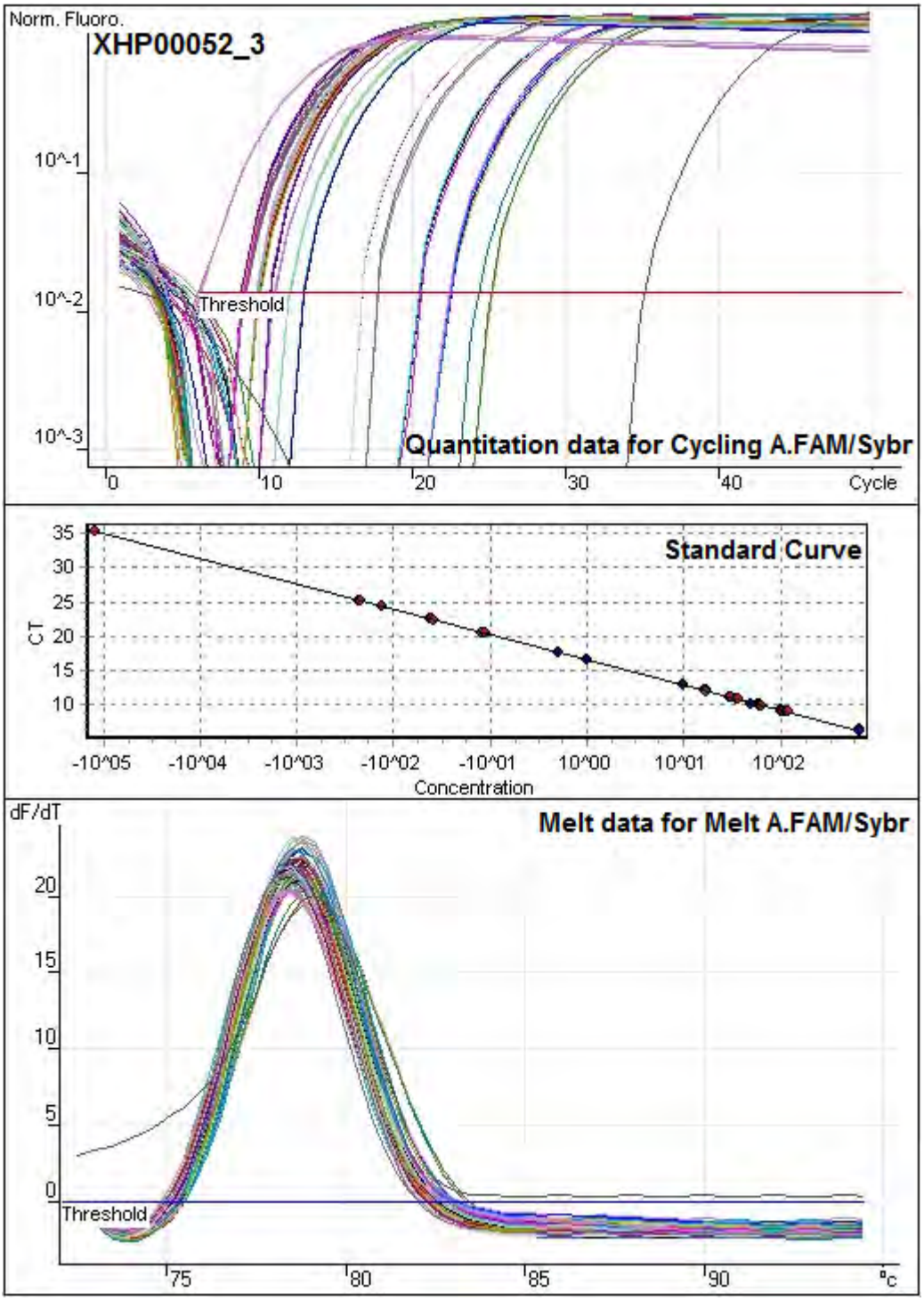
A.4.2.8.


Norm. Fluoro.
XHP00230_2
10^-1
10^-2
10^-3
Threshold
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
Standard Curve
25
20
15
10
CT
-10^03
-10^02
-10^01
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
3
2
1
0
Threshold
75
80
85
90
°c

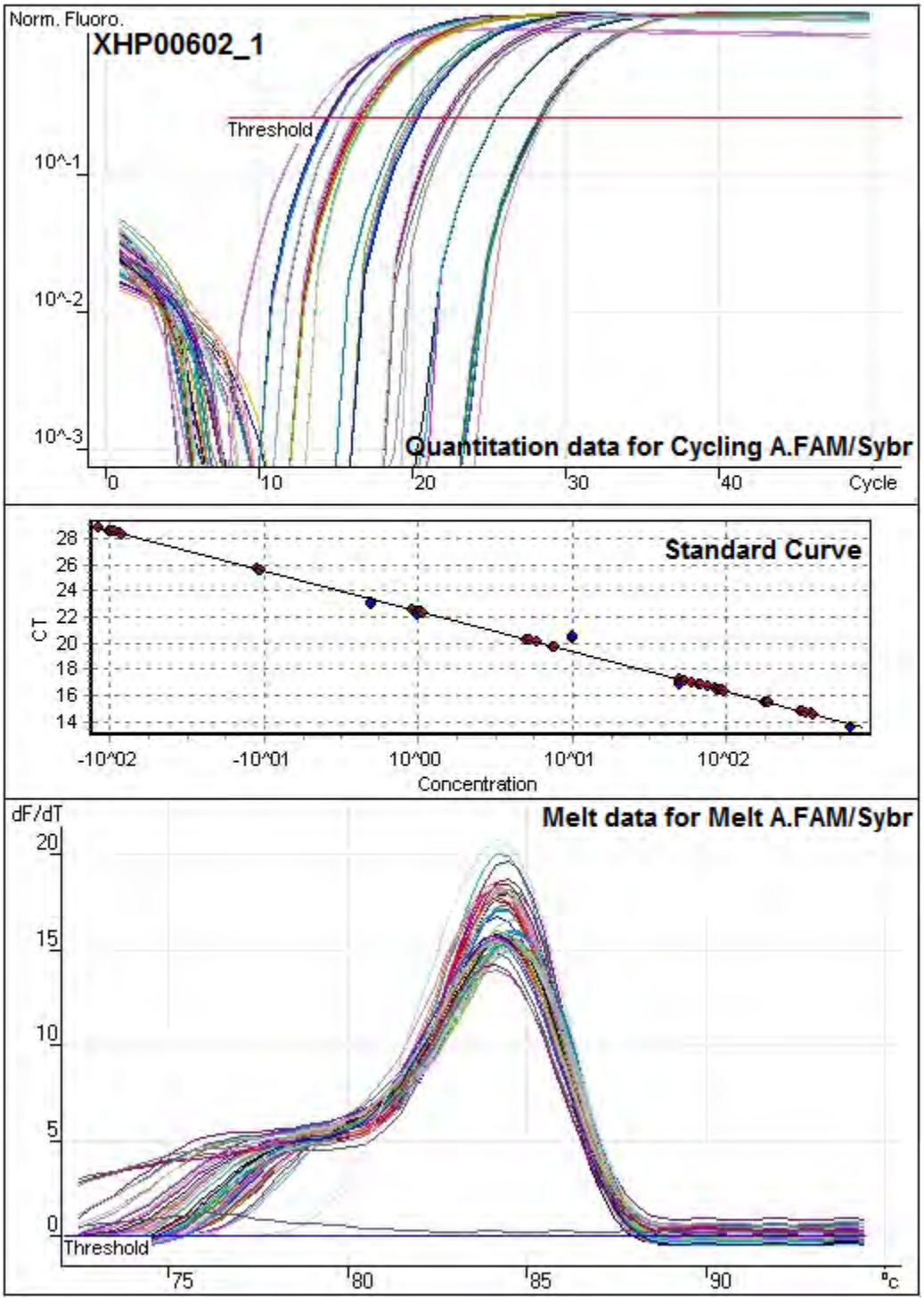
A.4.2.9.


Norm. Fluoro.
XHP01261_1
Threshold
10^-1
10^-2
10^-3
0
10
20
30
40
Cycle
Quantitation data for Cycling A.FAM/Sybr
42
40
38
36
34
32
30
28
CT
Standard Curve
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
10
5
0
Threshold
75
80
85
90
°c

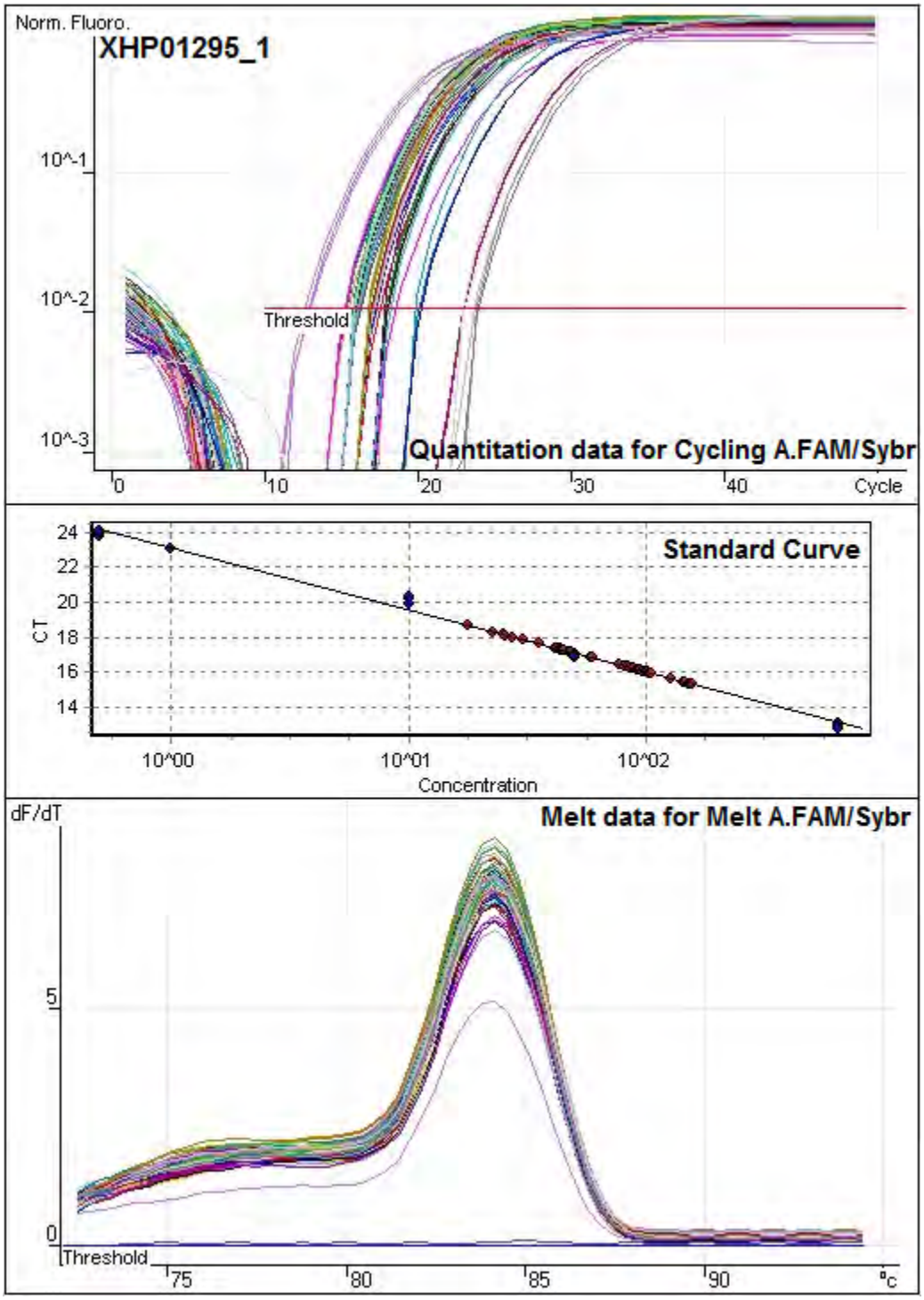
A.4.2.10.


Norm. Fluoro.
XHP00052_3
10^-1
10^-2
10^-3
Threshold
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
Standard Curve
CT
35
30
25
20
15
10
-10^05
-10^04
-10^03
-10^02
-10^01
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
20
15
10
5
0
Threshold
75
80
85
90
°c

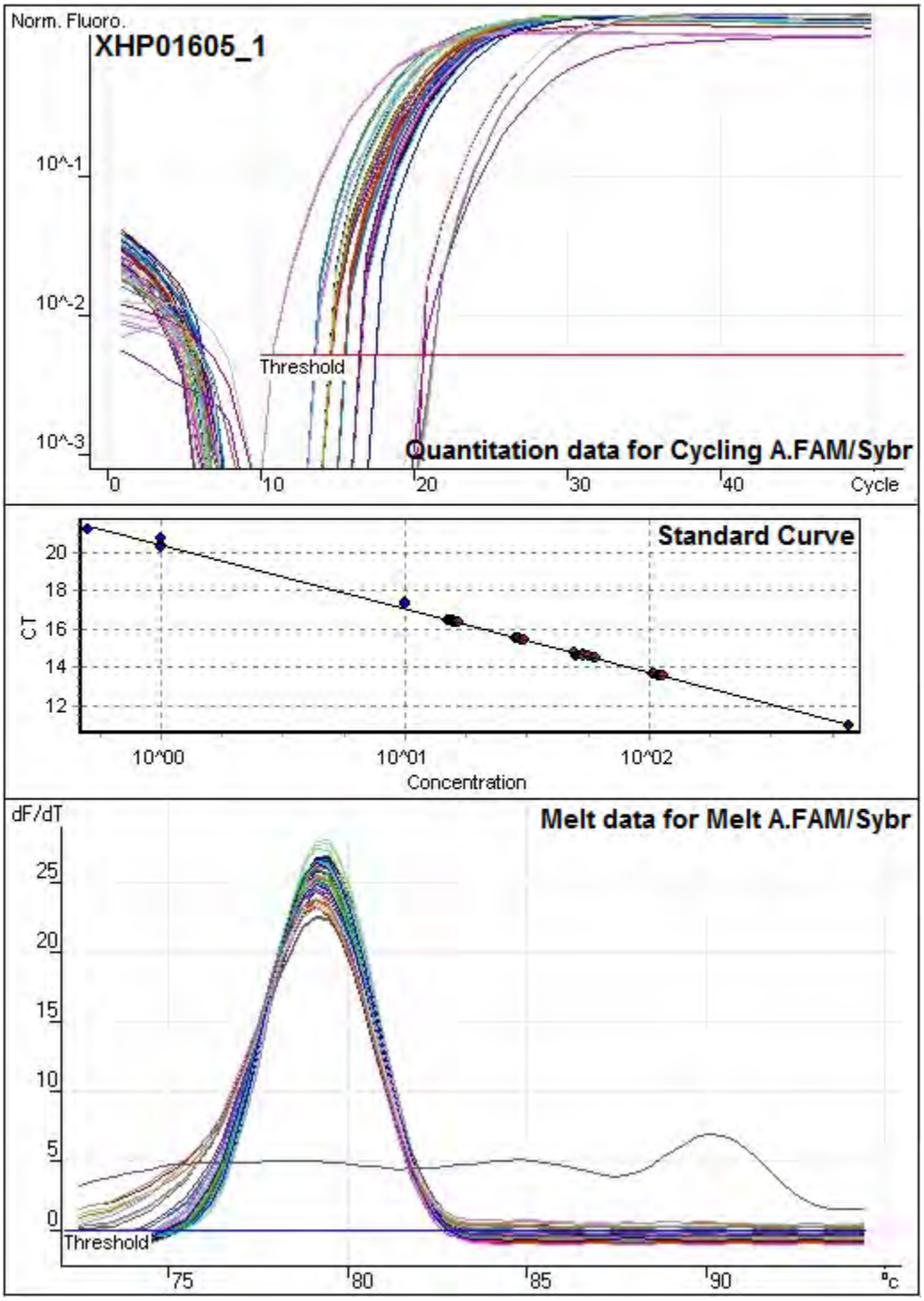
A.4.2.11.


Norm. Fluoro.
XHP00602_1
Threshold
10^-1
10^-2
10^-3
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
Standard Curve
CT
28
26
24
22
20
18
16
14
-10^02
-10^01
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
20
15
10
5
0
Threshold
75
80
85
90
°c

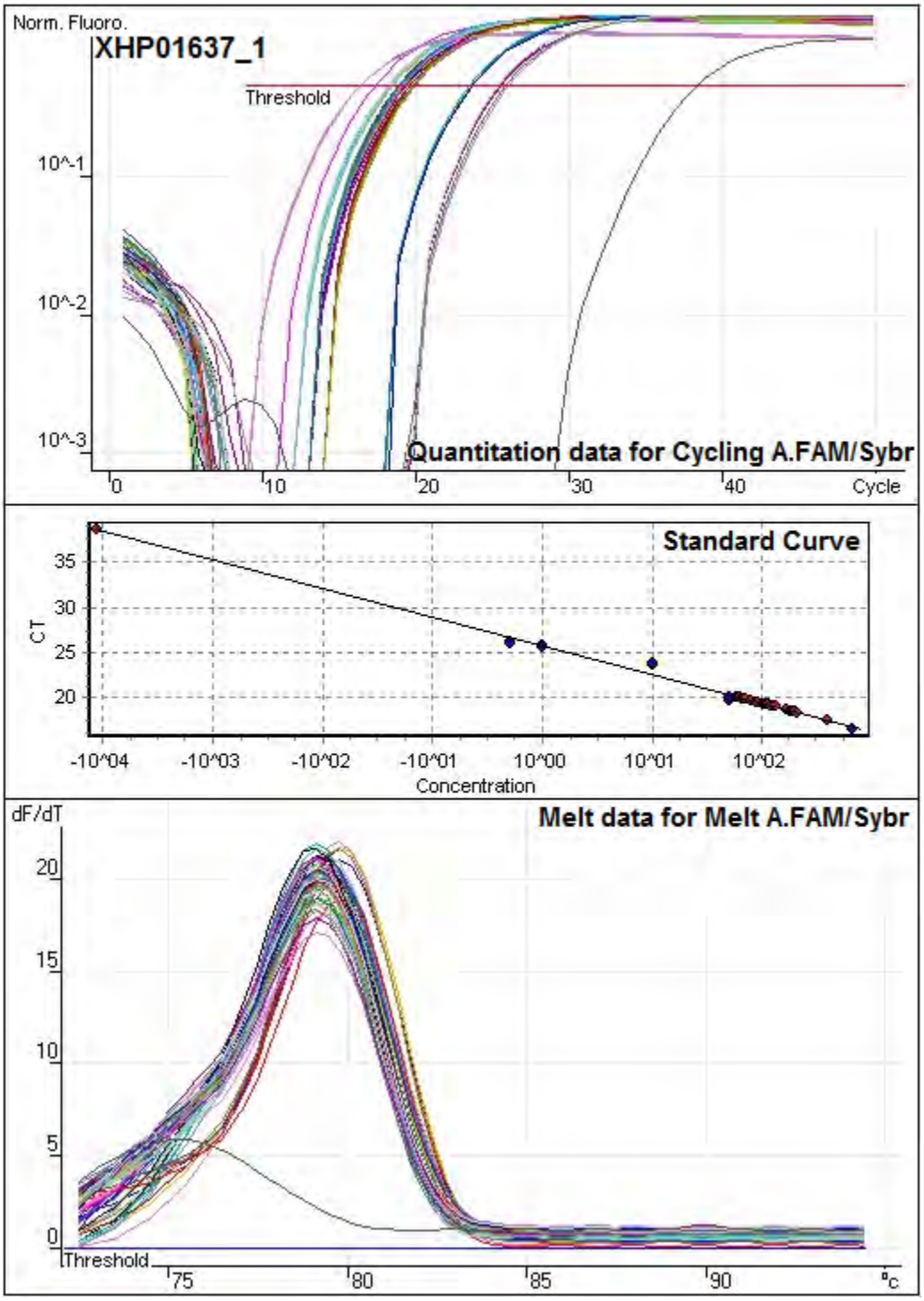
A.4.2.12.


Norm. Fluoro.
XHP01295_1
10^-1
10^-2
10^-3
Threshold
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
24
22
20
18
16
14
CT
Standard Curve
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
5
0
Threshold
75
80
85
90
°c

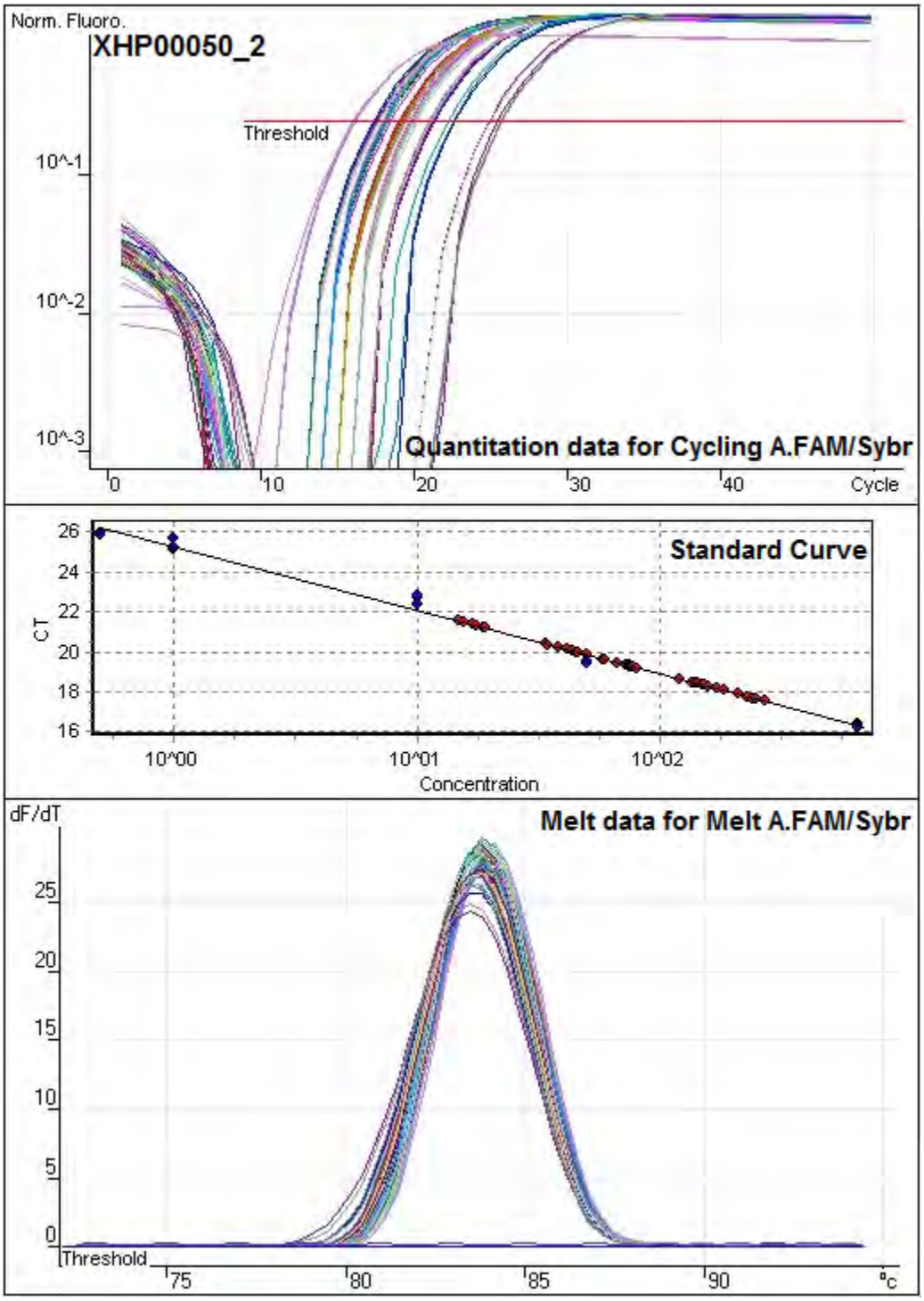
A.4.2.13.


Norm. Fluoro.
XHP01605_1
10^-1
10^-2
10^-3
Threshold
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
Standard Curve
CT
20
18
16
14
12
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
25
20
15
10
5
0
Threshold
75
80
85
90
°C

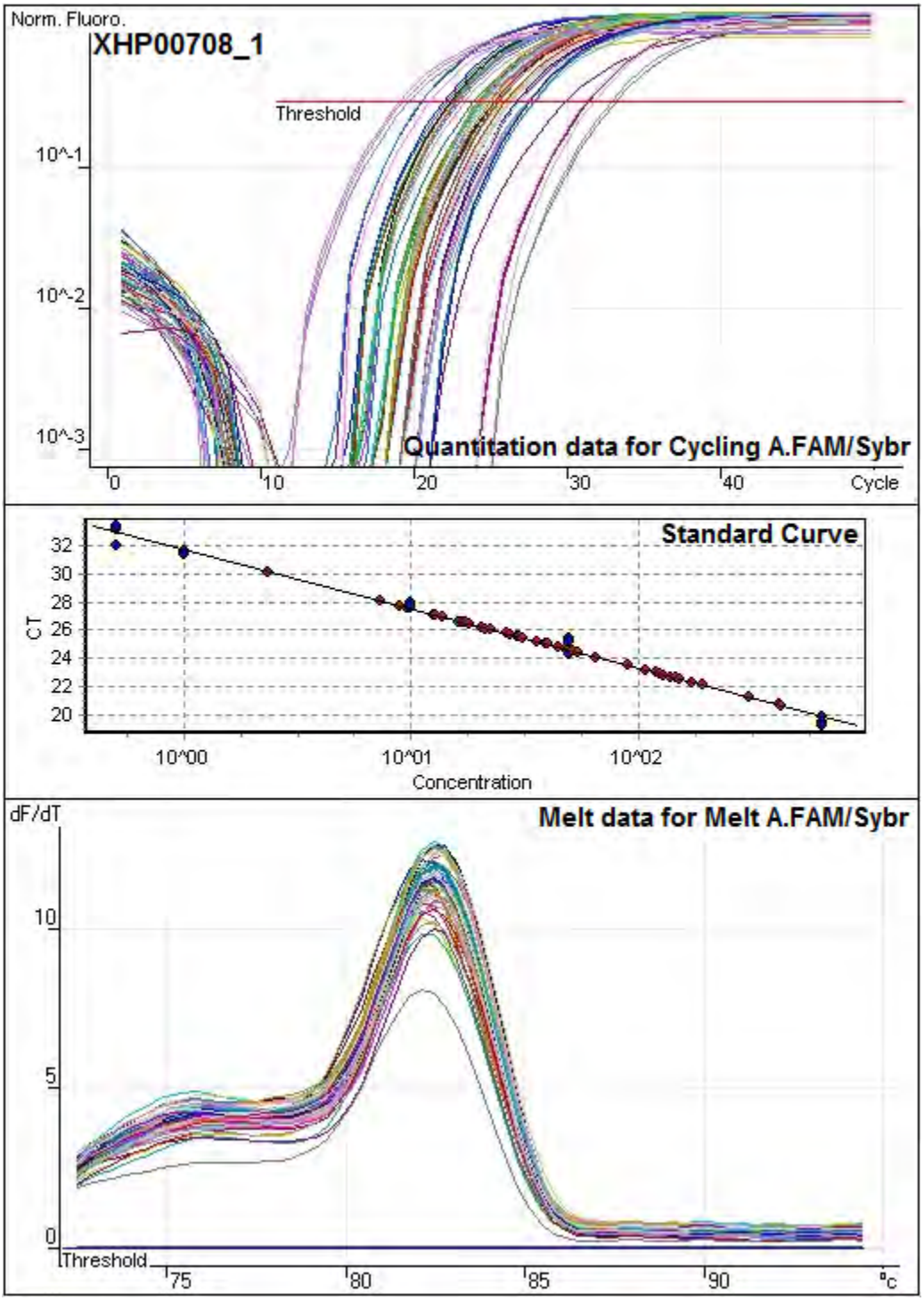
A.4.2.14.


Norm. Fluoro.
XHP01637_1
Threshold
10^-1
10^-2
10^-3
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
Standard Curve
35
30
25
20
CT
-10^04
-10^03
-10^02
-10^01
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
20
15
10
5
0
Threshold
75
80
85
90
°C

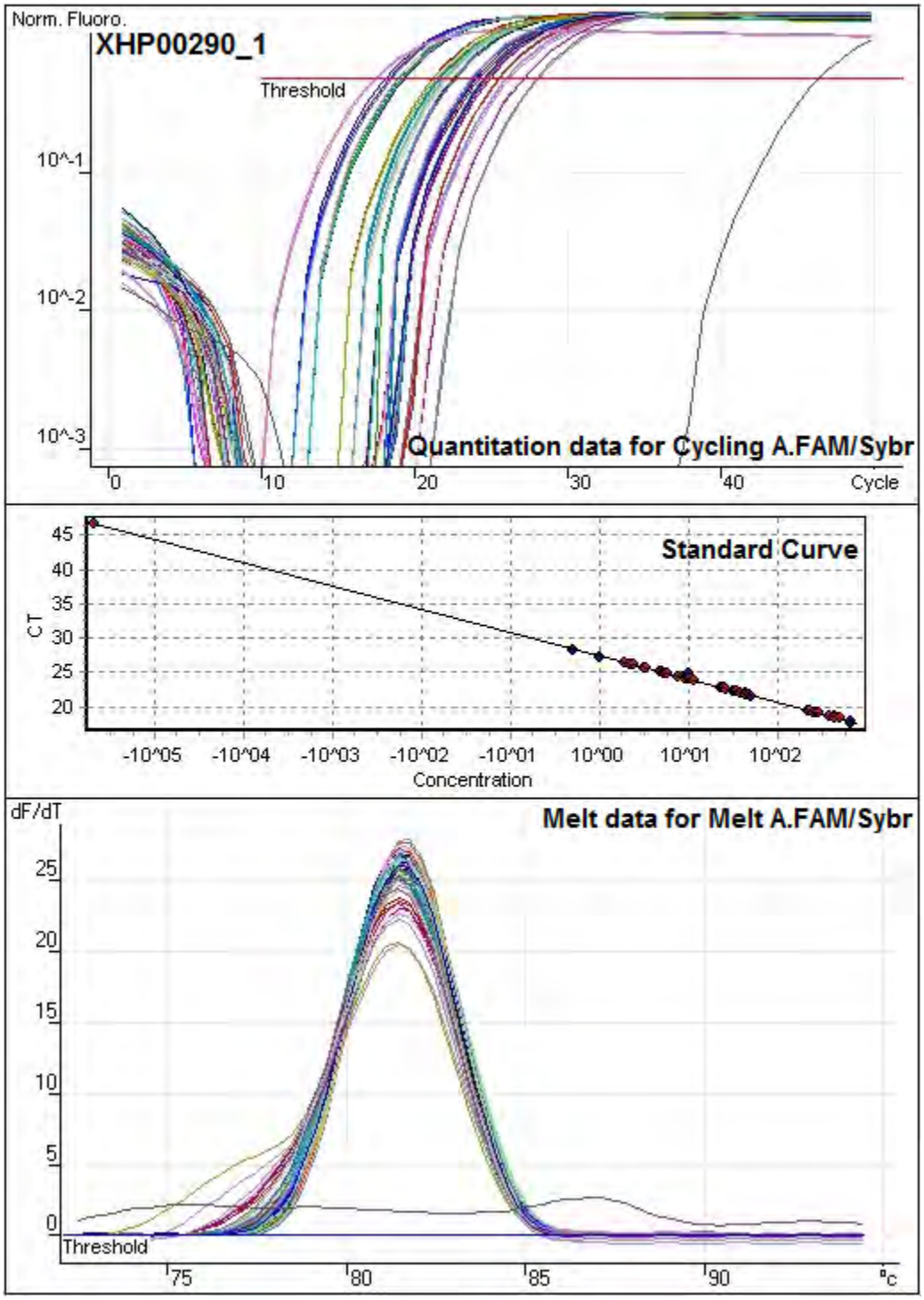
A.4.2.15.


Norm. Fluoro.
XHP00050_2
Threshold
10^-1
10^-2
10^-3
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
26
24
22
20
18
16
CT
Standard Curve
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
25
20
15
10
5
0
Threshold
75
80
85
90
°c

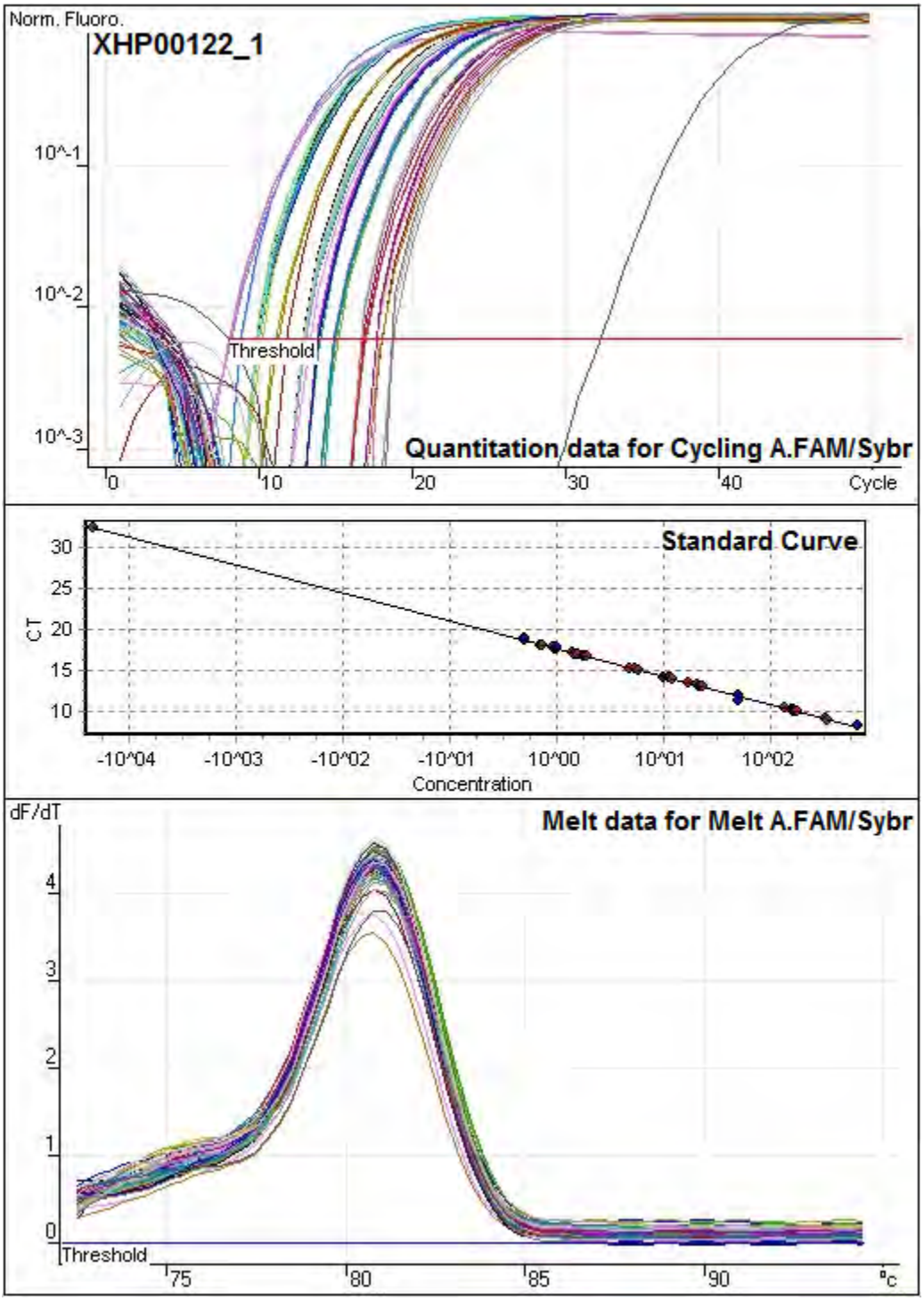
A.4.2.16.


Norm. Fluoro.
XHP00708_1
Threshold
10^-1
10^-2
10^-3
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
Standard Curve
CT
32
30
28
26
24
22
20
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
10
5
0
Threshold
75
80
85
90
°c

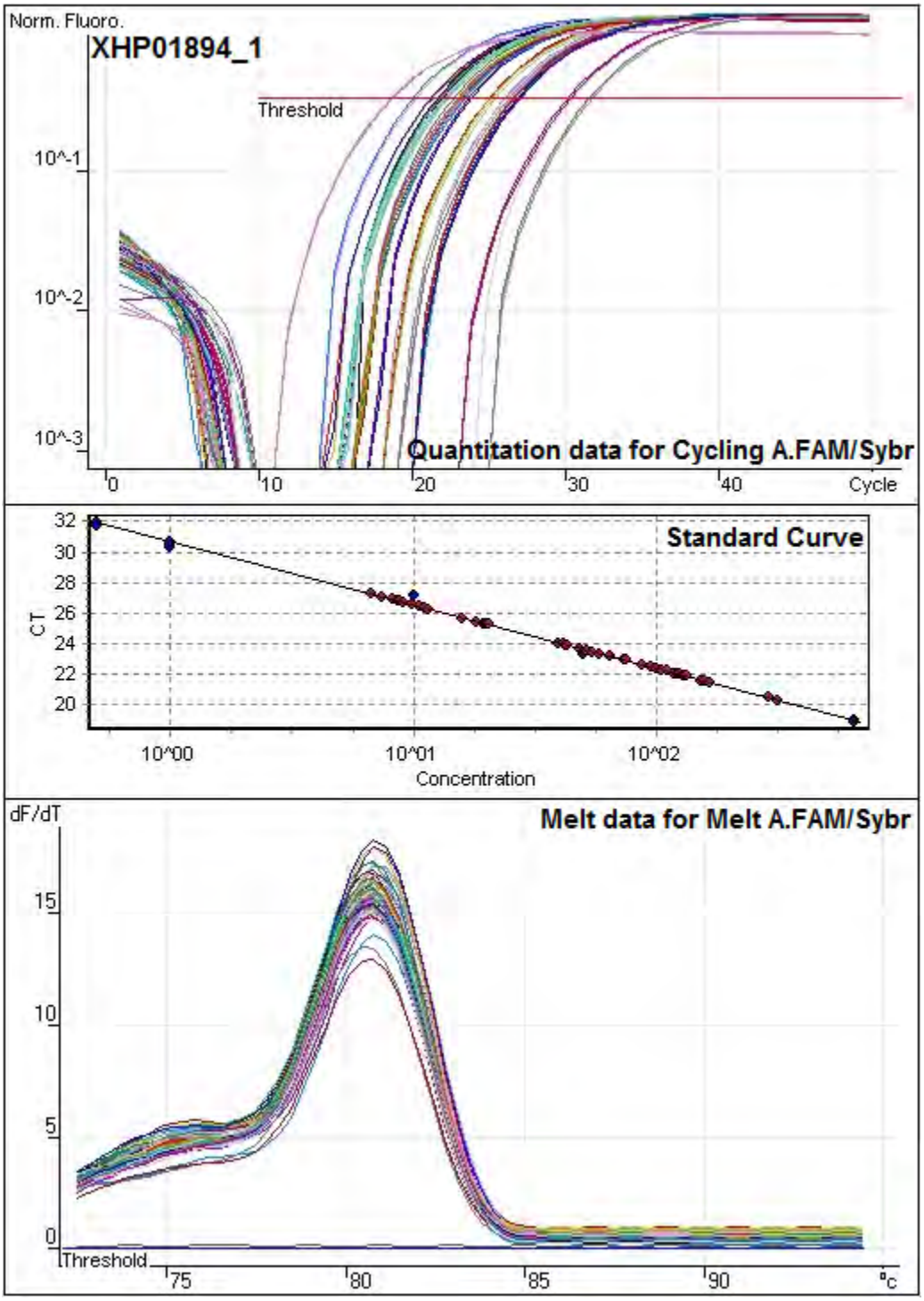
A.4.2.17.


Norm. Fluoro.
XHP00290_1
Threshold
10^-1
10^-2
10^-3
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
Standard Curve
CT
45
40
35
30
25
20
-10^05
-10^04
-10^03
-10^02
-10^01
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
25
20
15
10
5
0
Threshold
75
80
85
90
°c

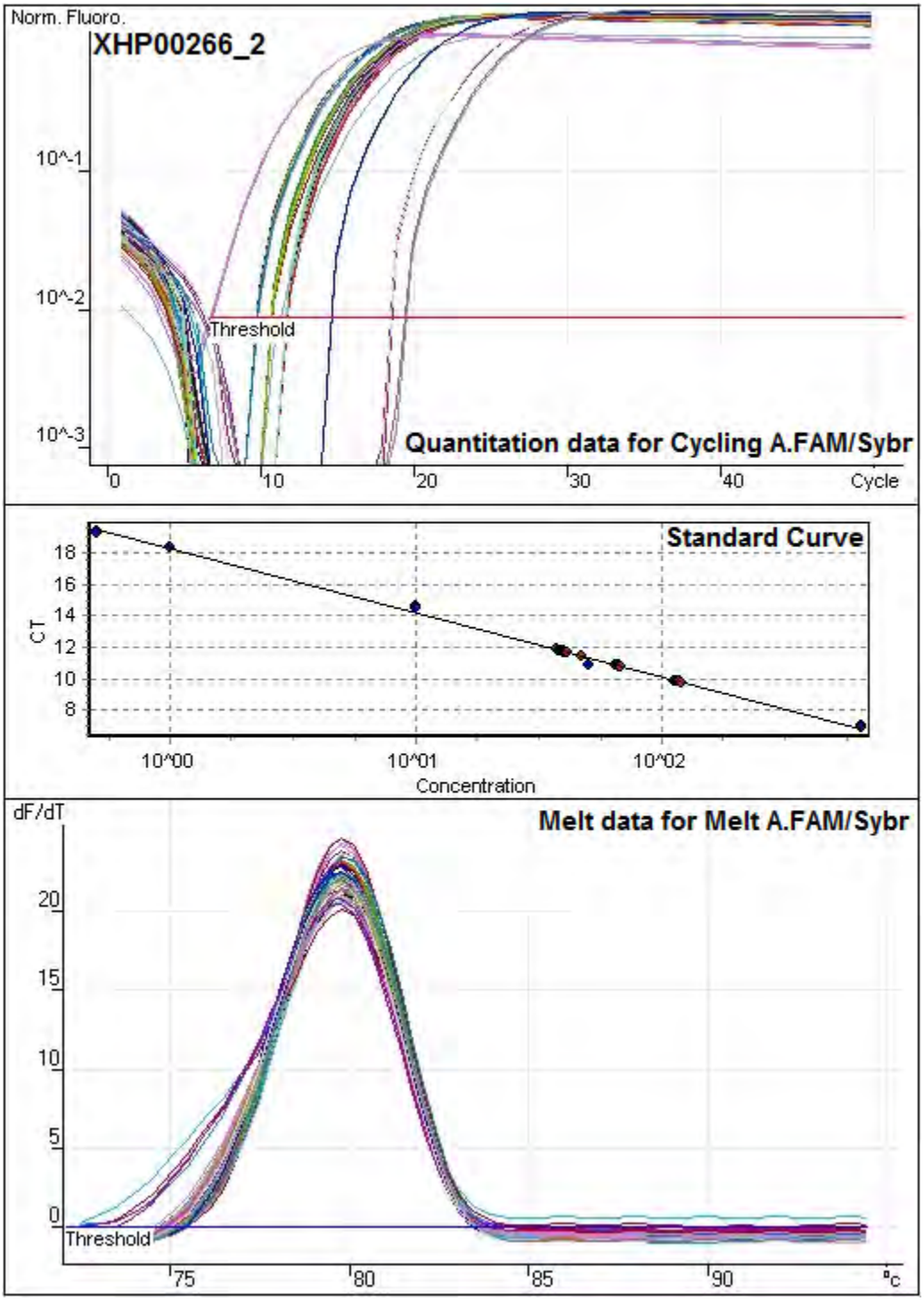
A.4.2.18.


Norm. Fluoro.
XHP00122_1
10^-1
10^-2
10^-3
Threshold
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
Standard Curve
CT
30
25
20
15
10
-10^04
-10^03
-10^02
-10^01
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
4
3
2
1
0
Threshold
75
80
85
90
°c

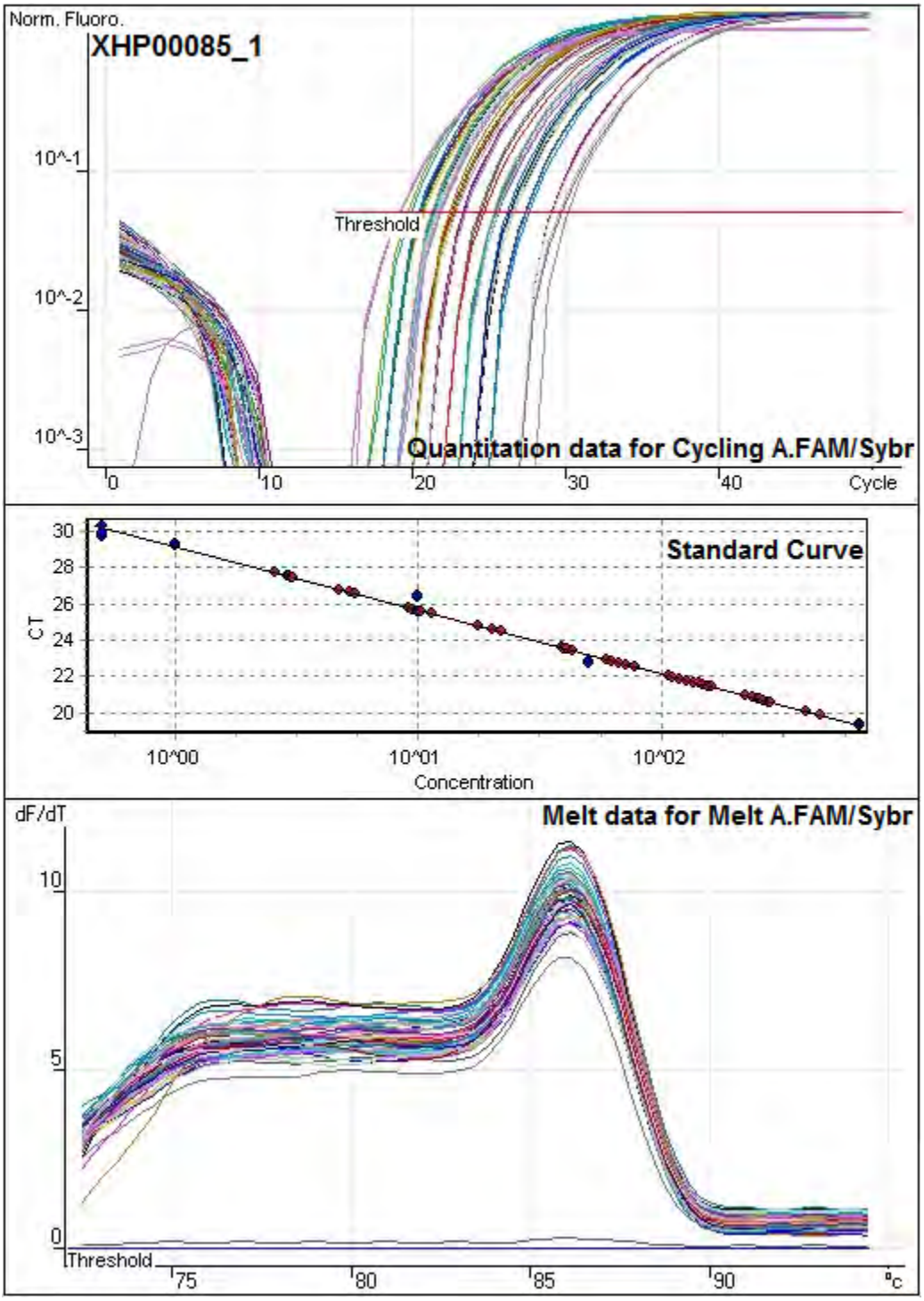
A.4.2.19.


Norm. Fluoro.
XHP01894_1
Threshold
10^-1
10^-2
10^-3
0
10
20
30
40
Cycle
Quantitation data for Cycling A.FAM/Sybr
32
30
28
26
24
22
20
CT
Standard Curve
10^00
10^01
10^02
Concentration
dF/dT
15
10
5
0
Threshold
75
80
85
90
°c
Melt data for Melt A.FAM/Sybr

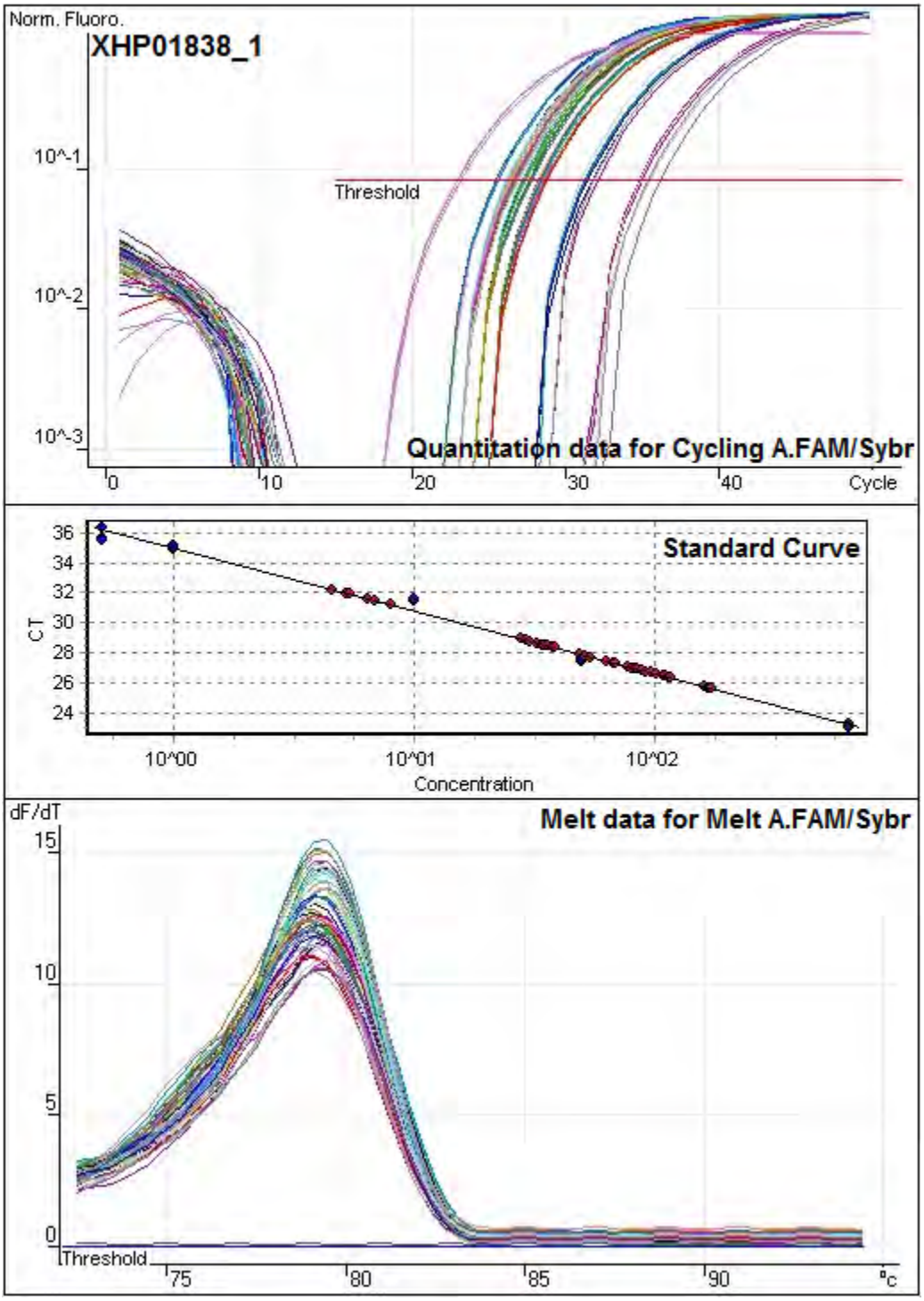
A.4.2.20.


Norm. Fluoro.
XHP00266_2
10^-1
10^-2
10^-3
Threshold
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
Standard Curve
CT
18
16
14
12
10
8
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
20
15
10
5
0
Threshold
75
80
85
90
°c

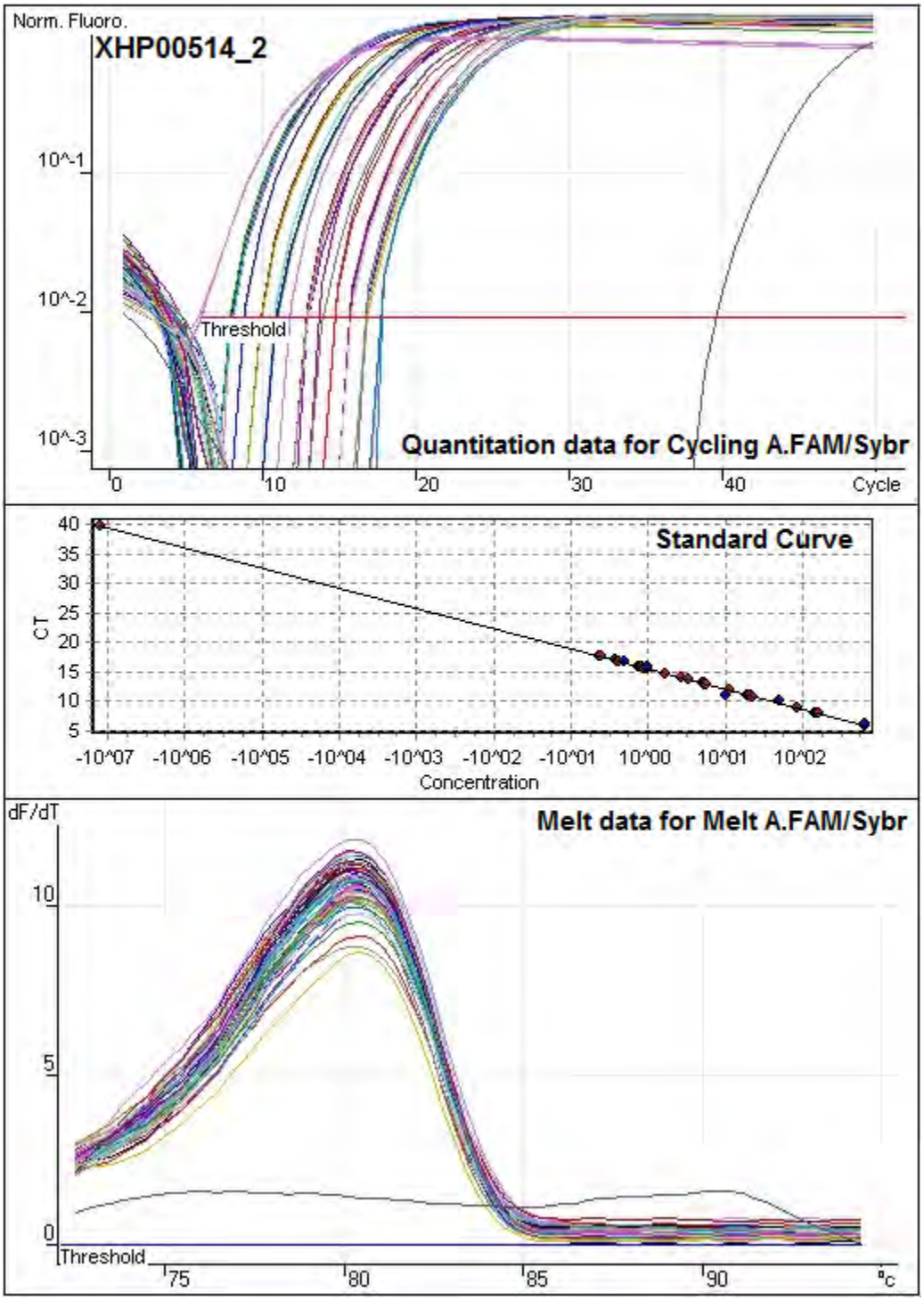
A.4.2.21.


Norm. Fluoro.
XHP00085_1
10^-1
10^-2
10^-3
Threshold
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
30
28
26
24
22
20
CT
Standard Curve
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
10
5
0
Threshold
75
80
85
90
°c

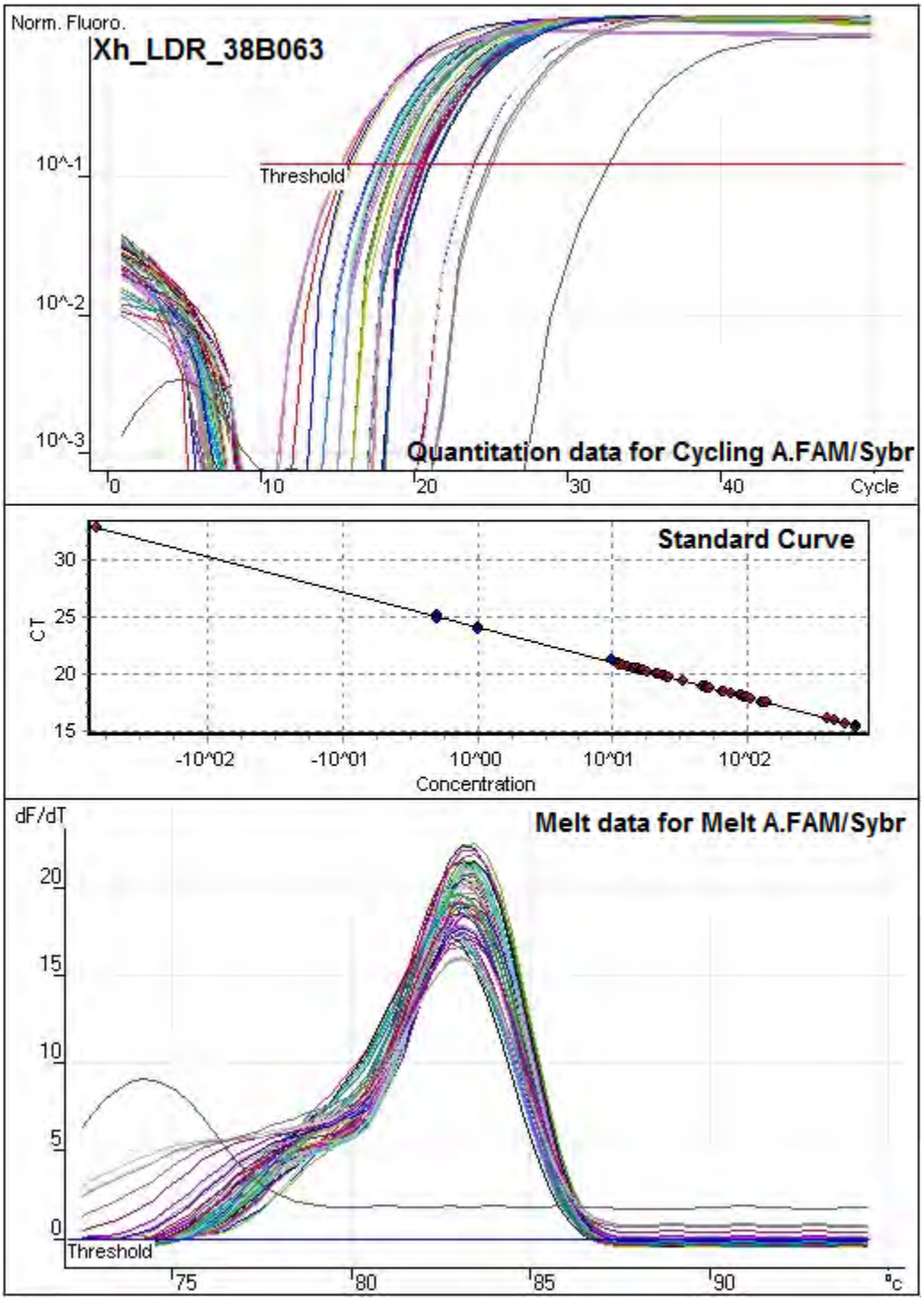
A.4.2.22.


Norm. Fluoro.
XHP01838_1
10^-1
10^-2
10^-3
Threshold
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
36
34
32
30
28
26
24
CT
Standard Curve
10^00
10^01
10^02
Concentration
dF/dT
15
10
5
0
Threshold
Melt data for Melt A.FAM/Sybr
75
80
85
90
°c

A.4.2.23.


Norm. Fluoro.
XHP00514_2
10^-1
10^-2
10^-3
Threshold
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
Standard Curve
CT
40
35
30
25
20
15
10
5
-10^07
-10^06
-10^05
-10^04
-10^03
-10^02
-10^01
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
10
5
0
Threshold
75
80
85
90
°c

A.4.2.24.


Norm. Fluoro.
Xh_LDR_38B063
10^-1
10^-2
10^-3
Threshold
Quantitation data for Cycling A.FAM/Sybr
0
10
20
30
40
Cycle
Standard Curve
CT
30
25
20
15
-10^02
-10^01
10^00
10^01
10^02
Concentration
dF/dT
Melt data for Melt A.FAM/Sybr
20
15
10
5
0
Threshold
75
80
85
90
°C

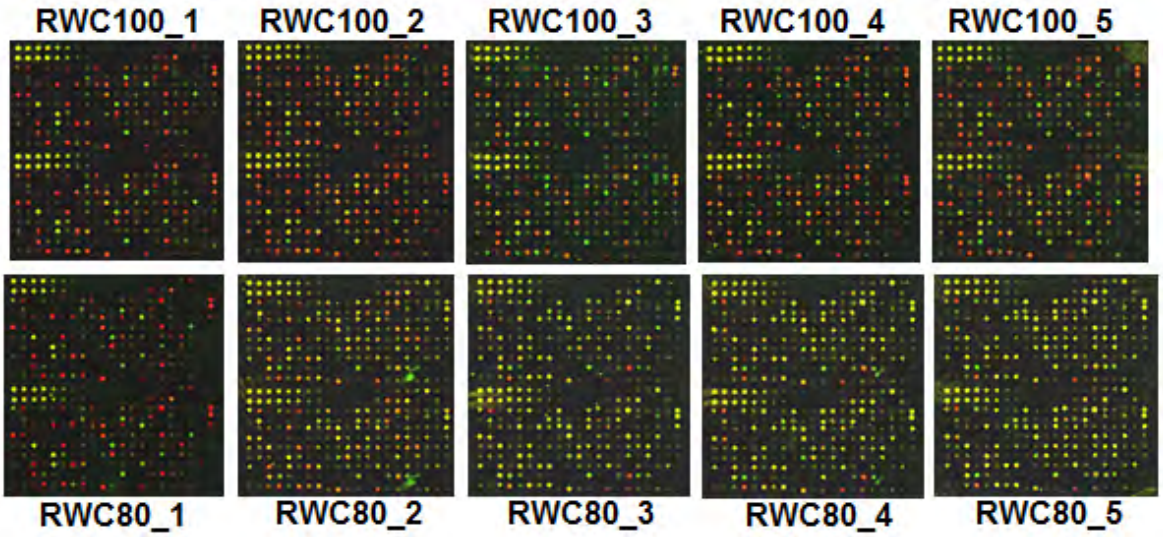


Figure A.4.3. Scanned images of microarray slides representing the 5 biological replicates of !00% and 80%RWC samples. Amplified mRNA transcripts isolated from leaf samples were labeled with Cy3 fluorescent dyes (green), and the common reference pool samples were labeled with Cy5 (red). Block 3 features of each slide image is enlarged and shown.

Table A.4.4. Efficiency and R^2 value of RT-qPCR reaction.

Contig ID	Contig description	PAMSAM cluster	Reaction efficiency	R^2
XHP00087_2	CCR2 (COLD, CIRCADIAN RHYTHM, AND RNA BINDING 2)	NDE	0.89	0.99
Xh_RRF_02C098	Translation initiation factor SUI1 family protein	NDE	0.99	0.99
XHP00531_2	dhn9	6	1.22	0.99
XHP00196_1	mitochondrial import inner membrane translocase subunit	6	1.03	0.98
XHP01603_1	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family protein	6	1.01	0.99
XHP00030_10	RAB18 (RESPONSIVE TO ABA 18)	5	0.93	0.99
XHP00474_1	unknown protein	5	0.90	0.99
XHP00230_2	unknown protein	5	0.96	0.99
XHP01261_1	LEA domain-containing protein	5	0.69	0.99
XHP00052_3	LEA14	4	0.87	0.99
XHP00602_1	low-molecular-weight cysteine-rich 68	4	1.12	0.98
XHP01295_1	dual specificity protein phosphatase-related	4	0.91	0.99
XHP01605_1	adenylate cyclases	1	0.99	0.99
XHP01637_1	S phase kinase-associated protein 1	1	1.07	0.97
XHP00050_2	UBQ4; protein binding	1	1.07	0.99
XHP00708_1	Lhca2 protein	2	0.73	0.99
XHP00290_1	Transketolase	2	0.98	0.99
XHP00122_1	photosystem II light harvesting complex gene 2.3	2	0.96	0.99
XHP01894_1	PS II oxygen-evolving complex 1	2	0.74	0.99
XHP00266_2	UBQ10 (POLYUBIQUITIN 10); protein binding	3	0.76	0.99
XHP00085_1	galactinol synthase 4	3	0.93	0.99
XHP01838_1	no match	3	0.74	0.99
XHP00514_2	RuBisCO	3	0.97	0.98
Xh_LDR_38B063	unknown protein	7	1.11	0.99

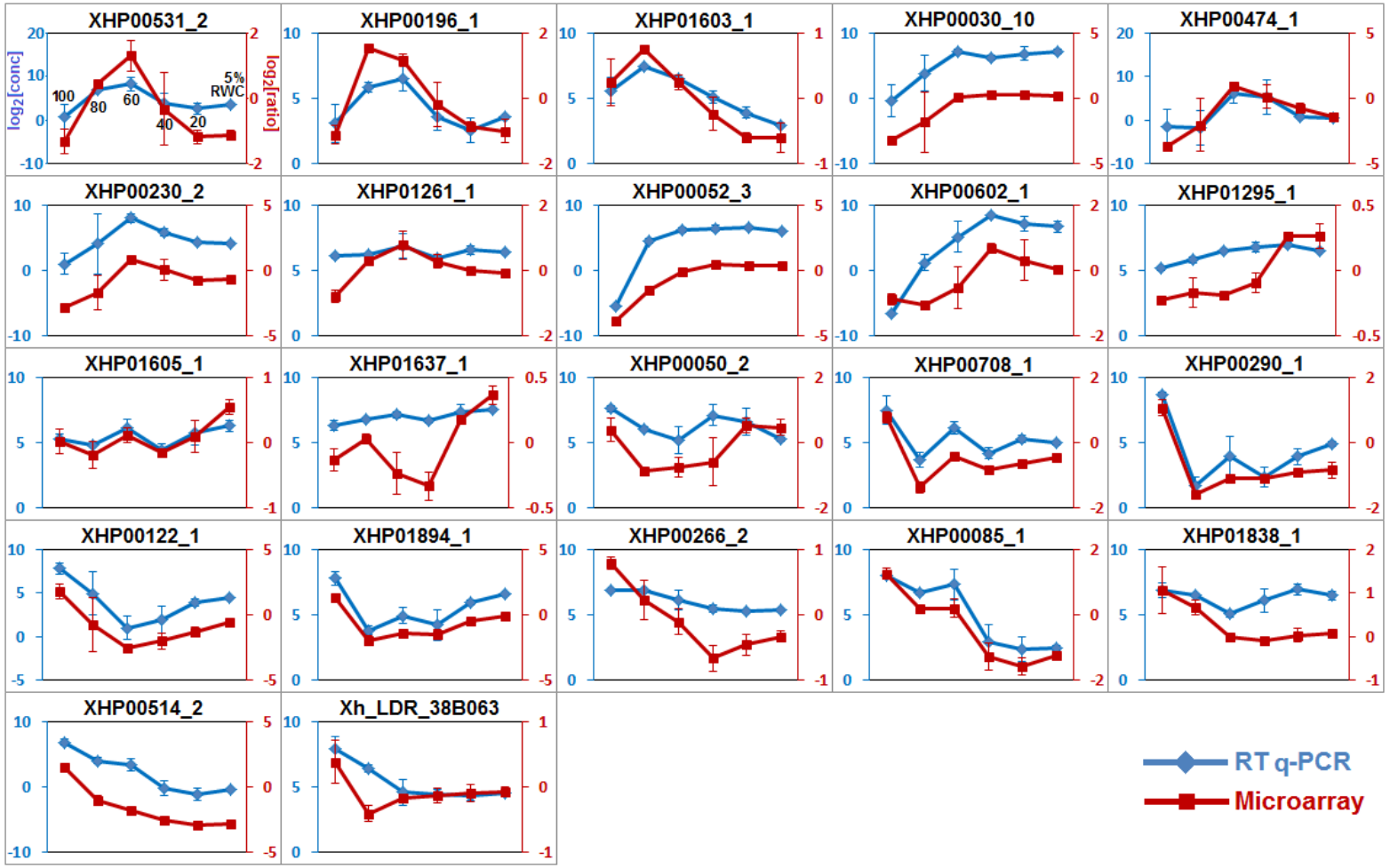


Figure A.4.5. Validation of *X. humilis* microarray expression data by RT-qPCR analysis. Transitional changes in mRNA transcript abundance of 22 contigs in *X. humilis* leaves (two biological replicates) at 6 different stages of water loss during desiccation assessed by microarray analysis was validated by RT-qPCR analysis normalized against cDNA input amount. Microarray expression values were represented as the averaged $\log_2$ [Cy3/Cy5] from the two biological replicates of leaf samples analyzed (red y-axis), and RT-qPCR expression values were represented as averaged $\log_2$ [contig] (blue y-axis). Error bars represented the expression values identified in the two biological replicates of the leaf samples. Contig descriptions can be found in Appendix (Table A.4.4).

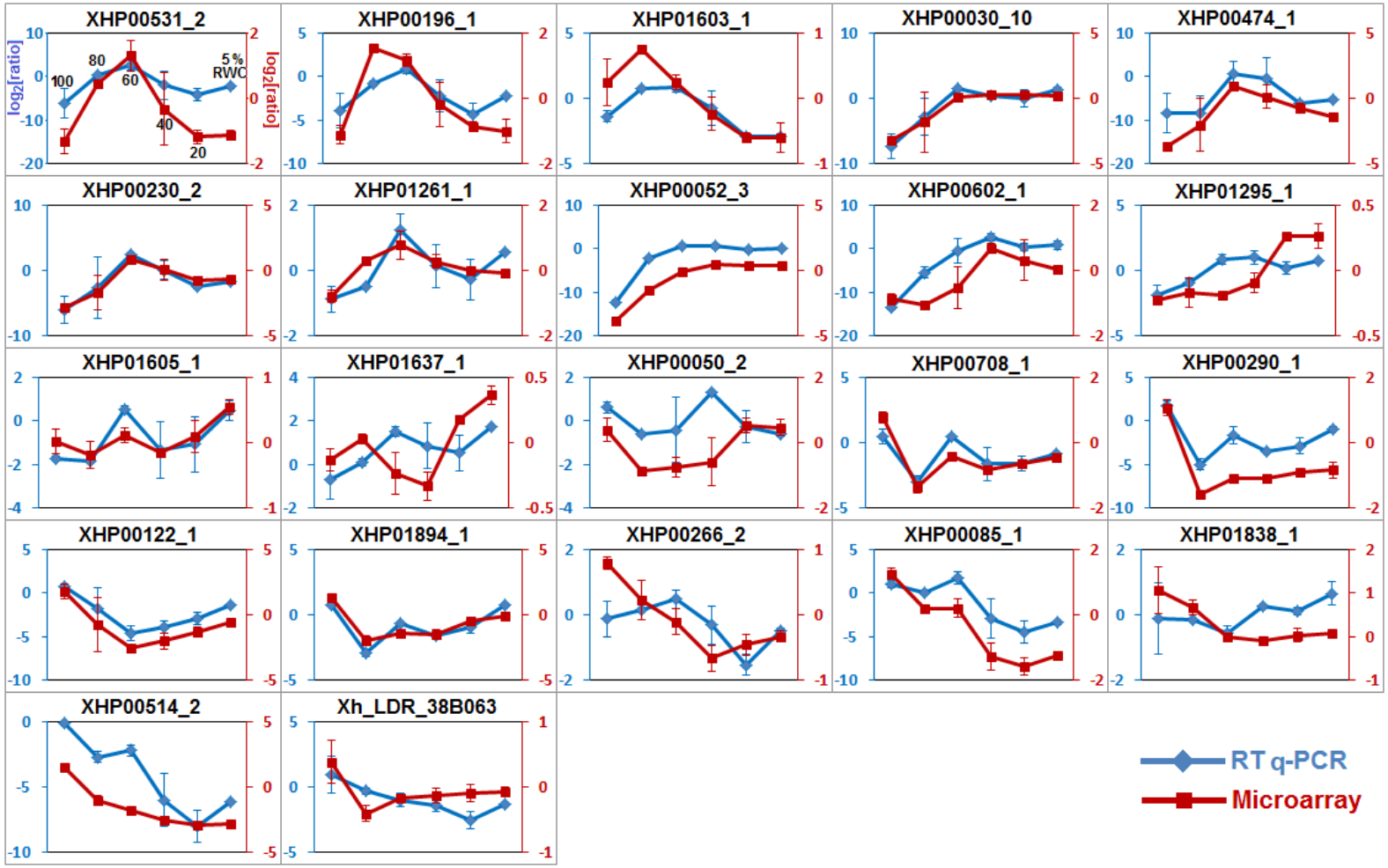


Figure A.4.6. Validation of *X. humilis* microarray expression data by RT-qPCR analysis. Transitional changes in mRNA transcript abundance of 22 contigs in *X. humilis* leaves (two biological replicates) at 6 different stages of water loss during desiccation assessed by microarray analysis was validated by RT-qPCR analysis normalized against expression of XHP00087_2. Microarray expression values were represented as the averaged $\log_2$ [Cy3/Cy5] from the two biological replicates of leaf samples analyzed (red y-axis), and RT-qPCR expression values were represented as averaged $\log_2$ [ratios of contig concentration over that of XHP00087_2] (blue y-axis). Error bars represented the expression values identified in the two biological replicates of the leaf samples. Contig descriptions can be found in Appendix (Table A.4.4).