Transcriptomic analysis of Douglas-fir megagametophyte development and abortion

by

Ian Boyes

B.Sc., University of Victoria, 2009

A Thesis Submitted in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Biology

Transcriptomic analysis of Douglas-fir megagametophyte development and abortion

by

Ian Boyes

B.Sc., University of Victoria, 2009

Supervisory Committee

Dr. Patrick von Aderkas, Co-Supervisor

(Department of Biology)

Dr. Jürgen Ehlting, Co-Supervisor

(Department of Biology)

Dr. Steve Perlman, Departmental Member

(Department of Biology)

Supervisory Committee

Dr. Patrick von Aderkas, Co-Supervisor
(Department of Biology)

Dr. Jürgen Ehlting, Co-Supervisor
(Department of Biology)

Dr. Steve Perlman, Departmental Member
(Department of Biology)

ABSTRACT

Douglas-fir develops a megagametophyte regardless of the pollination state of the ovule, whereas many other conifers develop a megagametophye in response to pollination. Megagametophytes in unfertilized ovules degrade two weeks following fertilization of the surrounding population. This is mediated by programmed cell death (PCD). Pollinated and unpollinated megagametophytes were dissected from Douglas-fir cones and extracted for RNA, which was then used as input for sequencing. A transcriptome was assembled from this data and expression levels were calculated. The data were fitted to quadratic regressions to produce coexpression groups. There is no clear upregulation of PCD effectors in the unpollinated megagametophyte. Potential regulators of megagametophyte fate are present in the data. Some are associated with ABA signalling and proanthocyanadin biosynthesis while others share similarity to known regulators of PCD. Seed development processes are represented

in the expression data, which support current knowledge of conifer seed development and provide targets for research.

Contents

List of Tables

List of Figures

List of Abbreviations

ABC	ATP-binding cassette
AGO1	Argonaute 1
Apaf-1	apoptotic protease activation factor 1
BAM	binary alignment/map format
BLAST	Basic local alignment search tool
CHS	chalcone synthase
CTAB	cetyltrimethylammonium bromide
CUC	CUP-SHAPED COTYLEDON
DCL3	dicer-like 3
DSEL	DAD1-like seedling establishment-related lipase
FBW2	F-box with WD-40 2
HPLC	High performance liquid chromatography
HSP	heat shock protein
JA	jasmonic acid

JAR1	jasmonate resistant 1
LDOX	leucoanthocyanidin dioxygenase
LEA	late embryogenesis abundant protein
LMI2	late meristem identity
LN	liquid nitrogen
LRP1	lateral root primordia 1
NGS	next-generation sequenceing
PA	proanthocyanidin
PAK2	p21-activated kinase
PCD	Programmed cell death
PDAT	phospholipid diacylglycerol acyltransferase
PINK	PTEN-induced putative kinase
RIP	receptor-interacting protein RIP-1 and RIP-2
RISC	RNA-induced silencing complex
ROS	reactive oxygen species
RT-PCR	Realtime polymerase chain reaction
RuBisCo	Ribulose bisphosphate carboxylase oxygenase
SAM	sequence alignment/map
SHI	short internode

SPS3F	sucrose phosphate synthase 3F
STP7	Sugar transporter 7
STP7	sugar transporter 7
TE	tracheary element
TLP	thaumatin-like protein
TNF	tumour necrosis factor

ACKNOWLEDGEMENTS

I would like to thank:

Dr. Patrick von Aderkas, for giving me perspective and guiding me to clarity.

Dr. Jürgen Ehlting, for feeding my scientific imagination and coming to lab beers.

Dr. Steve Perlman, for always keeping tabs on my emotional well-being.

Dr. Stefan Little, for lending his wisdom and emotional support.

Kate Donaleshen and Julia Gill, for making my summer days at work happily bearable.

Dr. Belaid Moa and Westgrid, for his knowledge and patience and for the CPU time.

Lan Tran and Coung Hieu Le, for endlessly commiserating with me.

Julia Rudko, for loving me even when I'm in dire straits.

My parents, for supporting me when my thesis turned me into a child again.

Brett Nelson and Chris Bennett, for always being ready for beer when I needed it.

Gary Moore, Stevie Ray Vaughn, and Jeff Healey, for getting me through the last weeks.

Chapter 1

Introduction

1.1 Douglas-fir

1.1.1 The Tree

Douglas-fir (*Pseudotsuga menziesii* Mirbel.) is a monoecious conifer that can be identified by the tridentate bracts on its seed cones. It can grow ninety centimetres per year and reach heights of 100 meters (Vidaković, 1991; Grescoe, 1997). Douglas-fir forests occupy a large range in western North America, extending from southern British Columbia through the western United States. Smith and Darr (2004) reported Douglas-fir forests as covering an area of 144000 km^2 in the United Sates. Canada's national forest information system reports the area occupied by Douglas-fir in Canada to be 48910 km^2.

Douglas-fir is the most commercially valuable species of *Pseudotsuga* (Eckenwalder, 2009). The wood of Douglas-fir is strong, stiff, and often available in long dimensions (Bormann, 1984). These properties make it highly desirable for cultivation and harvest for structural applications. Its popularity in the British Columbia logging

industry led to the rapid reduction of the original old-growth Douglas-fir forests in the coastal plain.

Douglas-fir is extensively cultivated in Europe. It was first introduced to the United Kingdom in 1827 by botanist David Douglas with a seed lot he collected himself (Eckenwalder, 2009). Great Britain now possesses 452 km^2 of Douglas-fir forest (Smith and Gilbert, 2003). France has the largest area of Douglas-fir with 4000 km^2 (IFN, 2008), which is over twice the size of Germany's inventory of 1800 km^2. Many other European countries including the Netherlands, Belgium, Italy, Portugal, and Spain have over 50 km^2 (Hermann and Lavender, 1999).

1.1.2 Douglas-fir Reproduction

Douglas-fir cone development, fertilization, and seed development occurs over two seasons. In the first year, buds are initiated on lateral shoots in late spring (Owens and Smith, 1964; Allen and Owens, 1972). Microsporangia form predominantly on the proximal half of the shoot, while megasporangia are formed primarily on distal half of the lateral shoot. While numerous buds may be initiated, they can also be aborted or enter a latent state. The number of mature cones produced is dependent on the rate of bud abortion rather than the number of buds initiated (Owens, 1969). The immature buds grow over the course of the first year and become dormant in late November or early December before resuming development in mid-February (Allen and Owens, 1972).

Ovuliferous scales are the site of ovule development. Each scale supports two ovules, which are oriented towards the center of the cone and 1.1). A leaf-derived bract is

pressed against the outer surface of the scale (Figure 1.2. Megaspore mothercells within the scale undergo meiosis to produce four haploid megaspores each. One will become dominant and the other three will degenerate (Allen and Owens, 1972). The dominant megaspore undergoes a series of nuclear divisions forming a large coenocyte that is bounded by a megaspore wall. Subsequent formation of cell walls in the coenocyte produces unicellular prothallial cells (von Aderkas et al., 2005a), which then divide prolifically to form a mass of cells called the megagametophyte.

Some prothallial cells begin forming archegonia in the micropylar end of the megagametophyte in early May (Figure 1.3) (Allen and Owens, 1972). The prothallial cells divide to produce a layer of neck cells at at the base of the gametophyte and a large central cell extending into the gametophyte (Figure 1.4B) (Owens et al., 1991). The neck cells are the site of entry of the pollen tube into the egg cell (Fernando et al., 1998). Division of the central cell produces a small ventral cell, adjacent to the neck cells, and one large egg cell (Chiwocha and von Aderkas, 2002). The mature egg cell has a large nucleus with many mitochondria in the perinuclear region (Owens and Morris, 1990). The rest of the megagametophyte is composed of many small, thin-walled cells.

Pollination commences in early to mid-April with the release of pollen grains from the male microsporangium. Prior to pollination, growth in the female cone causes the bracts to open, allowing passage of the pollen to surface of the scale. At the micropylar end of the ovule, the stigmatic tip passively collects pollen, which is then drawn into the micropyle (Owens et al., 1981). The pollen germinates three weeks later and continues to grow towards the nucellus for six weeks (Owens and Morris, 1990). Two male gametes develop at the growing end of the gametophyte. When the pollen reaches the nucellus, localized cellular degradation occurs in the nucellus,

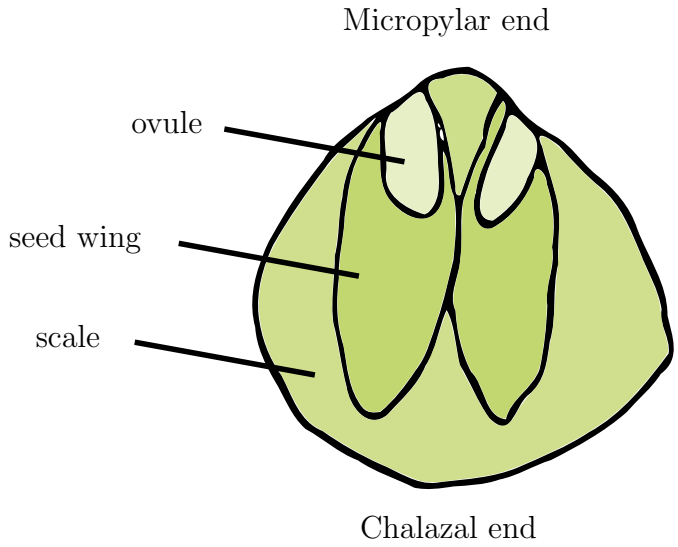


Figure 1.1: The surface of the Douglas-fir scale that is tightly appressed to the cone axis. The scale is affixed to the cone at its end nearest to the micropyle, which is oriented towards the apex of the tree. Pollen enters the ovule through the micropyle.

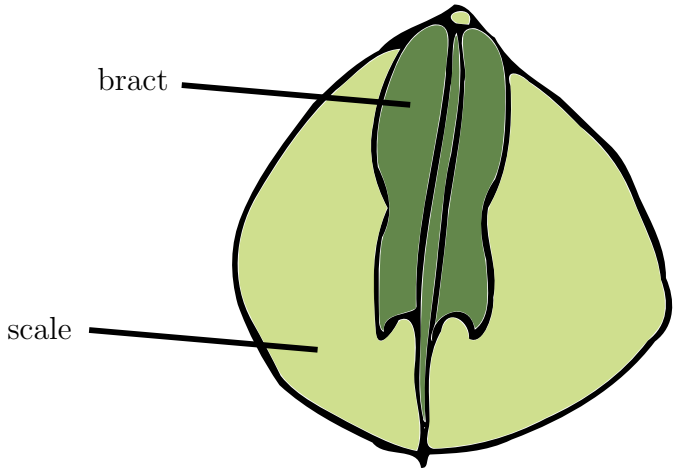


Figure 1.2: The surface of the Douglas-fir scale facing out from the cone axis. The tridentate bract is in direct contact with the scale surface.

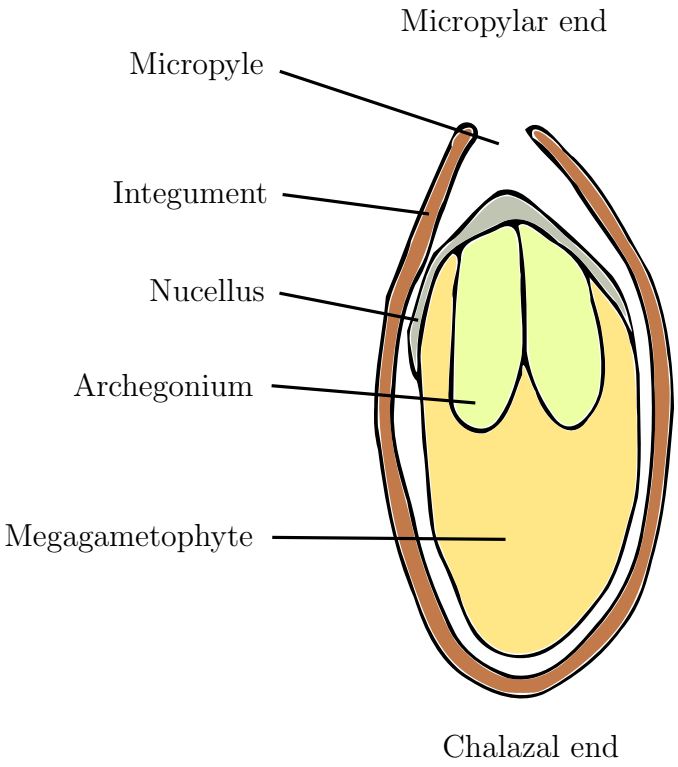
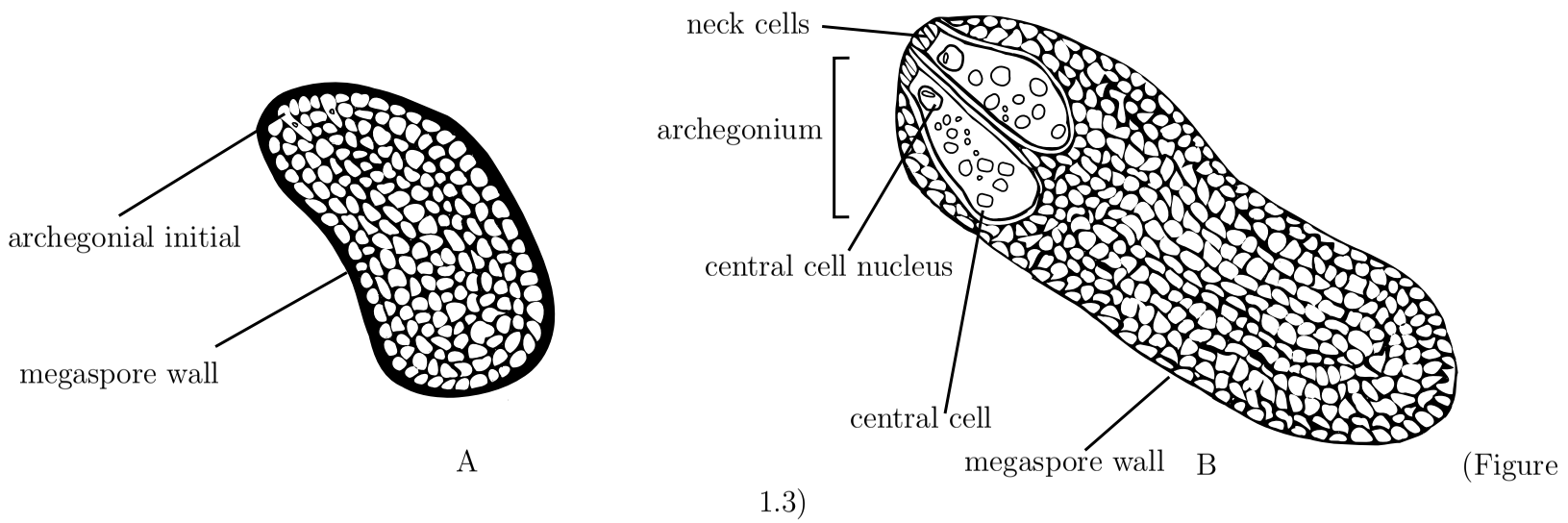


Figure 1.3: An archegonia-bearing ovule. Pollen has entered the micropylar canal and the entrance to the micropyle is sealed. The nucellus is a thick layer surrounding the micropylar end of the megagametophyte. No corrosion cavity has formed yet.



(Figure 1.3)

Figure 1.4: The progress of the development of the A) nascent archegonia (archegonial initials) into central-cell containing archegonia. Migration of the nucleus to the micropylar end of the archegonium B) is followed by the formation of the neck cell and a vacuolate central cell. The egg arises from the central cell (Chiwocha and von Aderkas, 2002; Owens et al., 1993)

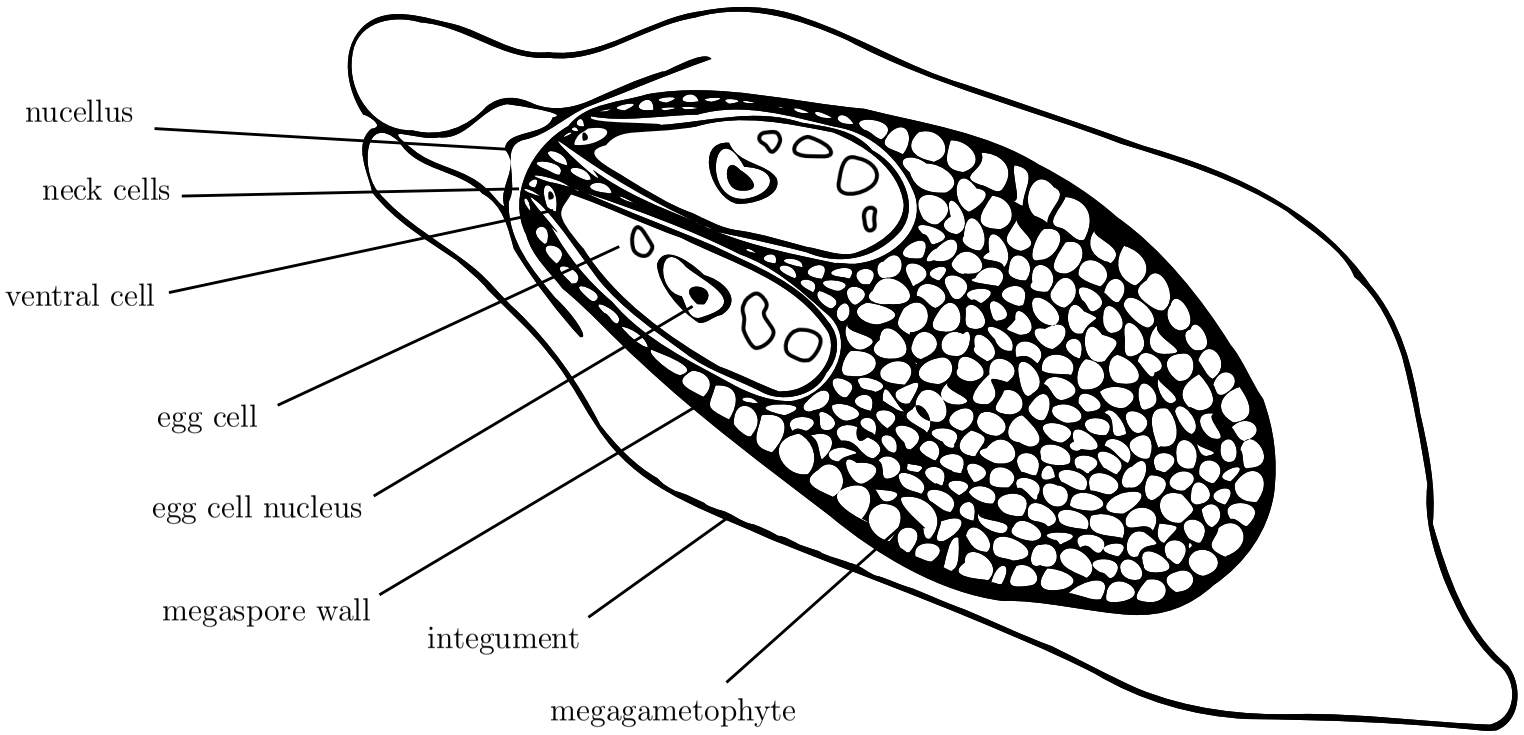


Figure 1.5: An ovule ready for fertilization. The central cell has divided to produce the ventral cell and the egg cell. When the pollen tube enters the archegonium, it will have to pass through the nucellus and neck cell to reach the nucleus. After fertilization embryo development will depend on the nutritive capacity of the megagametophyte.

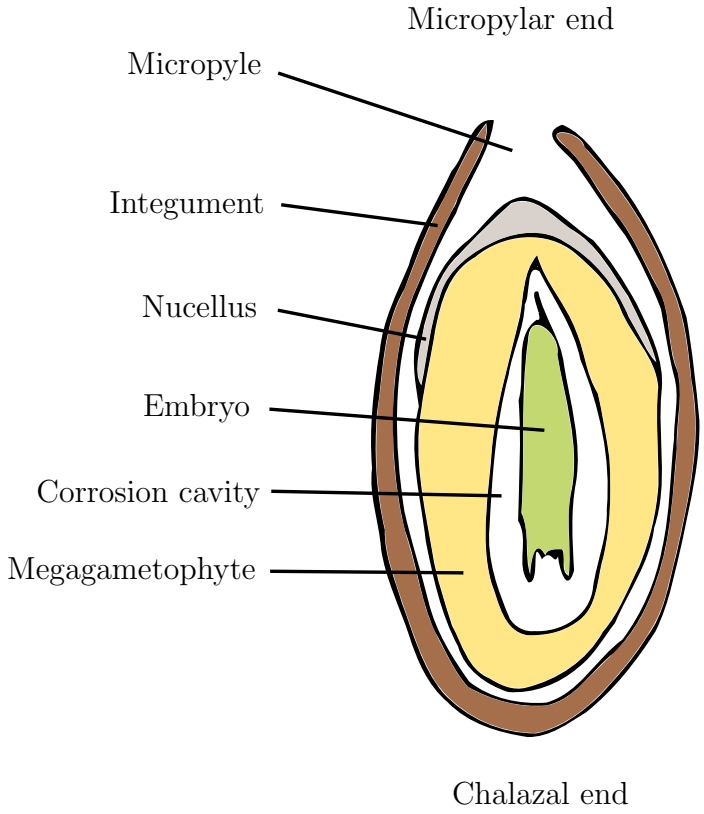


Figure 1.6: An embryo-bearing ovule. The archegonia given way to the advancing embryo. A corrosion cavity has formed to house the embryo.

facilitating pollen tube penetration (Owens and Morris, 1990). Sperm are released into the egg cytoplasm and they migrate to the nucleus (Fernando et al., 1998). Fertilization occurs in late May to mid-June, six to ten weeks after pollination (von Aderkas et al., 2005b).

1.1.3 Embryogenesis

Embryogenesis includes three anatomical stages: the proembryo, the early embryo, and the late embryo. During proembryogeny, nuclear divisions occur at the basal end of the zygote to form tiers of cells that will constitute the suspensor and embryo (Allen and Owens, 1972). Elongation of the suspensor cells forces the proembryo from the archegonium into the megagametophyte (Chiwocha and von Aderkas, 2002). Early embryogenesis consists of rapid growth the embryonic cells and elongation of the suspensor, which pushes the embryo further into the corrosion cavity, a fluid-filled space in the center of the megagametophyte (Figure 1.6. The body plan of the embryo develops during late embryogenesis. As the megagametophyte reaches maturity it becomes heavily loaded with lipid and protein bodies (Owens et al., 1993), giving the megagametophyte a white-yellow colour. Like the endosperm of flowering plants, these storage products provide nutrition to the seedling following germination.

1.1.4 Seed Abortion

The reproductive process in Douglas-fir does not always produce a viable seed. Common reasons for seed loss are a ovular abortion due to an absence of fertilization or developmental problems such as selfing (Owens et al., 1991). In *Picea* Mill., *Pinus* L., and *Thuja* L., eggs only develop if pollination occurs. Conversely, Douglas-fir develops egg cells and a megagametophyte regardless of the pollination status of the ovule (Rouault et al., 2004). If fertilization does not occur, the ovule aborts

approximately two weeks after fertilization would have occurred (von Aderkas et al., 2005b).

Megagametophyte abortion occurs by coordinated programmed cell death (PCD). PCD occurs within the megagametophyte at other points in its development. Before the embryo can develop, a corrosion cavity must form in the megagametophyte. In Scots pine (*Pinus sylvestris* L.), this forms by PCD characterized by cell rupture and the release of intracellular material into the corrosion cavity (Vuosku et al., 2009). After fertilization, multiple embryos can form because Douglas-fir ovules develop four archegonia. One of these embryos becomes dominant while the others degrade (Chiwocha and von Aderkas, 2002). Multiple mature embryos developing within one seed would pose problems due to the limited space in the ovule and limited seed reserves. Filonova et al. (2002) have demonstrated that the survival of the dominant embryo is supported by PCD of the subordinate embryos. PCD is integral to seed storage mobilization during germination in white spruce (*Picea glauca* Moench). The process displays hallmarks of plant PCD including internucleosomal DNA fragmentation, intracellular vacuolation, and caspase-like protease activity (He and Kermode, 2003a,b). Megagametophyte abortion requires a similar mobilization of nutrients, which are absorbed by the tree rather than the embryo and seedling.

1.2 Programmed Cell Death

Programmed cell death is the intentional suicide of a cell in a multicellular organism as a result of an internal or external stimulus. It is essential to development and immune system function in multicellular organisms. PCD has been studied intensively in animal systems and the study of PCD in plants has grown dramatically in the past two decades. The general functions of PCD are shared by animals and plants. In

animals, cancerous cells and virus-infected cells are removed through apoptosis induced by cytotoxic lymphocytes (Thompson, 1995). Viral and bacterial proliferation in plants often triggers localized PCD, called a hypersensitive response (Coll et al., 2010). PCD is also important in development in both animals and plants. In animal development apoptosis is responsible for sculpting of organs and tissues or sex-specific or stage-specific deletion of structures (Fuchs and Steller, 2011). Plants use PCD to remove cells during reproductive development (Vuosku et al., 2009) and during development of vegetative structures such as aerenchyma (Schussler and Longstreth, 2000).

1.2.1 Programmed Cell Death in Animals and Yeast

PCD in animals has traditionally been divided into three types. The first, apoptosis, is a form of PCD with conserved morphological hallmarks and known signalling pathways for its activation. Its functions in animal immune systems and development are well-studied. In contrast to apoptosis, necrosis has been regarded as an unprogrammed and catastrophic form of cell death triggered by physical or chemical stress. This view has been proven to be inaccurate by recent research demonstrating receptor-inducible necrosis and cross-talk between necrotic and apoptotic signalling pathways. New research is also challenging the classification of autophagy as a form or method of PCD. Autophagy is the vacuolar uptake of cellular contents and their transport to a lytic organelle. Its continual function in cell survival and maintenance (Degenhardt et al., 2006) makes it difficult to say whether PCD-associated autophagy is a pro-death or pro-survival process.

Apoptosis

Apoptosis is the best studied form of PCD and is the primary form of PCD in animals. Apoptosis is a highly controlled form of PCD that results in small membrane-bound packets of degraded cell components that can be consumed by phagocytes. Apoptosis occurs as a series of events: changes in the cell mmebrane surface, a gradual detachment of the apoptotic cell from its neighbours, the formation of thread-like protrusions (blebs) of the cytosol and cell membrane, and the condensation and internucleosomal fragmentation of chromatin (Häcker, 2000). Blebs eventually separate from the cell to form apoptotic bodies—small membrane-bound packages of cell debris. These are scavenged by phagocytes thereby preventing the release of inflammatory factors into the extracellular environment. While these visible changes are occurring in the membrane, actin, myosin, tubulin, and dynein are proteolytically degraded within the cell (Taylor et al., 2008).

In addition to the conserved morphology of apoptosis, there are conserved core signalling pathways responsible for regulating the process. These core pathways are called the intrinsic and extrinsic pathways. The intrinsic pathway of apoptosis is centred around the mitochondria and is primarily induced by intracellular stresses. These can include DNA damage, exposure to ultraviolet and γ-radiation, or growth factor deprivation (Wang, 2001; Li and Yuan, 2008; Brenner and Mak, 2009). The extrinsic pathway is activated by extracellular ligands that are transduced across the cell membrane by receptors (Thorburn, 2004). The result of both pathways is activation of caspase-3, a protease responsible for breakdown of cellular components and activation of other effectors. There is a common set of major players involved in apoptosis: death receptors found at the plasma membrane (Wilson et al., 2009), the members of the Bcl2 family of proteins (Martinou and Youle, 2011), cytochrome

c (Suen et al., 2008), and caspases (Taylor et al., 2008).

The Bcl2 family of proteins act on the mitochondria to regulate cell fate. Some members actively promote cell survival by maintaining mitochondrial integrity (Yang, 1997), while other favour apoptosis by inducing a rapid increase in mitochondrial permeability (Chipuk et al., 2010). Thus Bcl2 proteins can be labelled as either anti-apoptotic or pro-apoptotic. Pro-apoptotic Bcl2 proteins can induce mitochondrial permeability by forming multimeric pores in the outer membrane that are large enough to allow discharge of the intermembrane contents (Martinou and Youle, 2011). They can further increase membrane permeability by opening ion channels (Shimizu et al., 1999). Anti-apoptotic Bcl2 proteins inhibit their pro-apoptotic siblings. Apoptosis is partially activated by interruption of this inhibition (Yang, 1997; Li and Yuan, 2008). When the integrity of the mitochondrion is compromised by the activity of pro-apoptotic Bcl2 proteins, cytochrome *c* is able to leave the intermembrane space and enter the cytoplasm. Cytochrome *c* forms complexes with a docking protein, apoptotic protease activation factor 1 (Apaf-1), and caspase-9 (Li et al., 1997). Formation of this complex converts caspase-9 into an active protease that is able to activate caspase-3, the primary effector of apoptosis (Li et al., 1997).

Death receptors transduce pro-apoptotic signals from the extracellular environment to activate the extrinsic pathway. These signals are specific ligands. Upon binding, death receptors cluster, then cleave caspases 8 and 10 to active forms, which in turn activate caspase-3 (Wilson et al., 2009). Ligands for the death receptors can include cytokines such as tumour necrosis factor (TNF)- α and surface markers of cytotoxic lymphocytes (Ju et al., 1995; Wilson et al., 2009).

The two pathways of apoptosis are not distinctly separated. The extrinsic pathway can strengthen its activation of apoptosis by inducing the intrinsic pathway. After caspase-8 has been activated by a death receptor complex, it can activate Bid, a Bcl2-related protein that is able to promote the release of cytochrome *c* from the mitochondria (Luo et al., 1998). Loss of Bid reduces the capacity of the extrinsic pathway to induce apoptosis (Yin et al., 1999). Whether apoptosis is triggered by the intrinsic pathway or the extrinsic pathway, the mitochondrion has a central role in apoptosis.

Caspases are cysteine proteases that recognize four-amino acid motifs. They always cleave after a C-terminal aspartic acid residue (Li and Yuan, 2008). Almost all caspases are involved in apoptosis and have roles either as initiators or executioners. Initiators activate downstream executioner caspases by proteolysis, but do not effect changes in the cellular structure. These enzymes include caspase-8, -9, and -10 (Pop and Salvesen, 2010). The executioners include caspase-3, -6, and -7 (Pop and Salvesen, 2010); they directly bring about apoptosis by proteolytic degradation of cell components. This is an oversimplified view of the caspace cascade, because initiators and executioners do not consistently fit their labels. Executioner caspases are able activate both other executioners as well as initiators (Inoue et al., 2009), creating a strong positive feedback loop that contributes to the irreversibility of apoptosis.

Caspase-3 is the primary executioner of apoptosis and is necessary for the nuclear degradation and gross changes in morphology observed during apoptosis (Lakhani et al., 2006). CAD endonuclease is responsible for intranucleosomal DNA cleavage (Porter and Ja, 1999), a hallmark of apoptosis. Caspase-3 cleaves an inhibitor of CAD, ICAD/DFF-45, thus initiating DNA degradation. Caspase-3 is also responsible for triggering cytoskeletal destruction. It cleaves gelosolin into an enzyme fragment that

is a potent actin depolymerization enzyme (Kothakota et al., 1997). Caspase-3 also cleaves off the regulatory domain of PAK2 (p21-activated kinase), a kinase involved in the cytoskeletal effects of apoptosis (Rudel and Bokoch, 1997). This cleavage causes PAK2 to become strongly activated. Caspase-3 is central to apoptosis because it can activate other executioner caspases and upstream initiator caspases (Inoue et al., 2009).

Autophagy

Autophagy is the internal degradation of portions of the cytosol in lytic vacuoles (Fuchs and Steller, 2011). Its primary role is the promotion of cell survival and health. During starvation, autophagy digests intracellular components to provide nutrients. It also continually breaks down aging organelles and misfolded protein (Mizushima, 2005; Degenhardt et al., 2006). Autophagy is broken into two subtypes: microautophagy and macroautophagy. These two types differ in the mode by which materials are transported to degradative lysosomes. Microautophagy is the direct uptake of cytosol into the lysosome by membrane invagination. Macroautophagy is the engulfment of cytoplasm by double-membranes vesicles called autophagosomes (Levine and Klionsky, 2004). It is the predominant form of autophagy (Rabinowitz and White, 2010). The terms *autophagy* and *macroautophagy* are used often interchangeably.

Autophagy has been studied most extensively in several yeast species (Levine and Klionsky, 2004). The autophagic process is triggered by depriving the cells of key nutrients. It is triggered in as little as 30 minutes (Takeshige et al., 1992). At this point, the nutrient-deprived cells form autophagosomes that engulf portions of cytoplasm before fusing with a lysosome. The contents of autophagosomes (autophagic bodies) tend to be derived from a diverse array of organelles (Takeshige et al., 1992). Yeast accumulates large numbers of ribosomes and metabolic enzymes during nutrient-rich

periods. Autophagy makes use of these cellular components as a nitrogen source during starvation (Takeshige et al., 1992; Rabinowitz and White, 2010).

Autophagy is rarely used to mitigate cellular starvation in animals. It is more important for recycling cellular components. Organelles that have lost structural integrity can be eliminated by autophagy. Autophagy is also the primary method for regulating populations of organelles. Large amounts of misfolded cytosolic protein are removed by autophagy. The continual removal of cell components in this manner is referred to as *basal autophagy* (Klionsky, 2000).

In mammals, autophagy is the primary mechanism for maintenance of normal peroxisome populations. Excessive peroxisome proliferation can be artificially induced by di-(2-ethylhexyl) phthalate. Autophagy quickly returns peroxisome counts to normal level (Oku and Sakai, 2010). Decrepit mitochondria are also eliminated by autophagy. When the mitochondria begins to lose membrane polarity, the marker protein PTEN-induced putative kinase 1 (PINK) accumulates on the its surface (Narendra et al., 2010). PINK recruits a ubiquitin-ligase (Parkin) that induces specific autophagic removal of the organelle (Narendra et al., 2008, 2010). Protein misfolding and glycosylation errors in the endoplasmic reticulum trigger autophagy (Yorimitsu et al., 2006).

Autophagy's ability to consume large quantities of cytosolic protein represents a bulk alternative to proteasomal degradation. Ubiquitinated protein can accumulate in ordered cytoskeleton-associated structures called aggresomes (Johnston et al., 1998) or in agglomerations referred to as protein inclusions (Kirkin et al., 2009; Pankiv et al., 2007). Aggresomes are believed to be a form of long term storage for damaged protein (Kraft et al., 2010) while inclusions are aggregations formed by hydrophobic

interactions between misfolded proteins (Kirkin et al., 2009). Both structures are polyubiquitin-rich and are removed by autophagy. Though once believed to be separate, the proteasomal and autophagic pathways of protein degradation seem to be linked. Polyubiquitination is evidently involved in the formation of concentrated collections of unwanted protein suitable for engulfment (Johnston et al., 1998) and in the actual activation of autophagy (Pankiv et al., 2007). Furthermore, some regulators of autophagy share sequence similarity with proteins involved in ubiquitination (Kirkin et al., 2009; Kraft et al., 2010). Parkin is required for autophagy of mitochondria, but is an E3-ligase that also ubiquitinates misfolded protein (Kraft et al., 2010).

Autophagy as a mode of programmed cell death in animals is under increasing scrutiny. Autophagy is triggered by potentially lethal events such as high levels of protein misfolding, viral and bacterial invasion, loss of mitochondrial integrity, and starvation. It is unclear whether this autophagic response is intended to mitigate lethal factors or to kill the cell. In most cases, a causative role for autophagy in programmed cell death is questionable. In cases of PCD believed to be autophagic in nature, inhibtion of autophagy pathways tends not to prevent the death of the cell. Conversely, inhibition of autophagy-associated cell death does not necessarily prevent autophagy (Levine and Yuan, 2005). During *Drosophila* metamorphosis, large numbers of autophagosomes populate the cells prior to PCD (Krömer and Levine, 2008). PCD can be prevented by mutating regulators of apoptosis and by applying caspase inhibitors. During this inhibition autophagy still occurs (Lee and Baehrecke, 2001), regardless of cell death programming. The fact the autophagy-associated PCD is prevent my inhibiting caspases suggests that cell death may be carried out by cellular components more closely tied to apoptosis. Many autophagy-associated genes (ATG) have wide ranging roles that include interaction with apoptotic regulators and regulation of cell structure and membranes (Krömer and Levine, 2008). Mutation or

inhibition of ATG genes could have unforseen effects on cell dynamics, making it difficult to draw conclusions about autophagic cell death by inhibiting autophagy.

Autophagy's roles in starvation, immunity, and recycling of cell contents are well-established. Its involvement in animal PCD is not. As it is now, autophagic cell death is more aptly described as PCD accompanied by autophagic activity. Autophagy accompanying PCD has possible functions aside from being an effector of cell death that could temporally associate it with PCD. It could be a last-ditch survival method or a preprocessing step before apoptosis and phagocytosis.

Necrosis

Necrosis is very different to apoptosis. The cell swells and ruptures, releasing its contents into the extracellular space. This causes inflammation and recruits immune cells. Apoptosis instead produces apoptotic bodies that do not disturb the surrounding tissue. Necrosis also results in DNA fragmentation, but it is random, unlike the internucleosomal degradation in apoptosis. Bypassing the costly steps of apoptosis makes necrosis energetically cheap. It is commonly associated with physical and chemical damage to cell. These cells may not have the time or energy required for apoptosis. Necrosis has been viewed as an uncontrollable and undesirable form of PCD. This idea has been challenged by new research showing that necrosis can be a regulated process.

Necroptosis is a term used to distinguish regulated necrosis from unregulated necrosis. It produces the same morphological result as necrosis. The primary inducible pathway for necroptosis begins at the cell surface with the TNF receptor (Vanden Berghe et al., 2010). Binding of TNF-α results in receptor activation and intracellular activation of two kinases: receptor-interacting protein (RIP) 1 and 2 (Christofferson

and Yuan, 2010). These proteins are believed to have interactions with caspases and Bcl2 family members (Galluzzi and Krömer, 2008), providing a regulatory interface between necrosis and apoptosis.

The signalling events downstream of RIP activation are unknown. However there are other intracellular events that coincide with necroptosis. The mitochondrial membrane becomes hyperpolarized in cells treated with TNF-α (Vanden Berghe et al., 2010), contrary to the depolarization integral to apoptosis. Hyperpolarization can be rapidly induced by exposing cells to hydrogen peroxide (Vanden Berghe et al., 2010). TNF receptor activation can also induce mitochondrial hyperpolarization as well as endogenous generation of reactive oxygen species (ROS). Disruption of the normal mitochondrial membrane potential increases oxygen consumption leading to a build-up of ROS (Goossens et al., 1999).

An intracellular increase in ROS is a major executive step in necrosis and necroptosis. Lipid oxidation caused by ROS disrupts the integrity of cellular membranes. Increased membrane permeability in lysosomes allows hydrolytic enzymes to escape into the cytoplasm (Zdolsek and Svensson, 1993). This can induce apoptosis and necrosis. While a slight increase in permeability favours apoptosis, a large synchronous increase can induce necrosis (Krömer and Jäättelä, 2005). The combined release of lysosomal hydrolases and weakening of the plasma membrane by lipid oxidation causes the cell damage and swelling observed during necrosis. ROS appear to be an end effector shared by necrosis and necroptosis.

1.2.2 Programmed Cell Death in Plants

While PCD research in animals is beginning to strain the apoptosis-autophagy-necrosis classification system, PCD is even more diverse in plants. Some form of DNA

fragmentation occurs in most instances of plant PCD. It can be internucleosomal, producing the DNA-laddering characteristic of animal apoptosis. But, it can also be an apparently random process with no specific degradation products. Changes in nuclear morphology may also occur, such as shrinkage, ordered subdivision, or disappearance. Changes in cytoplasm appearance, such as aggregation, gelling, or shrinkage, are common. They can occur anytime between the initiation of PCD and the death of the cell. Seemingly regulative increases in mitochondrial permeability in plant PCD have excited researchers looking for parallels to apoptosis, but in some cases the mitochondria outlast many other organelles and even maintain their function after the demise of the nucleus. Although plant PCD is complicated in its diversity, it can be unified by major involvement of the central vacuole.

The Central Vacuole in Plant PCD

A fundamental role for the central vacuole in PCD is almost ubiquitous in plants. Hara-Nishimura and Hatsugai (2011) have suggested that plant PCD be broadly categorized based on the role of the central vacuole. They propose two categories: non-destructive vacuole-mediated cell death and destructive vacuole-mediated cell death.

Non-destructive vacuole-mediated cell death is the fusion of the central vacuolar membrane (tonoplast) with the plasma membrane, that results in release of the vacuolar contents into the extracellular space (Hatsugai et al., 2009). The vacuole is loaded with antimicrobial proteins and secondary metabolites that kill bacteria and induce PCD in releasing plant cell within 12 hours (Hatsugai et al., 2009; Hara-Nishimura and Hatsugai, 2011). It appears to be a hypersensitive response to bacterial pathogens that infect the extracellular space of the host. Destructive vacuole-mediated cell death is better known as autolysis. It is the most commonly described form of plant PCD in

the literature. Autolysis involves the enlargement and rupture of the central vacuole, which results in discharge of hydrolytic enzymes into the cytoplasm. These enzymes effect the degradation of intracellular components and lysis of the cell.

Central vacuole enlargement and rupture is a common event in many forms of plant PCD, but its timing and function varies. PCD is required for complete maturation of xylem tissue. Fibres and tracheary elements (TE) are highly lignified cells that are dead at maturity. While xylem fibres in *Populus tremloides Michx.* × *P. tremula L.* exhibit cytoplasmic degradation and nuclear fragmentation prior to lysis by the central vacuole (Courtois-Moreau et al., 2009), the TEs of *Zinnia elegans* Jacq. first undergo tonoplast rupture, followed by postmortem chromatin degradation by hydrolases released from the vacuole (Groover and Jones, 1999; Obara et al., 2001). Though both processes would be classified as autolysis, the similarities begin and end at the destruction of the central vacuole.

Postmortem retention of an intact cell wall is necessary in fibres and TEs, but is not a universal feature of vacuolar cell death. Aerenchyma is a porous structure that facilitates movement of air through the shoots and roots of some plants. During its formation the middle lamella degrades, causing cells to separate from their neighbours. These cells undergo autolysis and their cell walls often collapse (Schussler and Longstreth, 2000). Cell wall degradation is completed by cellulases released by the dying cells (Jackson and Armstrong, 1999). The lace plant (*Aponogeton madagascariensis*) Mirbel.), is an aquatic plant whose leaves lose most of their interveinal tissue by maturity. This tissue is removed by PCD, followed by efficient cell wall degradation. Cell walls completely disappear within 24 hours of cell collapse in lace plant leaves (Wertman et al., 2012).

The nucleus can also undergo morphological changes during autolysis. In some cases, the nucleus remains intact and is removed by vacuolar engulfment. In others, the the nucleus undergoes gross morphological changes before the cytoplasm becomes vacuolated. Rapid PCD in barley (*Hordeum vulgare* L.) is immediately preceded by the formation of a large proteid vacuole. The nucleus does not fragment or become lobed (Bethke et al., 1999). In the endosperm of wheat (*Triticum aestivum* L.), there are marked changes in nuclear structure during PCD. Early in the process, the nucleus becomes condensed and the nuclear membrane is invaginated and lobed (Li et al., 2004).

Autolysis has many similarities to autophagy. Autolysis is commonly associated with an accumulation of small vacuoles in the cytoplasm (Filonova et al., 2000; Xiong et al., 2006; Courtois-Moreau et al., 2009; Wertman et al., 2012; Xiong et al., 2006). Cytoplasmic aggregation is a common feature of plant PCD (Yamada et al., 2000; Serrano et al., 2010) that suggests formation of protein inclusions or aggresomes similar to those formed during animal autophagy. Aggregation and vacuolation in the cytosol is typically accompanied by an increase in the volume of the central vacuole, indicating that vacuoles fuse with the central vacuole, delivering their cargo for hydrolytic degradation. Eventually this build up in vacuolar volume ends in rupture of the tonoplast, discharge of hydrolases into the cytoplasm (Bassham, 2007), and destruction of the plasma membrane.

There are few forms of PCD that do not fall under the vacuole-mediated label. Oat *Avena sativa* L. undergoes an irregular PCD process in the presence of victorin, a toxin produced by the pathogenic fungus *Cochliobolus victorae* Nelson. Contrary to most plant PCD processes, a reduction in cell size occurs in the absence of vacuole rupture or plasmolysis (Curtis and Wolpert, 2004). The process is also preceded by

loss of mitochondrial transmembrane potential.

Similarites and Differences to Animal PCD

Plant biologists frequently use the term *apoptosis* to refer to plant PCD. Hallmarks of apoptosis such as nuclear segmentation and internucleosomal DNA fragmentation do occur in plants, but only inconsistently. Other key hallmarks of apoptosis, including membrane blebbing and formation of apoptotic bodies have never been reported in plant PCD (van Doorn and Woltering, 2005). A major barrier to this is the cell wall of plants. Autolysis shares more similarites with autophagy and necrosis, than with apoptosis.

Autophagy in animals and autolysis both involve vacuolation of the cytoplasm and fusion of these vacuoles to a lytic organelle. In plants it is the central vacuole and in animals it is the lysosome. *Arabidopsis* uses autophagy to recycle its cellular components. To do this, it implements homologues of yeast and animal genes (Thompson et al., 2005; Liu and Bassham, 2010). The expansion and rupture of the vacuole during plant PCD parallels the increased permeability of the lysosome during necrosis and necroptosis. Membrane disintegration in both organelles results in discharge of hydrolytic enzymes into the cytosol that are believed to be instrumental in the destruction of the cell.

Autolytic cell death presents morphological properties of both autophagy and necrosis. There are many reports of autolysis, necrotic-like cell death, and autophagic cell death in plant PCD. These descriptions artificially separate forms of PCD that are very similar. The most common mode of PCD in plants is an autolytic process involving mass autophagy, central vacuole rupture, and cell lysis.

Roles in Development

PCD has many roles in plant development and reproduction. It is integral to the formation of some structures and is required for the elimination of others. Genes potentially involved in plant PCD are described in table 1.1.

In xylem, PCD is necessary for the formation of fibres and TEs, though the sequence of events is not shared (Courtois-Moreau et al., 2009; Groover and Jones, 1999). Fibre PCD is a highly coordinated event involving the synchronous death of many neighbouring fibres (Bollhöner et al., 2012). DNA degradation occurs early in the programmed cell death cycle (Courtois-Moreau et al., 2009), but lignification continues even after cell death. In TEs, a regulated build-up the cysteine proteases XCP1 and XCP2 in the central vacuole prepares the cell for death by autolysis (Avci et al., 2008). Upon vacuole rupture these proteases are distributed into the cytoplasm, degrading the cellular components. DNA degradation occurs after vacuolar rupture (Obara et al., 2001). Rupture of the tonoplast appears to be the key event in the final steps of TE formation in zinnia, a model for xylem development (Fukuda et al., 1998).

Senescence is the controlled break-down of unnecessary organs. It requires coordinated nutrient reclamation and PCD (Roberts et al., 2012). Leaf senescence is well-studied and is the subject of several large gene studies (Quirino et al., 2000; Roberts et al., 2012). PCD in a senescing leaf starts with degradation of the photosynthetic components of the leaf including the chloroplast and ribulose-1,5-bisphosphate carboxylase oxygenase (RuBisCo) (Roberts et al., 2012). This photosynthetic disassembly is followed by typical markers of autolytic plant cell death including nuclear degeneration, cytoplasmic vacuolation, and cell lysis (Lim et al., 2007).

During angiosperm seed development and germination, PCD is an important player. The wheat nucellus is removed by PCD during early endosperm development. Organelle depletion and a high level of cytoplasmic vacuolation followed by lysis suggest autolysis (Domínguez et al., 2001). Later in wheat seed development PCD occurs in the starchy endosperm. The storage cells have indicators of PCD including chromatin aggregation, nuclear fragmentation, and mitochondrial degradation. Though much of the cell degrades, the endoplasmic reticulum continues to function and no disruption of the cell membrane is evident. Both starch synthesis and accumulation are able proceed long past complete destruction of the nucleus (Li et al., 2004). This process occurs at random throughout the endosperm. When it is complete, the endosperm is dead and the aleurone layer is the only live tissue in the seed (Young and Gallie, 2000).

In *Euphorbia lagascae*, PCD is utilized to reclaim nutrients from the cells remaining in the seed after the storage reserves are depleted during germination. In concert with PCD, an upregulation of lipid transfer proteins (LTP) is suggested as a method for scavenging of membrane lipids (Eklund and Edqvist, 2003).

Roles in Conifer Reproduction

PCD is integral to several events in conifer reproduction. The morphology of PCD in conifers is similar to that in angiosperms, occurring by vacuolation of the cytoplasm and subsequent cellular rupture. DNA fragmentation and caspase-like proteolytic activity has been detected in some cases.

Fertilization is dependent on the pollen tube successfully penetrating the nucellus. This is not accomplished purely by force, as localized PCD occurs in the nucellus ahead of the elongating pollen tube (Hiratsuka et al., 2002). In Douglas-fir, the

nucellar cells proximal to the pollen tube become thoroughly vacuolated. These vacuoles coalesce into larger vacuoles whose contents are secreted from the cell (Owens and Morris, 1990). This process hasn't been positively identified as PCD, but the cells become collapsed and lose cell-cell adhesion after vacuolation (Owens and Morris, 1990). The nucellus is softened by PCD in *Pinus densiflora* Siebold Hiratsuka et al. (2002). It occurs by autolysis with organelle degradation and DNA fragmentation (Hiratsuka et al., 2002).

The pollen tube passes through the nucellus and delivers sperm to the archegonial space. Successful fertilization leads to embryogenesis. Because four archegonia commonly develop in Douglas-fir (Fernando et al., 1998), it is possible for more than one embryo to develop in a single ovule–a condition called simple polyembrony (Korbecka et al., 2002). In *Pinus sylvestris* L., multiple embryos can develop from a single zygote (Filonova et al., 2002). Both species produce only one mature embryo, as others are eliminated by PCD. Autolysis begins in the suspensors of the subordinate embryos and proceeds to their apices. DNA degradation can be visualized using terminal deoxynucleotidyl transferase dUTP nick end labeling (TUNEL). TUNEL staining accompanies autolysis, first appearing in the suspensor and eventually spreading through the entire embryo (Filonova et al., 2002).

The dominant embryo obtains nutrition from the megagametophyte through a nutritive fluid in the corrosion cavity that is rich in amino acids and sugars (Carman and Reese, 2005). In *Pinus sylvestris*, the corrosion cavity forms by PCD (Vuosku et al., 2009). As the embryo forms, the lining of the corrosion cavity is continually shed to provide nutrition Vuosku et al. (2009). TUNEL staining is localized to the lining of the corrosion cavity, while the rest of the megagametophyte remains intact.

The remaining tissue in the megagametophyte is mobilized during seedling growth. By this time, the megagametophyte is rich in lipids and seed storage proteins (Owens et al., 1993). Post-germination PCD in white spruce (*Picea glauca* Moench) involves internucleosomal DNA fragmentation (He and Kermode, 2003a) and caspase-like proteolytic activity (He and Kermode, 2003b). The cells die by autolysis (He and Kermode, 2003a).

PCD is essential to the production of a viable conifer seed. It is also essential to the abortion of failed seeds and recovery of nutrients from the megagametophyte. PCD in other conifer reproductive processes has been studied by histology, assessment of DNA integrity, and protease studies. Little is known about the genes that mediate PCD in conifers. Genetic studies of conifer seed development are also scarce. RNA-Seq is a novel technology that could provide clues regarding the genetic basis of conifer seed development and megagametophyte abortion.

1.3 Objectives and Hypothesis

I conducted an experiment to study the transcriptional differences between fertilized and unfertilized megagametophytes and between reproductive megagametophyte tissue and vegetative tissues from cone scales and bracts. By controlling pollination in Douglas-fir cones, I was able to collect RNA from fertilized and unfertilized megagametophytes at four dates over the course of one month. I used this RNA to produce Illumina sequencing data. Using this data, required the construction of a *de novo* transcriptome assembly, read mapping, and statistical analysis of the read alignments. My objective is to document the process of RNASeq analysis, to study the involvment of PCD in Douglas-fir ovular abortion, and to contribute to current knowledge of the genetics of Douglas-fir and conifer seed development.

Table 1.1: Possible genes of interest in Douglas-fir PCD during abortion

Gene	Function
ATG family	Essential for autophagosome formation in eukaryotic organisms
Nix	Involved with MPT and engulfment of defective mitochondria by autophagy
Bcl2 family	Key regulators of apoptosis, similar sequence transcript found in Arabidopsis
Apaf1	Required in the intrinsic apoptosis pathway for activation of executioner caspases
VEIDase	Plant proteases linked with PCD
VPE	An enzyme implicated in PCD effected by tonoplast-plasmalemma fusion
Metacaspase family	A family of caspase-like enzmyes in plants and animals
Caspase-like proteases	Proteases with substrate profiles similar to those of mammalian caspases
Cell wall degrading enzymes	Pectinases and cellulases are involved in pollen tube penetration

I hypothesize that transcripts for effectors and regulators of PCD will be highly, differentially expressed in megagametophytes undergoing abortion. I would also expect transcripts associated with normal seed development to be highly expressed in the fertilized megagametophytes when compared to the unfertilized megagametophytes. Transcripts similar to those currently described in angiosperm seed development are also likely to be expressed during known seed development events in the fertilized megagametophytes.

Chapter 2

Analytical Steps in RNA-Seq

2.1 Introduction

2.1.1 Next-Generation Sequencing

Next-generation sequencing (NGS) describes a group of recently-developed sequencing technologies that produce much more data than Sanger sequencing at a lower cost per base. The disadvantage of NGS is that while millions of reads are produced, they are considerably shorter than Sanger reads. This necessitates new methods for producing full length sequences from the read data.

454 Pyrosequencing

The first NGS technology was an array-based form of pyrosequencing developed by 454 Biosciences (Pettersson et al., 2009). The original technique produced approximately 500,000 reads per run with an average length of 108 base pairs (Margulies et al., 2005). Currently, the flagship 454 instrument is capable of producing 1 million reads per run with a maximum length of 1000 bases (Roche, 2011).

Pyrosequencing itself was a recently developed method that introduced a new concept: sequencing-by-synthesis. This process produces base calls by determining the identity of each nucleotide added during DNA polymerization. When DNA polymerase adds a nucleotide to a growing strand, pyrophosphate (PP_i) is produced (Ronaghi, 2001). During pyrosequencing, solutions of dATP, dCTP, dGTP, or dTTP are sequentially added to the sequencing reaction and PP_i production is measured. Successive rounds of nucleotide addition and detection of incorporation produce a sequence of base calls. PP_i is indirectly detected by including ATP-sulfurylase in the reaction. This enzyme converts PP_i to ATP, which then fuels generation of light by luciferase (Ronaghi, 2001). Apyrase is responsible for clearing dNTPs after each iteration. This has two functions as it removes each cycle of introduced dNTPs and removes the excess ATP produced by ATP-sulfurylase (Ronaghi, 2001).

Pyrosequencing has been massively scaled up in 454 sequencing technology. 454 is an array-based form of pyrosequencing that allows simulatenous pyrosequencing of many DNA templates. DNA is fragmented and single fragments are anchored to adapter-coated beads and replicated (Margulies et al., 2005). The product is a bead coated in replicates of the sequence of interest. These beads are then placed into wells on a flow cell containing millions of picolitre-sized wells that contain the enzymes required for pyrosequencing (Holt and Jones, 2008). The iterative cycles of dNTPs are carried over the flow cell in a buffer that also serves to remove free dNTPs and excess PP_i (Holt and Jones, 2008). A CCD imager derived from astronomy grade cameras records emissions from wells with excited luciferase (Rothberg and Leamon, 2008).

Illumina Sequencing

Illumina sequencing is the most commonly used sequencing platform for RNA-seq experiments. It is similar to 454 sequencing in its array-based approach and use of sequencing-by-synthesis. Instead of a surface with etched wells, Illumina relies on a lawn of oligonucleotide adapters anchored to a flat flow surface. DNA is fragmented and ligated to adapter sequences complementary to the adapters in the lawn. These adapters are allowed to hybridize, producing a surface coated in both free adapters and adapters hybridized to adapter-linked sequence fragments (Shendure and Ji, 2008). These fragments must be amplified to produce a detectable signal as DNA polymerase synthesizes a complimentary strand, so a process called bridge-PCR is used to create clusters of identical sequences (Figure 2.1).

In Illumina sequencing, detection of pyrophosphate-release is replaced by differentially labelled dNTPs that are reversibly terminal, meaning they reversibly prevent the addition of additional nucleotides to the growing strand (Turcatti et al., 2008). Each Illumina cycle consists of introducing a mix of labelled dNTPs over the flow cell and allowing incorporation of these into each cluster by DNA polymerase. The flow cell is then imaged to identify the incorporated nucleotide in each cluster based on the attached the flurophore. The fluorophore is removed following imaging to relieve inhibition of further nucleotide addition. The original possible read length was 36 (Holt and Jones, 2008) bases, but has now increased to 150 bases.

An additional advent in both 454 and Illumina technologies was paired-end sequencing. Rather than sequencing from one end of each fragment, both ends of each fragment are sequenced. This produces paired reads that are separated by the distance between the $3'$ ends of the reads. Because this distance can be deduced from the read and fragment lengths, paired-ends provide a major advantage during data

analysis. To produce a paired-end data set, different adapters are ligated to each end of the fragment (Figure 2.2-A) and hybridized to the adapter lawn (Figure 2.2-B). These are then bridge-amplified to produce clusters of adapter-ligated fragments in both directions with exposed adapters (Figure 2.2-C). The sequencing for each direction is done separately, using adapter-specific primers to chose from which direction to sequence (Figure 2.2-DE). The result is that each cluster produces two reads, one from each of end of the read. Because the fragment size is known, the approximate distance between the mate pairs is known and can be used for verifying assemblies, scaffolding assembled contigs, and increasing mapping stringency.

2.1.2 RNA-Seq

RNA-seq is a special use of NGS that extends the process beyond genome sequencing. Instead of fragmented genomic DNA, cDNA is used as the input for NGS sequencing. The read data produced by this process can be used either for mapping to an existing reference genome or used to produce a *de novo* assembly against which the reads can then be mapped. Transcripts, exons, or CDSs that may be of interest to a researcher can collectively be referred to as genetic features. The read mappings to each genetic feature can be quantified to produce expression values.

When both the read data and a reference sequence are available, expression profiling is quite simple. The process begins with mapping the reads to a reference using alignment tools specifically designed for the task such as BWA (Li et al., 2009), Bowtie (Langmead et al., 2009), or Bowtie2 (Langmead and Salzberg, 2012). The product is a file containing the location of every read in relation to the reference. The reads-per-feature are calculated and adjusted based on the length of the genetic features.

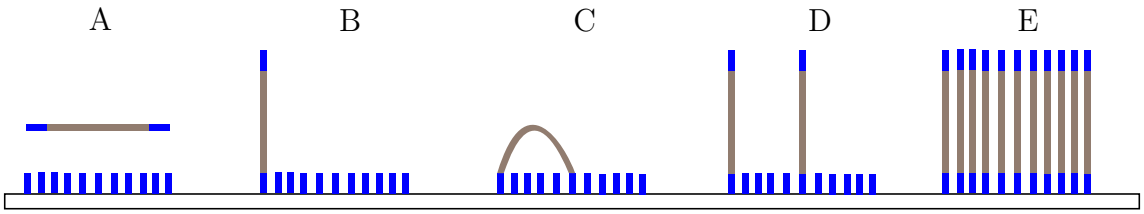


Figure 2.1: Illumina cluster generation. A) Adapters are ligated to both ends of the insert fragment. B) The adapters hybridize with complimentary adapters anchored to the flow cell surface. C) Further hybridization of free adapters with anchored adapters results in (D) bridge formation and successive PCR cycles create (E) clusters of identical sequences.

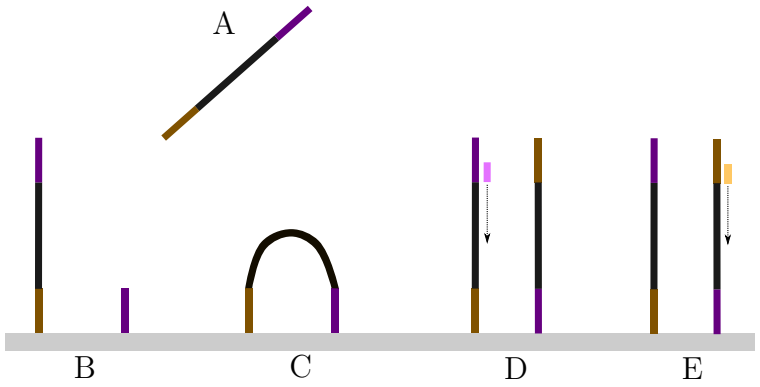


Figure 2.2: Schematic of the flow cell view of Illumina paired-end cluster generation: A) Different adapters (bronze, purple) are ligated to ends of fragments (black), B) The ligated adapters hybridize with adapter lawn, C. Fragments are bridge amplified to form clusters, D. Sequencing of one end of the paired reads is initiated with a specific primer E. Sequencing of the other mate is initiated with another specific primer.

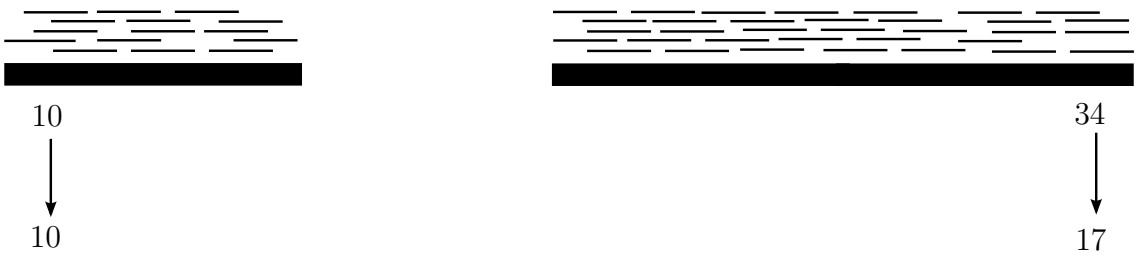


Figure 2.3: Read counting and expression values for two contigs. Reads are aligned to reference contigs or a genome reference. The reads aligning to each genetic feature are counted and expression values are calculated, taking into account the lengths of each feature (the right contigs is twice the length of the left).

RNA-seq presents a number of advantages over microarray technology. It can potentially detect and quantify all transcripts in the input data, whereas microarray results are limited to expression profiling only of the cDNA included in the array. Its dynamic range for resolving expression levels is far greater than microarrays (Wilhelm and Landry, 2009). Even very lowly expressed transcripts can be quantified with adequate sequencing coverage. RNA-seq also provides the opportunity to resolve isoforms (Wang et al., 2009). Because RNA-seq is based on sequencing each transcript it can potentially resolve individual transcripts with single-base resolution. Microarray technology is susceptible to non-specific hybridization, making its differentiation of similar transcripts or isoforms inferior to that of RNA-seq.

RNA-seq is not without disadvantages. It is more expensive than using an existing microarray to examine a cDNA sample. The computational power required for *de novo* assembly and the large amount of storage space required for the data adds cost to RNA-seq experiments. Like many cutting edge technologies, RNA-Seq can be challenging to use because there is a lack of standard analysis procedure and software. Many processing steps are required to find differentially expressed transcripts in an organism that doesn't have a reference genome (Figure 2.4). To make a reference transcriptome from Illumina sequencing reads, the reads must first be evaluated for low quality base calls and Illumina sequencing adapter contamination. *De novo* assembly of a transcriptome from Illumina reads is a very computationally intense process. A powerful server or computing cluster is essential. Once a reference sequence set is generated, the read data is mapped on to it sample-by-sample. The number of reads mapping to each reference transcript can be used to calculate normalized absolute expression values to allow comparison between read libraries (Oshlack et al., 2010). There are many algorithms available for normalizing RNA-Seq data and for finding significantly differentially expressed (DE) transcripts. However, there is no generally

accepted technique for normalization or DE analysis, so software must be selected by the researcher based on his or her needs.

Researchers must acquire a mastery of basic Linux commands as well as some programming to be able to do RNA-Seq analysis. Most current tools for RNA-Seq analysis are command line-driven. Optimizing and using these tools requires knowledge of both Perl and R statistical language. Shell scripting is the only way to handle the large numbers of files and lines of information made during RNA-Seq analysis. Although microarray analysis began as an unstandardized, complicated process, analysis tools have been developed that allow a set of standardized methods to be implemented using increasingly user-friendly software. As the use of RNA-seq increases, software will likely become more accessible, mirroring the developments in microarray analysis.

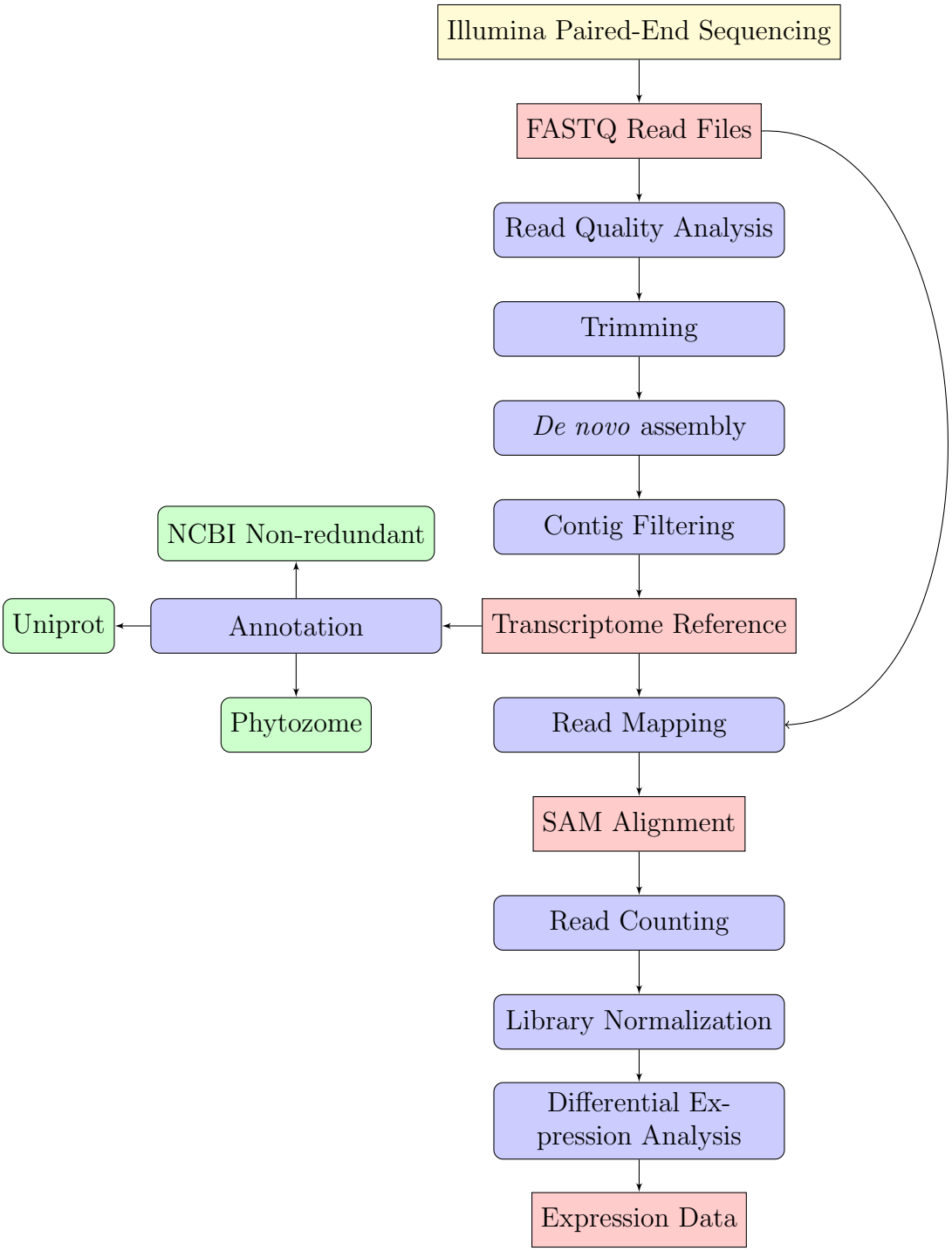
This chapter focuses on summarizing currently available algorithms and software for each step of an RNA-Seq workflow. The principals of data processing and analysis are discussed along with the advantages and disadvantages of available applications.

2.2 Computing Considerations

2.2.1 The Linux Environment

Most software available for analysis of NGS data is intended for use in a Linux environment. It is distributed either as source code that must be compiled prior to use or as precompiled program files. Many open source bioinformatics programs use Perl and Python interpreters, which are integrated into most Linux distributions. String manipulation utilities integrated into Linux, such as *grep*, *sed*, and *awk*, lend themselves very well to performing operations on sequencing data. *Grep* can extract

Figure 2.4: Steps in an RNA-seq workflow



lines from sequence files or count FASTA entries and *sed* can search and replace text in sequence files. *Awk* is a versatile text manipulation language that can quickly sort and extract information in tabular data. Linux shell scripting is useful for automating tasks for multiple read libraries.

Computing clusters are necessary to handle the high memory and processing demands of analyzing NGS data. These computers almost exclusively run UNIX-based operating systems. Standards for handling job management, parallel processing, and interconnect are well established for Linux and other UNIX-based operating systems.

2.2.2 Computing Strategies

Steps in RNA-Seq analysis, such as *de novo* assembly and read mapping are computationally demanding. Parallel processing is one way of meeting this demand. Almost all NGS software process data in parallel. The most common strategy is to exploit the multiple processing threads available in modern multicore processors. Tasks can be distributed between these threads to increase throughput. Programs that use this strategy include the *de novo* assemblers Trinity (Grabherr et al., 2011) and Velvet (Zerbino and Birney, 2008), and the read mappers Bowtie (Langmead et al., 2009) and BWA (Li and Durbin, 2009). The limits of this strategy are the amount of memory and number of processing threads that can be delivered in a single computer. *De novo* assembly benefits from hundreds of computing threads, and can require hundreds of gigabytes of memory. This scale of resources is only available in computer clusters.

Clusters are composed of hundreds to thousands of interconnected nodes. Each node is similar to a powerful desktop computer and has its own processors, memory, and storage. A cluster can be configured in two ways. In the first configuration, the

cluster is a collection of individual nodes that each possess dedicated memory. Programs that need more resources than are available on a single node must be specially programmed to distribute their processes over multiple nodes and pass information between these distributed processes. This task is handled by the message-passing interface (MPI) standard. MPI is freely available as an open source C++ library called OpenMPI (Gabriel et al., 2004). The second configuration strategy is to create a system in which all processors can address a Global Shared Memory (GSM) that is the distributed across all nodes (Dunigan et al., 2005). GSM access is simulated by an extremely fast interconnection between nodes. This obviates the need for MPI programming.

OpenMPI adds complexity to writing a *de novo* assembler. It is implemented in only two *de novo* assemblers: ABySS (Simpson et al., 2009) and Ray (Boisvert et al., 2010). All other assemblers are intended for multi-threaded computing on a single large node. The alternative is a GSM cluster. Blacklight, a computer built by SGI, has been used extensively for running demanding Trinity assemblies that require hundreds of gigabytes of memory (Henschel et al., 2012).

It is important to understand the architectures of computers and the programming methods used in the software. Pairing software with its intended hardware architecture makes running analysis faster and more efficient.

2.3 Data Files

Many bioinformatic file formats are problematic because they are unstandardized. NGS data can be large and complex, requiring well designed and standardized file formats. Today, the core file formats used in NGS data have published standard

formats. FASTQ is a read storage storage format derived from FASTA. SAM and BAM are related read-map storage formats.

2.3.1 FASTA

FASTA is a nucleotide and amino acid sequence format introduced by Pearson and Lipman (1988). It has been retained as the main sequence storage format for over two decades without any formal attempt at standardization. Each entry in the FASTA file consists of a descriptive line starting with '>' followed by information about the sequence. The lines following this constitute the sequence string (Figure 2.5) and can be provided as a single line or multiple lines of characters. The separation of entries in FASTA files relies on the '>' beginning each header line. Programs that assume each sequence is stored as a single line can incorrectly parse files where the sequence data is stored as multiple lines.

Figure 2.5: A sample of FASTA file content. Each FASTA entry is defined by an identifier line starting with '>' followed by the sequence line.

2.3.2 FASTQ

The FASTQ format stores sequence data and base quality information in the same file (Figure 2.6). It was originally derived from the FASTA format by the Sanger Institute. It too, lacked a formatting standard (Cock et al., 2010). The quality scores (Q) are log-transformed error probabilities (P_e) for each base called Phred scores. To save space, the ASCII characters with codes 33–126 are used to represent 93 possible values for Q (Cock et al., 2010).

Cock et al. (2010) realized the problems of undefined file formats and have since published a standard for the FASTQ format. Each entry of the FASTQ format consists of four lines. The first is an identifier line beginning with '@'. In Illumina FASTQ files, the identifier line contains instrument information, experimental conditions, and signal coordinates (Illumina, 2011). The second line contains the entire sequence and the third line contains only the character '+'. The final line contains ASCII characters that encode per-base quality scores. The sequence and quality lines must be equal in length and any parsing must ignore '@' and '+' as line identifying characters because they can occur in the ASCII quality line.

In the original Sanger FASTQ format, Phred quality scores (Ewing and Green, 1998) were calculated using the equation:

$$Q_{phred} = -10 \times \log_{10}(P_e)$$

Because the quality scores are represented by the ASCII codes from 33 to 126, a range of scores from 0 to 93 are possible. This allows a range for P_e of 1.0 to $10^{-9.3}$. Based on the Sanger calculation, a base call with a 0.1% chance of being incorrect has Phred quality score of 30. The associated ASCII code is the sum of the first

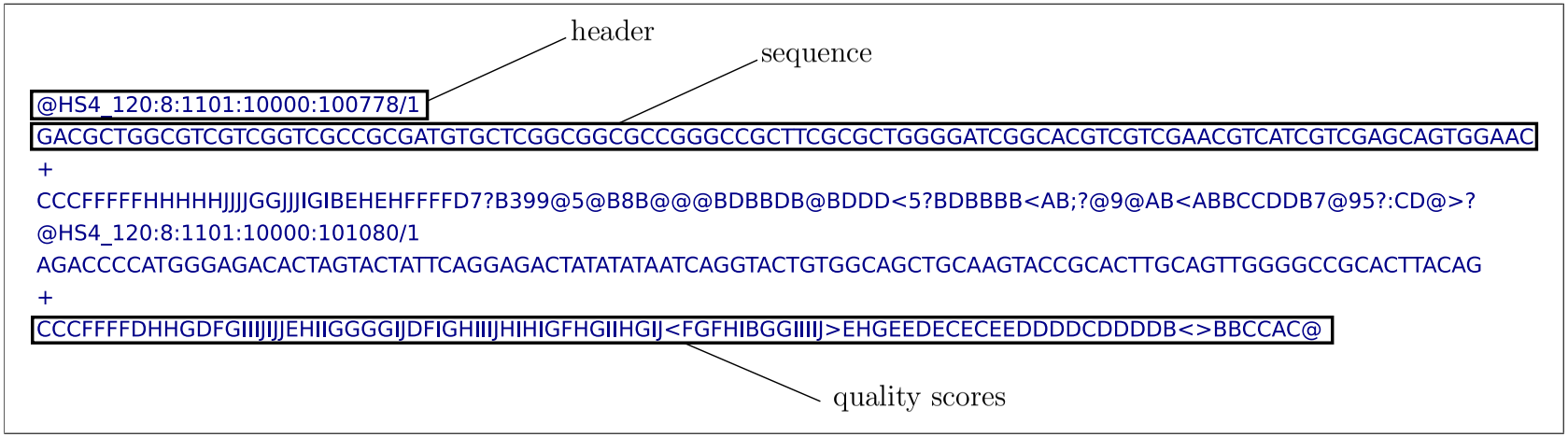


Figure 2.6: Two lines of a FASTQ file. The first line of each entry is a sequence identifier. The second line contains the sequence. The third line contains a single '+'. The final line is an ASCII string representing log probabilities of error.

human-readable ASCII code (33) and the quality score.

Illumina has changed the calculation and reporting of the error scores three times. The quality scores reported by the earliest Illumina instruments ranged from -5 to 62 and were represented by the ASCII characters with codes from 59 to 126 (Cock et al., 2010). The ASCII codes were determined by adding an offset of 64 to the quality score. The scores themselves were not calculated with the same equation as Phred scores. Instead they were calculated using the equation:

$$Q_{solexa} = -10 \times \log_{10}\left(\frac{P_e}{1 - P_e}\right)$$

Illumina changed their quality reporting again in version 1.3 of their analysis pipeline software, CASAVA. They returned to using the original Phred equation to produce quality scores. However, they only used a range of scores from 0 to 62 and the ASCII characters from 64 to 126 (Cock et al., 2010). Finally, CASAVA 1.8 changed the quality encoding of Illumina FASTQ files to use both the original Sanger equation and the ASCII offset of 33 for producing Q scores (Illumina, 2011). Because of these changes, any software that uses FASTQ quality scores in analysis must be provided with the quality encoding version. The last two Illumina Phred encodings are the most commonly accepted and are denoted as either Phred33+ or Phred64+ (Table 2.1).

2.3.3 SAM and BAM

Read data are typically mapped to a reference that can be a genome or a *de novo* assembly of a transcript set. Fast read aligners map millions of short sequences to a reference. The output is a map of all reads that successfully aligned to the transcriptome reference, the sequences of those reads, and the sequence of the reference. Li

Table 2.1: Quality scoring systems used in the FASTQ format.

Name	ASCII Range	Offset	Type	Score Range
Sanger	33 - 126	33	Phred33	0 - 93
Illumina 1.0	59 - 126	64	Solexa	-5 - 59
Illumina 1.3	64 - 126	33	Phred64	0 - 62
Illumina 1.7	33 - 126	33	Phred33	0 - 93

et al. (2009) have published two well-defined file formats for storing this information: SAM and BAM.

The SAM (sequence alignment/map)format acts as a simple container for storing the output of read mapping programs (Li et al., 2009). Every alignment is stored as a one-line entry including the read name and mapping information (see table 2.2 for detailed fields). SAM files also store the reference sequence and the sequences of all successfully mapped reads. The binary alignment/map (BAM) is a compressed, binary counterpoint to the SAM file. It is much smaller than the equivalent SAM file, but still permits fast retrieval of information (Li et al., 2009).

SAMtools is a software package for manipulating SAM and BAM files. SAM files can be converted to BAM format to reduce disk usage and it can also index BAM files, which improves query speed (Li et al., 2009). SAM and BAM files can be sorted such that the read mapping list is ordered by position in the reference. This reduces the complexity of the SAM file, and is a requirement for some programs that accept BAM and SAM files as input (Li and Dewey, 2011).

The BAM and SAM formats have become widely used since their introduction. They are the standard formats for storing read mapping results. Developers can easily generate SAM files because of the format's simplicity. The BAM format combined with SAMtools provides an easy way to compress human-readable data into a small package.

2.3.4 File Interconversion

Read data is often delivered in BAM format to minimize disk usage and file transfer time. The left and right mates of a paired-end data set must be extracted from the

Table 2.2: The data fields of a SAM line. The data provide comprehensive information about the mapping location of the read, quality of the mapping, fidelity of mate pairs. A bitwise flag contains codes indicating pair and mapping status for quick parsing.

Field	Example
Read identifier	`HS4_118:7:2306:9999:90099`
Bitwise flag	147
Identifier for left-read reference hit	contig4365577
Left-most hit position in reference	352
Phred score for read hit	42
CIGAR string indicating match discrepancies	91M
Right pair reference hit (= if the same)	=
Distance between mates	231
Inferred insert size	212
Query sequence	`ACAACTCTAACGGAC`
Query ASCII-33 string	`###############`

BAM file for assembly and read mapping. There are two tools for accomplishing this task: Picard Tools or bamToFastq.

Picard Tools is a software suite related to SAMtools (PicardTools, 2009). Like SAMtools, it is used to manipulate SAM and BAM files. One of its functions is to extract paired end reads from a BAM file and output them in FASTQ format. It requires a recent version of the Java Virtual Machine (JVM) to run. bamToFastq is a component of Hydra, a software package for locating genomic structural breakpoints using NGS data (Quinlan et al., 2010). It can quickly extract paired-end reads from a BAM file. BamToFastq is implemented in C++, which allows the user to avoid installing a JVM.

Two files are produced by extracting FASTQ files from a BAM file containing paired-end reads. One file contains the left-handed mates and the other contains the right-handed reads. To identify the mate files, left-handed file names are suffixed with '_1' and right-handed files with '_2'. Additionally, the identifier lines in the FASTQ files are terminated by `\1` if they are left-handed reads or `\2` if they are right-handed.

2.4 Processing Read Data

2.4.1 Read Data Assessment

Raw read data is not immediately ready for assembly. Since low quality base calls and sequencing artefacts can have a negative impact on the quality of a *de novo* assembly, it is necessary to evaluate the data set. Each read data set has idiosyncratic problems due to differences between instruments, runs, and the wet lab processing prior to sequencing.

FastQC

FastQC is a tool for assessing FASTQ read data (Kircher et al., 2011). It gathers a variety of metrics that can indicate problems in the data (Andrews, 2012). The most important of these include regions of low base quality, biased nucleotide composition, sequence duplication, and overrepresented sequences. This information can be used to optimize trimming.

Low base quality can hamper the construction of full-length contigs during assembly by causing misassemby and introducing ambiguity. FASTQC interprets the quality scores from a FASTQ file and represents them as box-and-whisker plots (Figure 2.7). The plots represent the distribution of quality values at each position over the length of the read. This can identify areas of where many reads have quality scores below the threshold quality score (20). The mean quality scores usually decrease as the read increases in length toward the 3′ end due to the limitations of the sequencing process (Liu et al., 2012a). These figures are valuable for identifying areas where trimming is necessary and assessing the overall quality of the dataset.

FASTQC also calculates the nucleotide composition of each position over the length of the reads (Figure 2.8). If the nucleotide composition is not equal over the length of the reads, a significant portion of them my contain overrepresented sequences. Usually the first several base positions in the read are biased towards certain nucleotides. This occurs due to the random hexamer priming used in Illumina library construction (Hansen et al., 2010). Nucleotide bias elsewhere in the read can indicate the presence of a biological sequence that occurs very frequently in the template or an error in sample preparation. Read sets with strong nucleotide biases over the length of the read may be heavily contaminated or may have been incorrectly amplified.

FastQC provides a list of sequences and k-mers that are overrepresented in the read data. FastQC attempts to identify them as commonly recurring sequences such as Illumina adapters or ribosomal RNAs. Adapter sequences can be used to guide trimming.

FASTQC is the most popular software package for assessing read data. It provides a good range of metrics for the data and in user-friendly output. FastQC produces plain-text summary files, PNG figures, and a HTML document that is easy to navigate. It is an essential tool in the transcriptomic workflow.

2.4.2 Read Filtering

The objective of read-trimming is to make a high-quality set of reads for *de novo* assembly. Several problematic elements can occur in read data and must be removed. Low quality bases, which tend to occur at the $3'$ end of the read, must be removed to avoid introducing sequencing errors into data analysis. Illumina adapter contamination at the $3'$ end of the read is very common, while contamination at the $5'$ end is rare. Adapter sequences must be trimmed from the data. The consequence of read filtering is losing data. Reads that become too short during trimming must be discarded. Trimming software is able to remove low quality bases using FASTQC quality scores, trim adapter sequences from reads, and maintain mates in paired-end data.

Base quality information is included with sequence data in FASTQ files (Cock et al., 2010). The benefit of this is that low quality bases can be removed from the read data before analyzing it. Because low base quality occurs primarily in the $3'$ end of the read, the simplest method is to remove bases at this end that fall below a quality cut-off. Although highly effective (Martin, 2011), it is unable to discriminate

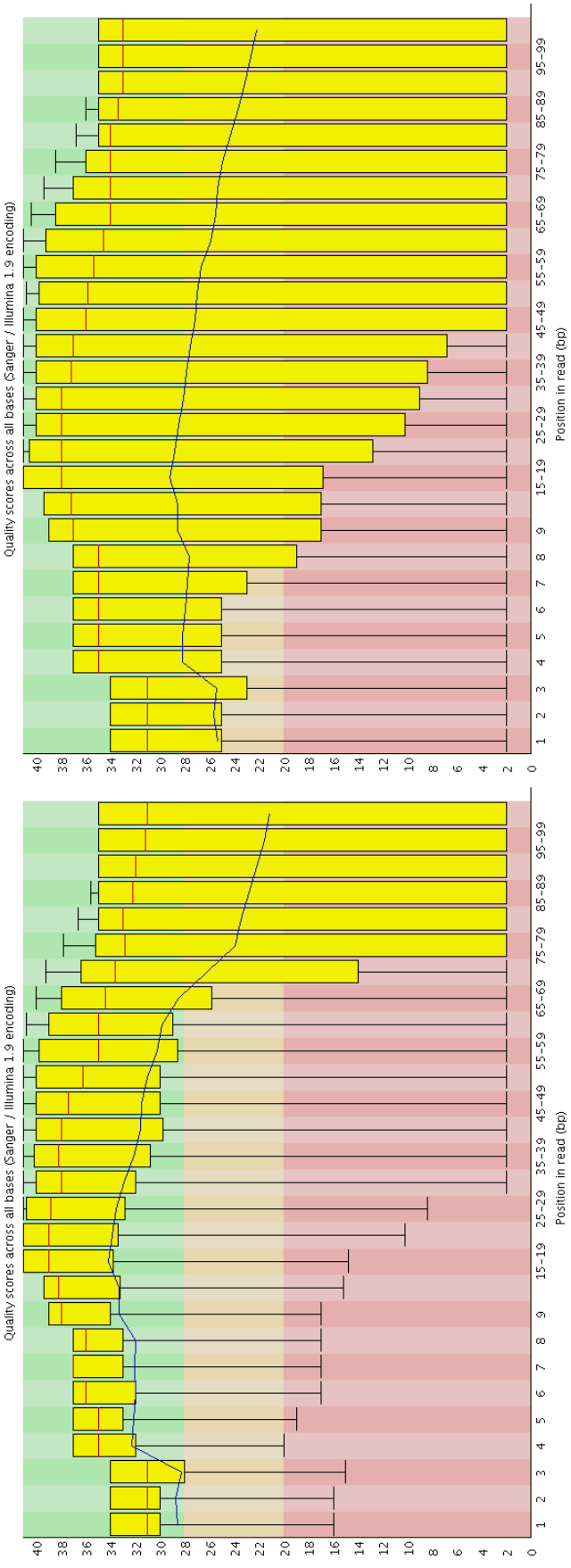


Figure 2.7: Sample per-base quality output from FastQC for left (L) and right (R) paired-end FASTQ files. Whiskers are 10 and 90% points, yellow boxes are inner quartiles (25–75%), red lines are median qualities, and the blue line is mean quality.

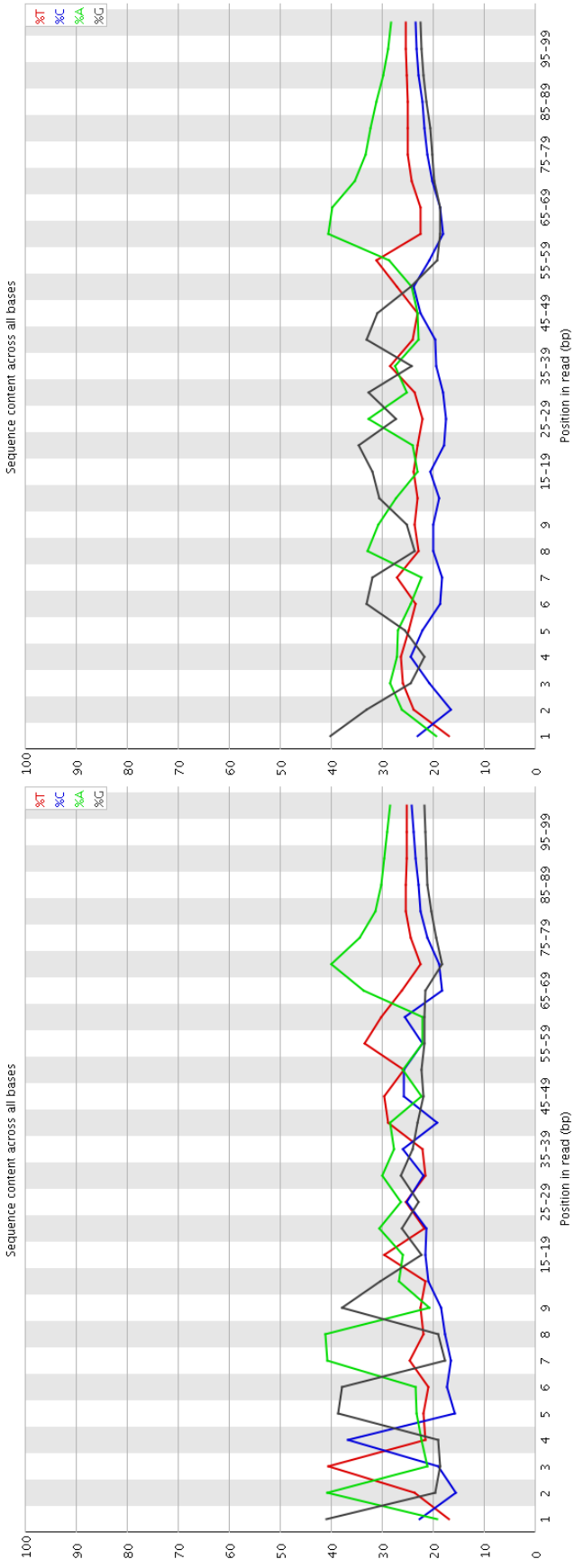


Figure 2.8: Sample per-base nucleotide content from FastQC for left (L) and right (R) paired-end FASTQ files. Red = thymine, blue = cytosine, green = adenine, black = guanine.

low quality in other regions of the read. Those regions can be trimmed using a sliding window method that is included in more advanced trimming software (Lohse et al., 2012). A window of user-defined length travels along the read and searches for areas where the average base quality is below a set threshold. When the average quality is too low, the read is cut and all preceding bases are discarded. Removing low quality bases helps to remove sequencing errors from the data.

Illumina library construction begins with a set of sheared fragments from the DNA of interest. Fragments are size-selected by gel electrophoresis so the length of the fragmented are within a known range. Adapters are ligated to both ends of each fragment. The adapter-fragment construct can be sequenced at both ends. Because the approximate fragment length and read sizes are known, the distance between each mate-pair can be calculated and exploited in assembly and mapping. The size-selection is not perfect and short fragments can introduce problems into the read data.

Ideally, a sequencing read begins at one end of the insert and ends somewhere within the insert (Figure 2.9A). Adapter sequences are found in 3′ end of the read when the insert between the flanking adapters is shorter than the read length (Martin, 2011). The read is sequenced from the 5′ end and continues into the 3′ adapter if the insert is too short (Figure 2.9B). Adapters are removed by aligning an adapter query sequence to the 3′ end and discarding bases between the adapter hit and the 3′ end. Even one-base alignments at a 3′ end should be trimmed because single base errors can introduce ambiguity into *de novo* assembly.

Paired-End Trimming

Inserts in paired-end sequencing are usually long enough that reads initiated from each end are separated by an unsequenced area in the middle of the insert (Figure 2.9C). Adapter contamination in paired-end data occurs only in very short inserts. In order for an extending read to incorporate part of the $3'$ adapter, it must also overlap its mate pair (Figure 2.9E). In this case the $3'$ contamination at the end of both mates should be identical and can be removed confidently (Lindgreen, 2012). The mates can also be collapsed into a single read (Lindgreen, 2012). Mate pairs can also overlap without incorporating adapter sequence. These mates can be merged into a single unpaired read. Paired-end data is easier to trim because $3'$ contamination is usually be duplicated in mate pairs, however using two paired input files also introduces a complication.

Maintenance of the link between two mates in paired FASTQ files is solely based on identical ordering of reads between the two files. If these files are trimmed separately the order can be disrupted by discarded reads. The mates will no longer be correctly paired, preventing operations that depend on paired ends. This problem is solved by removing both mates from the paired files if one mate is lost in the trimming process. Usable reads whose mates were discarded can be retained as unpaired data.

Common Parameters

There are several parameters that are shared between trimming programs. Trimming from the $3'$ end of a read and trimming using a sliding window both require a minimum quality cutoff. This is often set at a Phred score of $Q = 20$ (Liu and Bassham, 2012). For adapter removal, query adapter sequences can either be included in the software (Krueger, 2013) or provided by the user (Lohse et al., 2012).

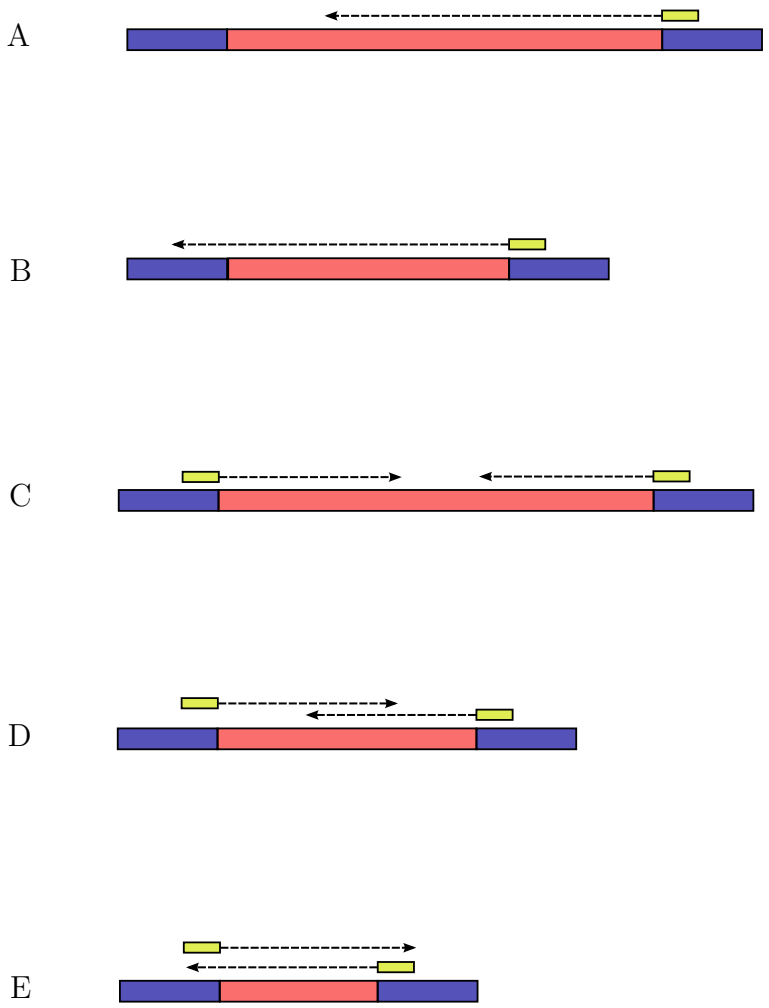


Figure 2.9: Possible events during Illumina sequencing that can be corrected by read filtering. (yellow = primer, blue = adapter, pink = insert) A) A single-end read without adapter contamination. The 3′ end of the read does not extend into the adapter. B) A single-end read that ran into the ′ adapter. C) A normal paired-end sequencing event. D) Reads share overlap at their 3′ ends. This overlap can be used to collapse the pair into a single unpaired read. E) Mates completely overlap and have identical adapter contamination at their 3′ ends. The adapter sequence can be trimmed and the overlapping mates can be collapsed.

The stringency for adapter alignments can be set by adjusting the mismatch allowance. After trimming, a length cut-off can be applied and this can be set by the user.

Software

There are many programs available for read trimming (Lindgreen, 2012). Often they are public releases of the in-house software of large bioinformatic and NGS consortia (Lindgreen, 2012; Lohse et al., 2012). These programs have evolved with the sequencing technology, incorporating paired-end trimming and increased efficacy in removing sequencing artefacts and low quality bases.

Cutadapt is a mature trimming program that is specialized for removing adapters and low quality bases from single-end data (Martin, 2011). It trims adapter sequences from both ends of each read. It can also remove low quality bases from either end of the read. It cannot perform sliding-window trimming over the whole read. This may not be a significant disadvantage since most low quality bases occur at the ends of the reads.

Trim Galore! is a recently released trimming program that was made to extend the functionality of Cutadapt (Krueger, 2013). It has the same quality and adapter trimming functions as Cutadapt, but it is capable of maintaining paired-end order in Cutadapt output. Trimmed reads that are shorter than a minimum threshold length can be filtered out of the data. When one mate in a pair is lost due to this threshold, the other mate can be discarded or saved as a separate unpaired output. Once the trimming is complete, Trim Galore! can be configured to automatically run FastQC to check the cleanliness of the trimmed FASTQ files.

Trimmomatic is an older trimming tool than Trim Galore! that was designed to trim paired-end data. It has several useful functions. It can remove adapters that are provided to it as FASTA sequences. Trimmomatic can only trim from the $3'$ end of the read where the majority of adapter contamination occurs (Lindgreen, 2012). Trimmomatic can use a sliding window method to trim over the entire length of the read and can also crop end from reads (Lohse et al., 2012). If reads fall below a length cutoff during trimming, Trimmomatic can discard them.

Btrim uses a bit vector algorithm to increase the speed of adapter trimming (Kong, 2011). It also improves on the memory efficiency of other trimming software. Btrim can be configured to remove both adapter sequences and low quality bases from the $3'$ end of the each read. Btrim doesn't have integrated paired-end trimming. Instead, it can output a detailed log file that records which reads are discarded during processing. An external Perl script can then be used to to read the log file and produce paired and unpaired FASTQ files to maintain the correct mate ordering.

AdapterRemoval is a new trimming program that aims to cover all of the shortcomings of other trimming software (Lindgreen, 2012). It can trim paired-end data and can merge mate pairs that share overlap. It is able to remove both $5'$ and $3'$ adapter sequences. For these merged reads, the overlapping region is given a new quality score based on the scores of the original mates. AdapterRemoval can trim low quality bases. It removes ambiguous base calls, represented by N, and trims low quality bases from both ends of each read using a minium Q threshold. AdapterRemoval succeeds in being a very complete trimming tool.

Read trimming software has improved greatly over the past year. The comprehensive tools, Trim Galore! and AdapterRemoval, make it easier to thoroughly trim

paired-end data. Read trimming is essential for generating a high-quality transcriptome assembly.

2.5 Transcriptome Assembly

2.5.1 The Overlap-Layout-Consensus Method

The method used for assembly of Sanger sequencing data is called overlap-layout-consensus (OLC) (Miller et al., 2010). The primary parameters for running a OLC assembly are word size, minimum overlap, and minimum percent identity. First, all possible words are generated from the reads. Then, all reads are searched for matching pairs of words. These words are used as seeds for extending overlaps between reads that meet the minimum overlap length and identity parameters. The overlaps are used to generate a directed graph connecting the reads (Miller et al., 2010). In the graph, each node is a read and each edge is a successful alignment (Figure 2.10). A path that passes through every node is calculated and a consensus sequence is generated by collapsing the graph (Compeau et al., 2011).

The OLC approach can be used to assemble Sanger data because a pairwise comparison of all reads is computationally possible. As the number of reads increases, this computation becomes orders of magnitude more time-consuming (Compeau et al., 2011). For 15,000 reads, 225 million pairwise alignments are required—a feasible calculation. Illumina sequencing can produce 100 million reads per flow cell lane. An OLC assembly of this read set would require 10^{16} pairwise comparisons, making this method inapplicable to *de novo* assembly of Illumina data (Boisvert et al., 2010).

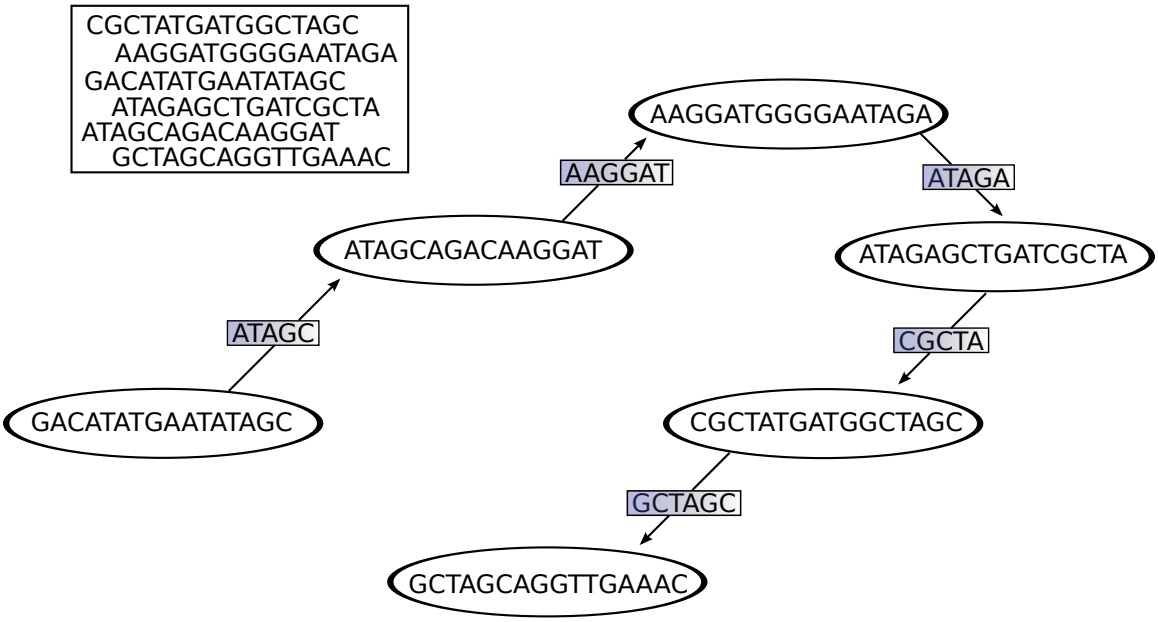


Figure 2.10: In the OLC method of sequence assembly, reads (top left) are used to build a directed graph. Each node (ellipse) represents a read and each edge (rectangle) represents the significant overlap shared by the connected nodes. Collapsing this graph creates a consensus sequence.

2.5.2 The De Bruijn Graph Method

A mathematical structure called a de Bruijn graph (DBG) provides a possible solution. The read data are first used to generate a catalogue of k-mers. These are all possible sequences of size k that can be derived from the read data. To generate all k-mers for a read, a sliding window of width k is moved along the read base-by-base and each sequence of length k is recorded. Then, all possible, non-redundant k-mer overlaps are collected and each is represented as $k-1$. A directed graph is constructed from the k-mers and overlaps with each edge being a k-mer and each node being a possible overlap of $k-1$ (Compeau et al., 2011). K-mer edges are linked if they share a node (overlap) of $k-1$ (Figure 2.11). Generating this graph imposes a much simpler calculation than calculating pairwise overlaps (Boisvert et al., 2010) for every read. The consensus sequence can be determined from its constructed DBG by walking along the edges and collapsing the graph.

The growing graph can have points of ambiguity due to heterozygosity, highly similar isoform sequences, and sequencing error and low coverage (Simpson et al., 2009). As unambiguous stretches of nodes and edges are collapsed, points of ambiguity cause the graph to diverge and form branches or bubbles (Figure 2.12). Branches are divergences in the graph that do not re-enter the graph later and can represent either sequencing errors or isoforms. Dead end branches or tips can be trimmed or retained based on arbitrary minimum length cut-offs or checking k-mer support of the branches (Simpson et al., 2009). Branches with high k-mer support can be kept as possible isoforms (Schliesky et al., 2012). Bubbles are also caused by sequencing error or nucleotide polymorphism (Schliesky et al., 2012). They differ from branches in that they eventually re-enter the main graph (Figure 2.12A). These can be 'popped' by retaining the side of the bubble with higher k-mer support, while deleting the other.

Bubbles with high k-mer support for both sides can be retained as separate contigs.

Assembly is more complex than constructing a DBG from the reads, collapsing the graph, trimming branches, and popping bubbles. The graph must take into account both strands of the sequence by using the reverse complement of each read in the calculation (Simpson et al., 2009; Boisvert et al., 2010). Paired-end reads can provide valuable information, but it also complicates the assembly process. They can be used to scaffold contigs. Contigs that are bridged by multiple read mates are likely derived from a single sequence (Grabherr et al., 2011) and can be combined into a larger supercontig.

ABySS (Simpson et al., 2009) and in Ray (Boisvert et al., 2010) use the message-passing interface (MPI) to efficiently distribute graph construction across a cluster. This provides are major increase in speed. Although assembly with DBGs is more efficient than the OLC method, the graphs can still consume hundreds of gigabytes of computer memory and demand a lot of processing power. Popping bubbles and trimming branches is also a computationally laborious process. The DBG has become a central theme in *de novo* assemblers, but the further processing and function of graph construction varies widely between assemblers (Grabherr et al., 2011; Boisvert et al., 2010; Simpson et al., 2009).

2.5.3 Transcriptome Assemblers

Most transcriptome assemblers are genome assemblers that have been re-engineered to better handle transcriptome data. There are differences between genome and transcriptome assemblies that must be addressed by the software. Because the genome exists as a single copy in each cell, reads generated from genomic DNA should provide even coverage over the entire assembly (Birol et al., 2009). Thus in genome assembly,

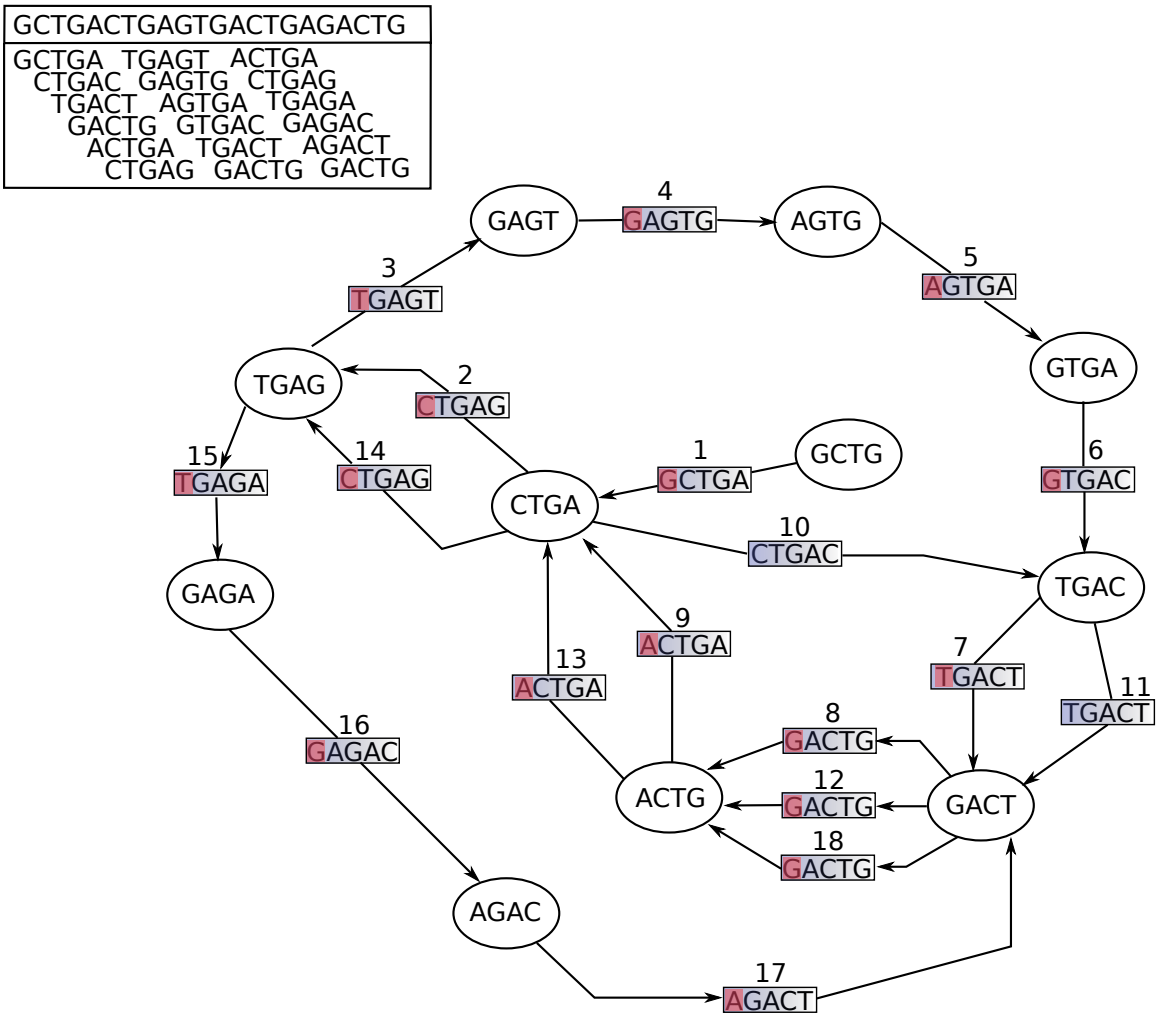


Figure 2.11: *K-mers* ($k = 5$) are generated by sliding a five base-pair window along the read and recording all k-mers (top left). A non-redundant list of all $k-1$ overlaps is generated and the k-mers (blue rectangle) are linked at nodes (ellipse) based on overlap, resulting in a de Bruijn graph. The consensus sequence is (red highlighted characters) generated by reading along the graph edges, collapsing the overlaps.

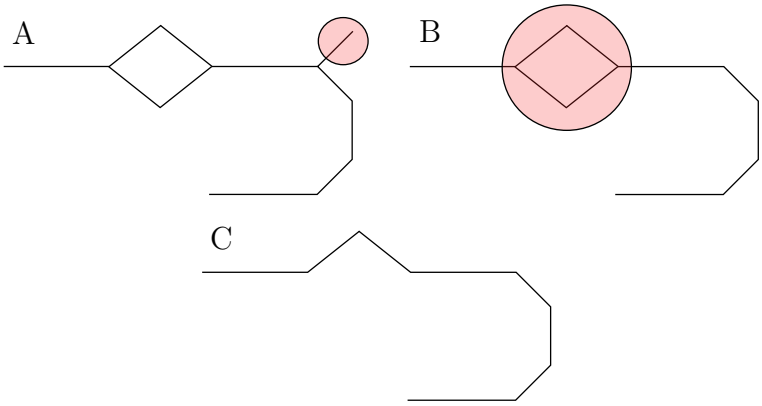


Figure 2.12: A subgraph resulting from large-scale collapsing of DBGs. A) Branches (tips) are divergences that do not later merge with the graph. They are removed by trimming if they are not supported by read data. B) Bubbles are divergences in the DBG that later merge with the main graph. They are eliminated by popping, which uses k-mer coverage to choose the valid side of the bubble. C) A collapsed and finished graph representing a contig concensus sequence.

low read coverage of a branch or bubble suggests that it is the result of a sequencing error. The transcriptome can have large variations in coverage between transcripts due to differential expression. This makes it difficult to confidently trim DBGs using k-mer coverage as a supporting parameter. Transcripts and isoforms with low, but significant expression levels would be removed in this process (Schulz et al., 2012). Some developers have released software to help their genome assemblers generate more robust transcriptome assemblies. Two examples are Velvet and ABySS. Trinity is the only assembler that is distinctly intended for transcriptome assembly.

Velvet was one of the first major DBG-based assemblers for Illumina sequence data (Zerbino and Birney, 2008). Its design focused primarily on accounting for sequencing errors and resolving repetitive DNA elements found in mammalian genomes. It was optimized for the very short reads (35 – 50bp) generated by early Illumina instruments. Oases, one of the most recent transcriptome assembly tools, is written by the developers of Velvet to process Velvet contig assemblies into probable transcriptional units.

Oases relies on Velvet to produce rough contig assemblies from read data (Schulz et al., 2012). Rather than using a single k-mer length, a range of low to high k-mer lengths are used in the interest of exploiting the benefits of both. Using shorter k-mer lengths can produce many, short contigs. Because the overlap required to connect two k-mers is shorter, incorrect paths can be made (Zerbino and Birney, 2008). The benefit of short k-mers is that they can also assemble contigs that have low read coverage (Schulz et al., 2012). Using long k-mer lengths produces a conservative set of contigs. The overlaps used to connect k-mers are long and therefore much more stringent than short ones. By carefully merging assemblies at a series of k-mer lengths, the sensitivity of short k-mers can be combined with specificity of long k-

mers (Schulz et al., 2012). In Oases, this set of assemblies is first scaffolded using paired ends, then similar contigs are grouped into clusters. Oases attempts to reduce these clusters by removing redundancy before extracting contigs from each cluster. The extracted contigs are merged in a final DBG with a k-mer length chosen by the user (Schulz et al., 2012).

ABySS is an genome assembler described by Simpson et al. (2009). It assembles contigs in a similar fashion to Velvet. The major advantage of ABySS is that its DBG contruction step can be distributed over a computing cluster using OpenMPI. Like Velvet, ABySS is not able to produce good transcriptome assemblies without an additional step. Robertson et al. (2010) have developed transABySS to allow ABySS to generate acceptable transcriptome assemblies.

TransABySS is an analytic pipeline published by the developers of ABySS that attempts to produce a set of complete transcripts from ABySS output (Robertson et al., 2010). It functions very similarly to Oases. The user must generate a series of ABySS assemblies ranging from short (21–25) k-mer lengths to long k-mer lengths that approach the length of the reads. TransABySS performs pair-wise alignments between assemblies with adjacent k values, removing contigs that are completely aligned within a contig in the other assembly. This proceeds through all of the assemblies until a final reduced set of contigs is produced (Robertson et al., 2010). The small contigs from the reduced set are then mapped to the rest of the contigs. Small contigs with no paired-end connection to the rest of the contigs are removed along with small contigs that did not have alignments to at least two other contigs.

Trinity is a *de novo* assembler that is built with the sole intention of constructing transcripts and isoforms from Illumina read data (Grabherr et al., 2011). Unlike

Velvet and ABySS, Trinity builds graphs at multiple steps. The first step, Inchworm, focuses on generating all possible transcripts and portions of transcripts by walking every possible path through directed graphs assembled from the reads. Chrysalis is the next step and clusters the Inchworm contigs into groups of contigs that share sequence similarity of at least $k-1$. The contigs are joined into DBGs at these points of similarity if the junctions are supported k-mer coverage (Grabherr et al., 2011). The DBG contructed for each Chrysalis cluster is called a 'component'. The final step is Butterfly. Graphs from Chrysalis are resolved into transcripts and isoforms based on read coverage and paired-end support. Sequencing errors are mitigated at this step by removing branches with low read support.

Velvet-Oases, ABySS-transABySS, Trinity, and other NGS assemblers can require large computing resources depending on the size of the input read data. They have been programmed to run on powerful single computing nodes or clusters. Velvet and Trinity can run only on single nodes or on a GSM system (Zerbino and Birney, 2008; Grabherr et al., 2011). Although Trinity can run well on a GSM system, its performance scales better with CPU-count on a single computer node. This is due to latency between the nodes in a GSM system (Henschel et al., 2012). ABySS uses OpenMPI to start hundreds of processes on different nodes to increase the speed of DBG-contruction (Simpson et al., 2009). After the graph is constructed branches and bubbles are processed on a single node. In this way, ABySS is only partially cluster-capable. Trans-ABySS is a collection of scripts and programs that are not capable of distributing processes across nodes (Robertson et al., 2010).

2.5.4 Output

The primary output of most *de novo* assemblers for NGS data is a FASTA file containing contigs. Assemblers can also leave a collection of intermediate and auxiliary data generated from the assembly process. In the case of ABySS, these files contain discarded graph paths, discarded bubbles and branches, and unscaffolded contigs (Simpson et al., 2009). These files can subsequently be useful for trans-ABySS, which uses the unscaffolded contigs and discarded paths in portions of its pipline. Trinity's output is more complex. Because it incorporates isoform prediction, it produces FASTA output in which each entry is an isoform that is linked to a gene. Each gene is linked to one or more isoforms. These data can be entered into counting software to produce isoform-specific expression levels without a reference genome (Trinity Team, 2013).

2.5.5 Further Assembly

The contigs produced from a *de novo* assembly of RNA-Seq data should represent transcripts, or in the case of a Trinity assembly transcripts and isoforms. This is not always the case. Often, far more contigs are produced than are expected transcripts (Schliesky et al., 2012). Expressed sequence tag (EST) assemblers can be used to assemble these contigs into larger sequences that may more closely resemble real transcripts (Surget-Groba and Montoya-Burgos, 2010; Duan et al., 2012; Liu et al., 2012b). TIGR Gene Indices Clustering Tool (TGICL) is the EST assembler used to produce the TIGR transcript assemblies (Pertea et al., 2003). It has not been maintained for several years. CD-HIT-EST (Li and Godzik, 2006) has offered a fast alternative. These tools can be used with high stringency parameters to reduce redundancy in the contig output of a *de novo* assembler.

A *de novo* assembly provides a base for further analysis including expression profiling and transcript annotation. Newer assemblers such as Trinity are able to resolve whole transcripts and isoforms as opposed to modified genome assemblers such as ABySS/transABySS and Velvet/Oases.

2.6 Annotation

RNA-Seq can sequence and quantify every transcript in a transcriptome. It is not limited to a set of known transcripts as microarrays are. The downside of this is that an annotation must be found for as many transcripts as possible. One way to do this for each RNA-Seq transcript is finding significant sequence similarity to a known transcript with an annotated function. This is commonly done using large scale BLAST database comparisons.

2.6.1 BLAST

Basic local alignment search tool (BLAST) is a very well-known and mature software package for fast local alignment of a query sequence against a database of subject sequences (Altschul et al., 1997). Performing BLAST searches on multiple databases for hundreds-of-thousands or millions of sequences can pose a major computational barrier that is most readily removed by using a computer cluster.

The standard NCBI release of BLAST does not support distribution of BLAST searches over a compute cluster. Two ways of harnessing the power of a cluster for BLAST searches are to either break up the queries in smaller, more manageable portions and process these in parallel or share a BLAST database in memory for multiple BLAST processes distributed across multiple computing nodes.

Breaking the query FASTA file into smaller parts and running many BLAST processes in parallel is relatively simple. The query file must be broken into parts that can be processed in a timely manner. Initiation of each BLAST job, constant submission of new jobs as others complete, and error checking of the results can be automated so that the entire query data set can be submitted and left to complete without user interaction.

Implementation of shared memory systems in parallel BLAST applications has been sparse. This may be because the previously described method is very simple while developing software for true parallel computing is quite complex. The most memory intensive part of running BLAST searches is loading of the reference database into memory (Altschul et al., 1997). By breaking a query file up and running the parts individually, the database is loaded into memory for each job. Both memory and time are wasted with this approach. Using MPI or other approaches allows the memory-committed database to be shared among processing nodes, reducing the total consumed memory.

mpiBLAST is an MPI-based implementation of BLAST (Lin et al., 2011). It uses the MPI standard to fragment the BLAST database across multiple nodes. Processes on the nodes are able to access the database to search for alignments. Results are gathered from all the processes and collected in a single output file. mpiBLAST is available for download, but is not currently maintained. It has not been updated to integrate more recent changes in the NCBI's BLAST releases. ScalaBLAST is similar to mpiBLAST in its approach. It distributes processes across nodes, but uses the Global Arrays toolkit(Oehmen and Nieplocha, 2006) rather than MPI. It also has the advantage of continued regular updates.

2.6.2 Databases

BLAST searches across multiple databases increase the chance of finding a possible annotation for each contig.

The Universal Protein Resource (UniProt) is a database intended to provide a low-redundancy set of accurately annotated protein sequences (Bairoch et al., 2005). Because of the high quality of the annotations and the redundancy of the sequence entries, UniProt is an excellent resource for finding robust functional annotations. Uniprot's high standard for functional annotations means that it should be combined with a database with a larger set of annotations.

The NCBI's nr (non-redundant) collection is a composite of several protein sequence resources. It contains non-redundant CDS translations of Genbank sequences excluding environmental sequences (NCBI, 2009). It also brings together several protein resources: PDB a database of structurally defined proteins, SwissProt which is also a component of UniProt (Bairoch et al., 2005), the protein information resource (PIR) a collection focusing on classification by evolutionarily significant domain similarities (Wu, 2003), and the Protein Research Foundation database. Annotating with the nr database gives a higher chance of finding hit for more transcripts.

Phytozome is a collection of plant genomes with gene models that is intended for evolutionary comparisons (Goodstein et al., 2012). The Phytozome resources also include the TAIR sequence database (Goodstein et al., 2012), a curated resource containing characterized sequences from *Arabidopsis*, the best-characterized model plant. High-scoring *Arabidopsis* BLAST alignments provide the most robust plant-specific annotation.

2.7 Expression Profiling

2.7.1 Read Mapping

In order to generate quantitative expression data, reads from each experimental condition must be mapped to a reference sequence. In a few years, this has progressed from a process taking days or weeks to a matter of hours on a desktop computer (Langmead and Salzberg, 2012). The output of read mapping is a SAM file describing the sequences of the reference and the reads, the mapping location of each read, and the quality of the alignment (Li et al., 2009). Production of this file is the first step in generating quantitative expression data from read alignments.

Software

BWA is a read alignment program that uses a Burrows-Wheeler transform (BWT) to provide an index of the reference sequence. This index can be queried very quickly (Li and Durbin, 2009) and is the primary method used in modern read alignment software. BWA is paired with another program called BWA-SW. BWA-SW improves on the speed of alignment of reads between 200–1000 bases by using the Smith-Waterman alignment algorithm (Li and Durbin, 2010). This is intended for mapping reads from 454 sequencers.

Bowtie is a read aligner that can achieve high alignment speeds of 30 million reads per hour (Langmead et al., 2009). It is slightly outperformed by BWA, especially for single-end reads (Li and Durbin, 2009). Like BWA, it uses a Burrows-Wheeler transform to reduce alignment time. Bowtie is slightly limited in its ability to map longer Illumina reads quickly as it was designed for earlier generations of Illumina instruments that produced shorter reads. It can not produce gapped alignments, possibly increasing the number of unaligned reads. Bowtie2 is successor to Bowtie

that optimizes the aligner for longer reads and gapped alignments. Bowtie2 is faster than Bowtie and BWA(Langmead and Salzberg, 2012). Bowtie has an additional advantage in that it is well integrated into the read-counting package, RSEM. RSEM is specially designed for counting reads mapped to *de novo* transcriptome assemblies (Li and Dewey, 2011). It saves work for the user by performing Bowtie alignment, SAM sorting, and indexing automatically. RSEM cannot use Bowtie2 because it is incapable of counting reads with ungapped alignments. The combination of Bowtie and RSEM is excellent for easily generating accurate read counts based on mappings to a reference transcriptome. Bowtie2 is best if RSEM is not being used for counting or if the reads are being aligned to a reference genome. BWA-SW is the best choice for mapping 454 reads or other long read data.

2.7.2 Read Counting

Reads that are aligned to a reference transcriptome or genome can be used to calculate expression levels for each genetic feature. Though this appears to be a simple task, it can become more complicated when the reference is a transcriptome assembly. Some programs stack the reads on genetic features and count the number of reads per feature. This works well for genome mapping. Newer software takes into account reads that map to multiple reference sequences in a transcriptome reference.

Samtools

Samtools can be used to produce 'pileup' files (Li et al., 2009). These are files that describe the reads aligned to each genetic feature. This is in contrast to the SAM format where reads are listed with the location and feature that they map to. Simple line counting scripts can be used to calculate the number of reads mapped to each feature. This process does not take into account ambiguously aligned reads and the

mapping program is configured to print only the top read alignment into the SAM output.

HTSeq-count

HTSeq-count is an alternative to Samtools that directly counts the reads-per-feature from a BAM file (Anders, 2010). This bypasses the need for a pileup file or counting script. HTSeq-count is intended for quantifying reads mapped to a reference genome and requires a GFF or GTF file to define genomic features in the reference genome sequences. Utilizing the genome features definitions, HTSeq-count is able to detect reads that span non-coding areas and discard these reads from counts. Though this process is indispensable for genomic mappings, the requirement of a GFF/GTF file complicates the process for counting reads aligned to transcript sets.

RSEM

RSEM calculates the number of reads aligned to each genetic feature and accounts for ambiguous mappings. Reads can map ambiguously due to incorrect resolution of isoforms and read alignment to two closely related genes (Li and Dewey, 2011). These ambiguous mappings, called multireads, can comprise over 50% of aligned reads (Li et al., 2010a). Previous methods, such as Samtools pileup and HTSeq-count have based analysis only on the highest-scoring alignment for each read alignment.

RSEM starts by mapping reads to a reference file using Bowtie (Li and Dewey, 2011). Instead of recording only the best alignment for each read, Bowtie records every possible alignment for each read. RSEM uses these multiple alignments and their quality scores to determine count estimates for each reference genetic feature. If the reference contains both gene and isoform information, isoform- and gene-level count estimates can be calculated. The count estimates calculated by RSEM are

mainly non-integer numbers because counts are split between reference transcripts. This can complicate downstream analyses that require integers count values.

2.7.3 Normalization

There are two possible levels of normalization in RNA-Seq data. Within a library, counts must be normalized between transcripts because a longer transcript will generate more sequencing reads than a shorter transcript even if they are equally expressed. This normalization is done by weighting read counts based on transcript lengths. Expression levels must also be normalized between libraries to address differences in coverage between libraries. Expression data normalization tools are a diverse group of RNA-seq analysis software that are generally implemented in the R statistical language (Dillies et al., 2012; Sun and Zhu, 2012; Bullard et al., 2010).

RPKM

One of the earliest normalization techniques was reads per kilobase per million mapped reads (RPKM). It was initially designed to normalize transcript levels within a library (Dillies et al., 2012). Because it takes into account the total number of mapped reads in a library, it is also used for inter-library normalization to remove coverage bias. However, this method has been shown to introduce bias when performing differential expression analysis between libraries (Dillies et al., 2012; Bullard et al., 2010). The RPKM method is still popular despite this major shortcoming.

RSEM

RSEM performs normalization within libraries while producing count estimates (Li et al., 2010a). Therefore, counts calculated by RSEM must be subsequently normalized between libraries. Because RSEM's counts are non-integer values, any

inter-library normalization must be able to accept non-integers as input.

TMM

Trimmed mean of M values (TMM) is a method integrated into the edgeR package, a common RNA-Seq analysis package written in R (Robinson et al., 2010). TMM is considered to be a robust method, but has some limitations due to its integration into edgeR. The TMM algorithm is run at the same time as DE calculations (Dillies et al., 2012). Normalized counts calculated by TMM cannot be exported to a separate file. Calculating DE using edgeR is most robust when using data containing biological replicates, an expense that is too costly for many researchers.

DESeq

The DESeq normalization method is integrated into the DESeq RNA-Seq analysis package (Anders and Huber, 2010). Like edgeR it is implemented in R and both normalizes libraries and assesses the significantly DE transcripts. Inter-library normalization in DESeq is integrated into its accurate DE calculation pathway (Dillies et al., 2012). It can accept only integer input and is unable to accept count estimate data from software such as RSEM. DESeq is primarily intended for handling count data that include biological replicates (Anders and Huber, 2010).

Conditional Quantile

Conditional quantile normalization (CQN) method is a newer method introduced in the R-package *cqn* (Hansen et al., 2012). CQN can normalize expression data based on any biasing values chosen by the user. This is most often GC-content, a source of bias not accounted for by other normalization software. CQN is shown to have increased expression level precision between biological replicates (Hansen et al., 2012).

Unlike the TMM and DESeq methods, CQN's only function is count normalization and is not integrated into a complete DE analysis package. This allows the user to produce normalized $\log_2$(RPKM) values. CQN also accepts non-integer estimated count values that are the output of more advanced counting software such as RSEM.

Conclusion

Many RNA-Seq R packages attempt to perform multiple functions including read counting, normalization, and DE calculations. Only a few methods are devoted only to normalization. RPKM has been shown to be inferior to most other normalization methods, while CQN performs well and is versatile in what data it can use as input. The result of normalization is a set of absolute expression levels for each transcript in across samples.

2.7.4 Differential Expression

Normalized expression values provide the basis for answering biological questions. Usually, most interest lies in transcripts that are differentially expressed between samples. Statistically significant differential expression can be calculated by a number of R packages. These calculations are complicated by the common absence of biological replicates in many RNA-Seq experiments. A lack of biological replicates makes it impossible to apply statistical analysis to determine whether a transcript is significantly differentially expressed between two samples.

edgeR

edgeR is an R package introduced by Robinson et al. (2010) for calculating differential expression. It uses a parametric method (Tarazona et al., 2011) that attempts to fit count data to a Poisson distribution (Robinson et al., 2010). These distribu-

tion is used because edgeR's primary assumption is that differential expression is a rare event. Most transcripts or genes are not expected to be differentially expressed between samples (Tarazona et al., 2011). This assumption, while true in many cases, does not hold in every situation.

DESeq

DESeq takes a similar approach to edgeR. It also makes the assumption that DE is a rare event, but it attempts to fit differences in expression to a negative binomial distribution rather than a Poisson distribution (Anders and Huber, 2010). This is an attempt to improve upon the false discovery rates generated by Poisson-based DE software such as edgeR. EdgeR predicts more significantly DE genes because it underestimates variance in read counts (Anders and Huber, 2010). DESeq also improves over edgeR in its dynamic range of detection. While edgeR overestimates DE for transcripts with low expression levels and underestimates DE for transcripts with high expression levels, DESeq's DE estimates are not influenced by extremes of expression. DESeq is developed as a tool for taking integer read counts as input. It has a robust, built-in normalization algorithm that internally generated data for DE analysis. It is not able to accept non-integer read counts, such as those generated by RSEM. DESeq is able to calculate significant DE based on data lacking biological replicates, but it functions best with replicates.

NOISeq

DESeq and edgeR are both intended to be used for data with biological replicates. Although they can be used without replicated data, NOISeq has the advantage of being able to simulate technical replicates. These pseudoreplicates produce DE estimates that are very close to those produced using true technical replicates (Tarazona

et al., 2011). NOISeq offers another advantage in producing lower false discovery rates than both edgeR and DESeq. It accomplishes this by using a non-parametric test, rather than a parametric test. edgeR and DESeq find more significant DE as the size of the library increases. NOISeq produces stable DE calls over large ranges of coverage (Tarazona et al., 2011). NOISeq accepts non-integer input counts and can be configured to skip its internal normalization process if the input is already normalized.

2.8 Conclusion

Deciding on a sequence of software to use for RNA-Seq is a long process. The continual release of new software and updates for older software makes the process more challenging. Some steps in the workflow have many tools available that must be compared to determine the optimal choice. However, testing any software can take time because the data files are so large.

FastQC is the standard for evaluating the quality of Illumina read data. It is user-friendly and generates high-quality plots outlining read quality and nucleotide composition. Read trimming software is gradually improving. Trim Galore! and AdapterRemoval trim paired-end reads and effectively remove Illumina adapters and low quality bases. *De novo* assembly is most computationally intensive step in RNA-Seq analysis. Trinity has taken a dominant position in *de novo* transcriptome assembly because it is able to predict transcript isoforms rather than every possible contig.

Read-mapping is the subject of several well-developed programs, each of which is suited to different applications. The alignment file formats, BAM and SAM are

well-designed and standardized. They have already been extensively implemented by researchers. Counting read alignments from BAM and SAM files is becoming a task that entire applications are dedicated to. RSEM specializes in accounting for ambiguous read-mapping to transcriptome references. It has strongly improved on preexisting methods.

Normalization and DE analysis of read counts is an area dominated by R programs. Normalization is still a controversial issue and many different methods are available for intra- and inter-library normalization. Novel sources of bias in Illumina counts are accounted for in each new R package. The CQN package has many advantages. It is dedicated only to normalization, it accounts for bias introduced by GC-content, and it can take RSEM output as input. edgeR and DESeq are the primary R packages for assessing DE between samples. They are both intended for analysis of data with biological replicates and integers read count inputs. NOISeq is an alternative package that is partially optimized for non-replicated data.

Once an analysis workflow has been established, it only needs to be maintained and updated with superior software and newer software versions. Eventually this process may be supplanted by more automated systems as has occurred in microarray analysis. The development of RNA-Seq analysis tools is accelerating just as development for microarray tools accelerated towards a more robust, standard method. For now, the diversity of available software must be adapted by the research to each biological problem.

Chapter 3

Transcriptomics of Douglas-fir Ovular Development

3.1 Introduction

3.1.1 Seed Development in Douglas-fir

Pollination occurs in Douglas-fir in early to mid-April. Pollen grains land on the stigmatic tip at the micropylar end of the ovule. The pollen is taken into the micropylar canal and can remain there for up to three weeks prior to germinating (Owens and Morris, 1990). The pollen grain elongates and a pollen tube forms to penetrate the nucellus and archegonium (Owens and Morris, 1990). Within the archegonium, the pollen tube bursts, releasing sperm that then fertilize the egg cell.

Prior to fertilization, the ovule undergoes extensive development. The megagametophyte enlarges and forms cell walls and archegonia (Chiwocha and von Aderkas, 2002). Soon before fertilization, an egg cell forms within the archegonium (Allen, 1943). Douglas-fir develops a megagametophyte and egg cells regardless of the pol-

lination state of the ovule. The survival of the ovule is decided at fertilization. If fertilization does not occur, the megagametophyte begins to abort two weeks after fertilization would have occured (Rouault et al., 2004). Abortion appears to occur by PCD.

PCD is involved in a number of events during conifer seed development. These include PCD-mediated softening of the nucellus to ease pollen tube penetration prior to fertilization (Hiratsuka et al., 2002), the formation of the corrosion cavity and production of its fluid, elimination of subordinate embryos in polyembryony (Filonova et al., 2002), and breakdown of the megagametophyte after germinationc (He and Kermode, 2003a,b). These processes are associated with processes typical of autolysis as well as PCD hallmarks such as nuclear fragmentation and high levels of protease activity. None of the processes have been genetically characterized. RNA-Seq may offer a way to characterize differential gene expression during ovular abortion in Douglas-fir.

3.1.2 RNA-Seq

Illumina Sequencing is a high-throughput method for sequencing DNA. Illumina reads are currently limited to 150 bases (Illumina, 2012). The technology compensates for short read lengths by producing enormous coverage at a lower price. Reads from sequenced cDNA can be used to calculate expression values based on the number of reads aligning to regions within a reference sequence (Langmead and Salzberg, 2012). Genomes are frequently used as mapping references, but they are not available for all organisms. *De novo* transcriptome assembly is used to generate putative transcripts from read data (Grabherr et al., 2011). These read data can then be mapped back to the *de novo* transcriptome to create expression profiles. Chapter 2 provides a detailed description of an RNA-Seq workflow.

3.2 Methods

3.2.1 Material Collection

A population of Douglas-fir clones (genotype 970) at Mount Newton seed orchard (Victoria, BC) was selected for collection of megagametophyte tissue. Branches with high number of female reproductive buds were manually emasculated by removing male buds in March 2011. These branches were covered with paper exclusion bags to prevent pollination.

Following bud burst and the development of reproductive receptivity (April 2011), half of the bagged branches were injected with a mix of pollen collected from multiple genotypes and shaken to ensure distribution of the pollen. The other half were left unpollinated. Some weeks after pollination in all seed orchard and natural surrounding populations was complete, the paper bags were removed and replaced with mesh insect exclusion bags.

Megagametophytes (200-400 per treatment) were collected on four dates: June 10, June 22, June 30, and July 6These dates were chosen to coincide with events in seed development including corrosion cavity formation, fertilization, embryogenesis, and the abortion of unpollinated megagametophytes. Bracts and scales were collected to compare expression profiles in vegetative versus reproductive tissues. The cones were collected in small batches of two to four to minimize potential transcriptional changes due to wounding. Pollinated megagametophytes became opaque and developed embryos while unpollinated megagametophytes did not. Dissected tissues were immediately flash frozen in liquid nitrogen (LN) then stored in a -80°C freezer.

RNA Extraction

RNA extractions were performed using a method modified from Kolosova et al. (2004). Tissues were finely ground in liquid nitrogen with a mortar and pestle. Prior to RNA extraction, tissue powders were constanty kept at or below -80° in a freezer or LN2.

Tissue (200 mg to 400 mg) was added to 1.0 mL of extraction buffer (200 mM Tris-HCl pH 8.0, 1.5% lithium dodecylsulphate, 300 mM lithium chloride, 10 mM EDTA disodium salt, 1% sodium deoxycholate, 1% Tergitol NP-40), which was amended just prior to use with 1 mM aurintricarboxylic acid, 10 mM dithiothreitol, 5 mM thiourea, 2% w/v polyvinylpolypyrrolidone. The suspensions were immediately mixed by vortexing, then flash frozen in LN before being allowed to thaw at room temperature.

Thawed samples were centrifuged at 18000×g in a refrigerated centrifuge for 15 minutes at 4°C. The pellet was discarded and the supernatant was treated with 100 μL 3.3 M sodium acetate (pH 6.1) and 1000 μL ice cold isopropyl alcohol. The samples were mixed thoroughly and incubated at -80°C for one hour. After thawing the solutions were centrifuged at 8000×g for 30 minutes at 4°C. The resulting supernatant was drained from the pellet and discarded. The pellet was incubated on ice with 400 μL each of TE buffer (pH 8.0) and 5M sodium chloride. Frequent vortexing and pipette aspiration was used to resuspend the pellet.

The suspensions were mixed with 200 μL of 10% cetyltrimethylammonium bromide (CTAB), mixed thoroughly by vortexing, and incubated at 65°C for 5 minutes. The solutions were then extracted with 500 μL 25:1 chloroform-isopentanol by thorough homogenization followed by centrigfugation at 5000×g for 6 minutes at 4°C. The aqueous phase was extracted a second time, mixed with 8M lithium chloride, and

incubated at 4°C overnight.

The solutions were centrifuged at 18000×g at 4°C for 30 minutes to collect RNA pellets. The pellets were washed with 500 μL 70% ethanol and centrifuged again at 18000×g for 10 minutes at 4°C. The pellet was dried completely and resuspended in 30μL of nuclease-free water. RNA solutions were stored at -80°C.

DNAse Treatment

To eliminate contaminating genomic DNA from the extracted RNA, a DNA digest and subsequent RNA purification were performed. Reactions were 100 μL and contained one unit of rDNase I (Ambion, USA), 1X DNase Buffer (Ambion, USA), and 20 ug of total RNA.

Reactions were halted by the addition of 20 μL EDTA (pH 8.0) then extracted with 110 μL 25:24:1 phenol-chloroform-isopentanol, mixed, and centrifuged at 14000×g for 10 minutes. The aqueous extract was then extracted twice with 24:1 chloroform-isopentanol with 10 minute centrifugation (14000×g).

To precipitate RNA, the aqueous fraction was mixed with 10μL 3.3 M sodium acetate (pH 6.1) and 120 μL isopropyl alcohol. The solution was incubated at -20°C for 2 hours and centrifuged at 14000×g for 30 minutes at 4°C. The supernatant was discarded and the pellet was washed in 100μL of 70% ethanol and collected by centrifugation at 14000×g for 10 minutes. The pellet was dried, dissolved in 20μL nuclease-free water and stored at -80°C.

RNA Quality Assessment

RNA quality was assessed using using the BioRad Experion^{TM} system based on the manufacturer's instructions. This capillary electrophoresis system asses RNA quality based on RNA length distributions and rRNA quantity ratios.

3.2.2 Transcriptome Sequencing

RNA was sequenced at Canada's Michael Smith Genome Sciences Centre in Vancouver, Canada. Four multiplexed samples were sequenced on each of three Illumina HiSeq2000 lanes. Two of these samples consisted of Douglas-fir nucellus which was not used in the experiment. Paired-end reads were 100 bp in length with target insert sizes between 200 and 30 bp. The lanes were sequenced at different dates. The reads generated for each library were packaged in BAM files.

3.2.3 Data Preprocessing

The BAM files were converted to paired-end FASTQ files with bamToFastq from the Hydra-SV package (Quinlan et al., 2010). These FASTQ files were evaluated for base quality, nucleotide overrepresentation, and sequence overrepresentation using FastQC (Andrews, 2012). The quality scores and sequence overrepresentation data were used to direct FASTQ trimming.

Reads were trimmed with Trimmomatic (Lohse et al., 2012). A minimum Phred cut-off of 20 was used for quality trimming. To remove as much adapter contamination as possible, four conserved sequences from the Illumina adapters identified by FastQC were used as queries for trimming:

5′-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3′

5′-CACTGTCCTCAAGTCTGCACACGAGAAGGCTAG-3′

5′-ATCTCGTATGCCGTCTTCTGCTTG-3′

5′-CAAGCAGAAGACGGCATACGAGAT-3′

The first nine bases of each read were also removed because of very high GC-content reported in the bases by FastQC. Any reads that fell below 36 bp in length during the trimming process were discarded.

In some mate-pairs, one mate was discarded while the other was retained. These retained, unpaired mates were assessed with FastQC and trimmed again with Cutadapt (Martin, 2011) to remove any remaining adapter contamination and low quality bases. All adapters found in the each unpaired file were used as input trimming queries for Cutadapt. The paired and unpaired data for each library was assessed with FastQC after trimming with Trimmomatic and Cutadapt to ensure that Illumina adapter sequences were removed and there were no low quality bases.

3.2.4 De novo Assembly

The trimmed reads were pooled and used to assemble 31 ABySS 1.3.2 (Simpson et al., 2009) assemblies at odd k-mer values from 21 to 81. For the initial graph-building step of ABySS, the Westgrid parallel computer cluster, Nestor, was used. Each assembly was limited to 130 processors and 260 gigabytes of memory. The graph processing was performed separately on Breezy (Westgrid). Each assembly utilized 24 processors and 256 gigabytes of memory. Parameters for the ABySS assemblies were configured according to the transABySS user manual (Chiu and Nip, 2012). The options used were: `OVERLAP_OPTIONS='--no-scaffold'`, `SIMPLEGRAPH_OPTIONS='--no-scaffold'`, and `MERGEPATHS_OPTIONS='--greedy'`. The ABySS-transABySS assembly process

took approximately two weeks. Each completed assembly was assessed with abyss-fac. These data were used to assess the effect of k-mer length on the assembly of Douglas-fir transcriptome data. Figures were generated with R.

The output of the ABySS assemblies was merged into a single transcriptome assembly using the first step of transABySS 1.3.2 (Robertson et al., 2010). This process filters, extends, and merges the multiple input assemblies into one FASTA output file. This file was assessed with abyss-fac.

The large contig set produced by transABySS was processed to further remove sequence redundancy. CD-HIT-EST 4.6.1 (Li and Godzik, 2006) was used to cluster and merge sequences with greater than 90% similarity using a word-size of 10. The ABySS assemblies and the transABySS assembly were used as input to CD-HIT-EST. The merged assembly was assessed with abyss-fac.

3.2.5 Annotation

The merged assembly was annotated by querying each sequence against many sequence databases using BLAST 2.2.27+ (Altschul et al., 1997). BLAST databases were made for the BLAST non-redundant protein collection, UniProt, and 33 genomes stored in Phytozome. The assembly files were split into smaller (1000+) sets of queries that were than searched against each database in parallel on the Nestor/Hermes computing cluster. BLASTx was used for querying protein databses and BLASTn was used against the Phytozome genomes. For each query, the minimum e-value cutoff was set at 10^{-5} and only the best hit was retained.

The merged assembly was searched for open reading frames (ORFs) greater than 200 bp in length. This was accomplished with getorf from EMBOSS (Rice et al.,

2000). Sequences with a potential ORF or an annotation were compiled into a reduced sequence set that was used for further analysis.

3.2.6 Read Mapping and Counting

The raw FASTQ reads were mapped to the annotated reference using Bowtie and counted using RSEM (Li and Dewey, 2011). Bowtie's read-mapping functionality is integrated into RSEM. First, the reference FASTA file was prepared for RSEM using rsem-prepare-reference. The untrimmed reads for each library were then mapped and counted using rsem-calculate-expression. Sequences with estimated counts greater than 0 in at least one sample were joined into an expression table for the entire experiment.

3.2.7 Normalization

Inter-library normalization was calculated with the Bioconductor R package cqn (Hansen et al., 2012). Library size and GC-content were used as normalizing factors. Sequence length was not incorporated into the calculation because it is accounted for by intra-library normalization in RSEM. Normalized expression data were returned as log_2RPKM values.

3.2.8 Differential Expression Analysis

Differential expression analysis was performed in a pairwise fashion using NOISeq (Tarazona et al., 2011). Each library was compared to the each of the other libraries and transcripts with q-values greater than 0.9 were considered significantly DE. The mean expressions levels for the vegetative tissues and for the megagametophytes were compared to one another and the same q cutoff of 0.9 was used to select DE sequences between these sample groups.

Transcripts that were DE between the pollinated and unpollinated megagametophyte samples at one or more of the four sampling dates were used to generate large heat maps of the top fifty most differentially expressed genes for each treatment.

3.2.9 Quadratic Regression

To find sequences that followed defined expression patterns over the course of the experiment, the normalized expression data were fitted to linear and quadratic models. As an inital filtering step, only sequences that experienced a change in log_2RPKM of at least 2 were used in the analysis. For each transcript, the expression values were first fitted to a quadratic model and tested for significance. If the p-value of the F statistic was below 0.01, the sequence was classified as fitting a quadratic regression model. The expression values were fitted to a linear model if the quadratic fit wasn't supported. Sequences with p-values greater than 0.01 for both quadratic and linear models did not significantly fit either model and were not used for the rest of the analysis. Sequences identified in the regression models were filtered based on a minimum change in expression level between the time points.

Transcripts were further separated into expression categories based on the polynomial factors of the regression models: β_2 and β_1. For data that fit the quadratic regression, different signs for β_2 and β_1 were used to identify six different expression profiles: early or late increased expression (Table 3.1A,D), early or late decreased expression (Figure 3.1B,C), an increase in expression during the center sampling times (Table 3.1H), or a decrease in expression during the center sampling times (Table 3.1G). If data fitted the linear model, the sign of β_1 was used to identify linear increases and decreases in expression (Table 3.1E,F).

For each expression category, one heat map each was generated for the pollinated and unpollinated treatments. Different minimum changes in $\log_2$RPKM were used to find the around fifty of the most differentially expressed transcripts. These heat maps were used to identify transcripts possibly involved in seed development. Smaller heat maps were generated for these transcripts.

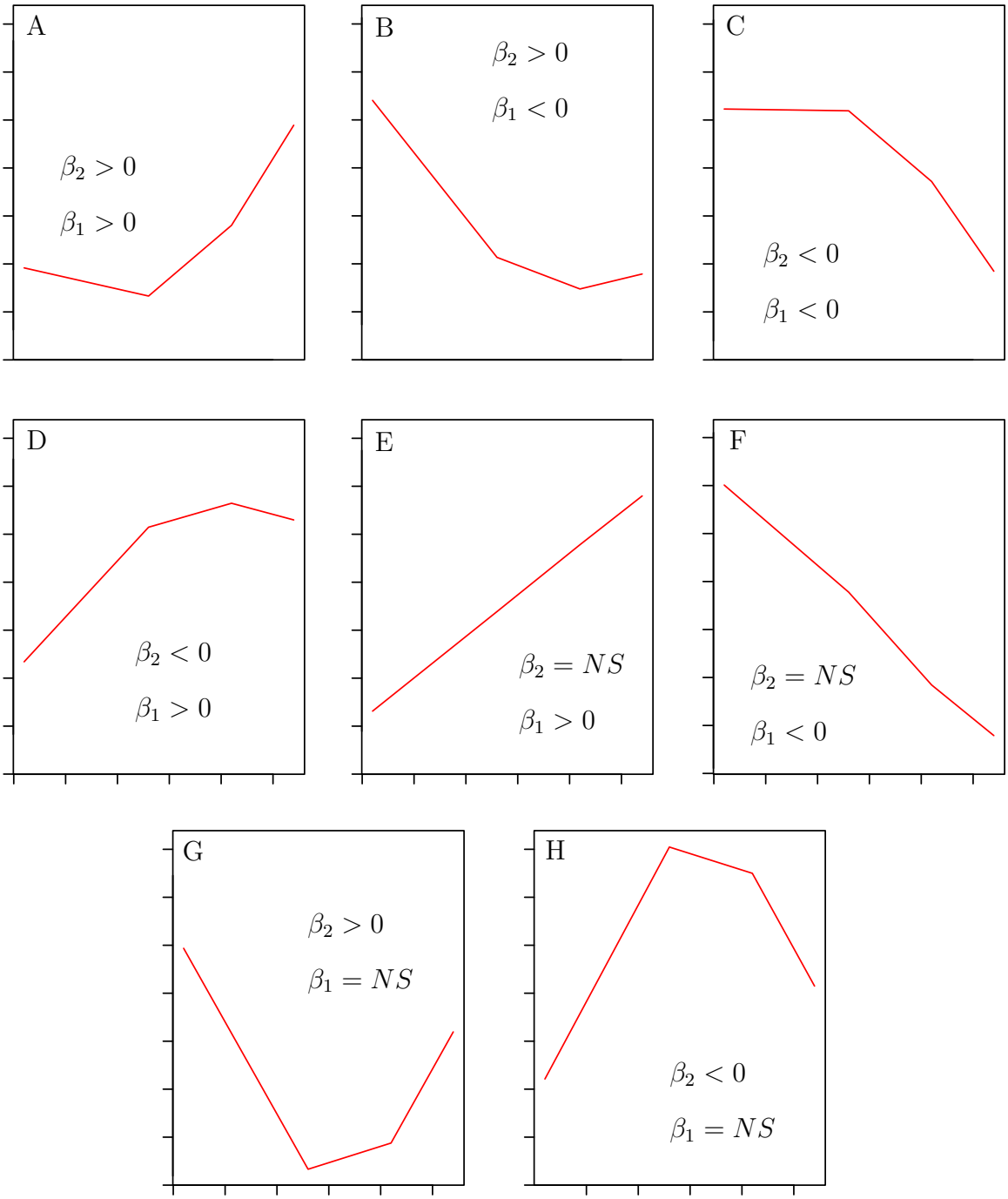


Figure 3.1: Example plots of quadratic regressions. A) In Expression profiles with late increases in expression, β_2 and β_1 are greater than zero. B) Expression profiles with early drops in expression fit regression with $\beta_2 > 0$ and $\beta_1 < 0$. C) Late decreasing transcripts fit regressions with both β_2 and β_1 being negative. D) Expression profiles with early increases in expression fit regressions with β_2 being negative and β_1 being positive. Linear increases (E) and decreases (F) fit regressions with $\beta_1 > 0$ and $\beta_1 < 0$ respectively; β_2 is not defined. Parabolic expression patterns have no defined β_1. Reduced expression midway through the experiment (G) fits a regression with a positive β_2 while increased expression (H) fits a regression with a negative β_2.

3.2.10 Finding PCD-related Genes

Candidate gene lists were populated in two ways. Protein sequences for genes known to be involved in plant and animal cell death processes (Table 1.1) were retrieved from Uniprot. These were supplemented with Swiss-Prot sequences from *Arabidopsis thaliana* and *Mus musculus*. Sequences with GO annotations related to programmed cell death (GO:0012501) and autophagy (GO:0006914).

The candidate genes list was used to construct a BLAST database against which all differentially expressed genes were queries. Candidate genes with significant differential expression between pollinated and unpollinated samples at each experimental date were used to construct a table of transcript annotations. This table was used to produce a conservative heat map of genes whose expression patterns indicate a possible role in megagametophyte abortion.

3.2.11 Heat Map Generation

Large heat maps were generated using Genesis (Sturn et al., 2002) and edited in Inkscape. In addition to heat maps of transcripts potentially involved in seed development and megagametophyte abortion, a heat map was generated showing putative transcripts expressed primarily in vegetative or megagametophyte tissues.

3.3 Results and Discussion

3.3.1 Data Analysis

RNA Quality Analysis

After extraction and DNAse treatment, the RNA extracts were analyzed on the Experion (Bio-Rad, USA) system. All samples were had high quality scores (RQI) of greater than 7.0 except for the July 6 unpollinated sample (Table 3.1). A decrease in quality appears to occur in the unpollinated samples over the course of the experiment, possibly indicating progressive RNA degradation.

Read Assessment and Trimming

The total read counts for each raw read library ranged from 61 million in June 10 P to 137 million in June 30 E (Table **??**). There was no visible trend in read count between the vegetative and megagametophyte libraries or between the pollinated and unpollinated megagametophyte libraries. Trimming the raw reads resulted in a loss of reads from each library. Libraries derived from the same Illumina run had similar decreases in read count. The bracts and scales had the lowest losses of 10.3% and 11.0% respectively. The four libraries from June 10 and June 22 had the highest losses of $32.8 \pm 1.28\%$. The final megagametophyte libraries, June 30 and July 6 had intermediate losses of $14.6 \pm 1.38\%$.

Transcriptome Assembly

ABySS was used to generate 31 contig sets using odd values of k from 21 to 81 (Table A.1). These assemblies are diverse in their sequence lengths and counts. The low k-value of 21 produced an assembly of 16.38 million contigs. Only 122,801 of these contigs were greater than 200 bp in length. The N50 length for this assembly

Table 3.1: Bio-Rad Experion results including 28S:18S ribsomal RNA ratios and RQI quality index from analysis of Douglas-fir megagametophyte RNA extracts. Samples were collected at four dates from pollinated (P) and unpollinated (E) cones. Extracts from Douglas-fir bracts, and scales are also shown. The highest possible RQI score is 10.

Date	26S:18S Ratio		RQI	
	P	E	P	E
June 10	1.31	1.03	8.4	8.2
June 22	1.50	1.27	8.7	7.6
June 30	1.76	0.99	8.9	7.3
July 6	1.23	0.91	8.6	6.7
Bracts	1.21		7.9	
Scales	1.24		7.7	

Table 3.2: Read counts assessed by FastQC. These include the counts from the raw libraries and the reads retained as pairs or lone mates after trimming. Counts are in millions.

Library	Raw	Paired	Unpaired	Reads Lost
Bracts	91.4	78.3	3.7	10.3%
Scales	150.0	128.2	5.7	11.0%
June 10 E	84.1	53.3	2.6	33.5%
June 10 P	61.6	40.7	2.1	30.6%
June 22 E	83.9	53.1	3.0	33.3%
June 22 P	95.5	60.2	3.0	33.8%
June 30 E	137.0	106.0	10.2	15.3%
June 30 P	81.5	65.1	5.7	13.2%
June 6 E	73.8	56.5	5.4	16.2%
June 6 P	88.5	70.9	6.6	13.7%
	947.5	710.9	47.9	19.9%

was 369 bp and 35,344 contigs surpassed it. The longest contig generated was 25,491 bp in length. These values change dramatically with increasing values of k.

As k increases, the total number of contigs drops (Figure 3.2). At $k = 81$, 173,788 contigs were produced–a fraction of the number at $k = 21$. The N50 length increases steadily with larger values of k as longer contigs make up a greater portion of the assembly. The maximum contig length plateaus at 269,991 bp between $k = 47$ and $k = 55$ before dropping to 106,337 bp. The sudden decrease in maximum contig length parallels large changes at $k = 55$ in the contig counts and N50 length (Figures 3.2 and 3.3). This suggests that there are a large number of reads that are not incorporated into contigs when the value of k is less greater than 55.

The ABySS assemblies were used as input for transABySS in order to produce a single merged assembly. The merged ABySS assembly consisted of 4,931,388 contigs with 743,739 having a length greater than 200 bp. The N50 length of the merged transABySS assembly was 1054 bp and there were 133,230 contigs over this length. The EST assembler CD-HIT-EST was used to reduce redundancy in this large set of contigs. The contig count decreased to 4,931,388 and the number of contigs over 200 bp in length was reduced to 356,498. The N50 length was 854 bp and 63,349 contigs were over N50. This assembly was used for downstream anaylsis.

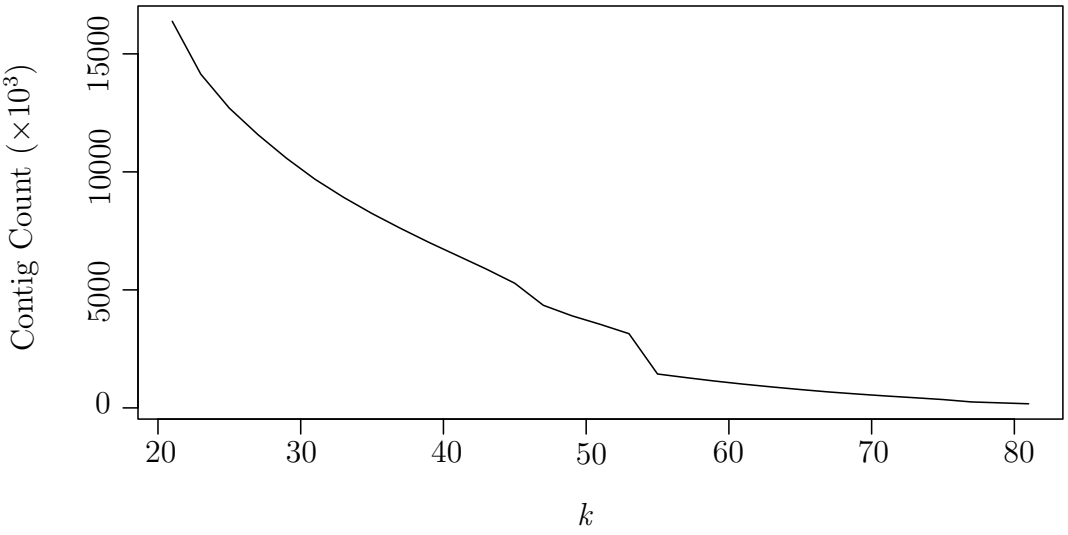


Figure 3.2: Kmer value versus contig counts

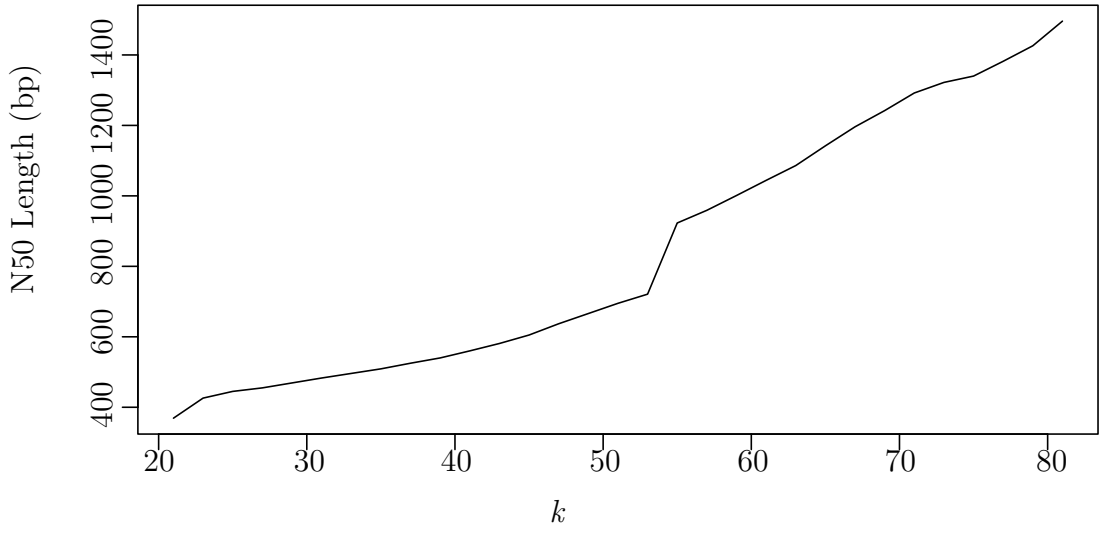


Figure 3.3: Kmer value versus contig counts

BLAST Annotation

The merged contigs from CD-HIT-EST were used as queries for BLAST searches. Three collections were used for these BLAST searches. The large NCBI non-redundant CDS translation and protein collection (nr) produced alignment to 319,696 contigs with e-values less than 10^{-5}. The smaller UniProt protein database produced 176,900 hits. The final BLAST searches used genomic sequences from Phytozome (Goodstein et al., 2012). The number of BLAST hits for each genome are detailed in (Table A.2). Of the 4,931,388 contigs used in BLAST searches, 392,276 had at least one hit in at least one of the databases. The contigs were combined with unannotated contigs with potential open reading frames (ORF) longer than 200 bp. The number of contigs that had an annotation, 200 bp ORF, or both was 500,886.

Expression Profiling

Reads from each library were mapped to the Douglas-fir reference transcriptome. The read mapping rates for each library are described in table A.3. There were 223,163 contigs with RSEM counts greater than one in at least one library.

Pairwise differential expressions between all megagametophyte samples are shown in table A.4. The largest numbers of differentially expressed (DE) contigs were found between different sampling dates. The largest number of DE transcripts (5,187) was found by comparing June 10 E and July 6 P. The fewest DE transcripts were found between pollinated and unpollinated samples from the same collection date. The mean expression levels for all megagametophytes and mean expression levels of bracts and scales were compared. Statistically significant DE in these sample groups is found in 12,490 transcripts (Table A.4).

Transcripts were fitted to quadratic regressions that represented expression profiles of biological interest (Figure 3.1). The number of transcripts fitting each regression category in pollinated and unpollinated samples are shown in tables 3.3 and 3.4 and curves are shown in figures A.1–A.16. The large number of transcripts fitting category 2 in the unpollinated samples is the result of bacterial contamination in the June 22 unpollinated library. The large contigs produced during *de novo* assembly could be bacterial genomes. This contamination could affect read counting and normalization. Further analysis of this data could benefit from removal of these bacterial reads. The fifty most differentially expressed transcripts in each regression category were further investigated and transcripts with strong annotations in other organisms were chosen for discussion.

Table 3.3: Transcripts fitting each regression in pollinated samples.

Category	Number of Transcripts	Profile
1	412	Late up
2	159	Early down
3	58	Late down
4	79	Early up
5	732	Linear up
6	629	Linear down
7	30	Positive parabolic
8	4	Negative parabolic

Table 3.4: Transcripts fitting each regression in unpollinated samples.

Category	Number of Transcripts	Profile
1	363	Late up
2	1488	Early down
3	92	Late down
4	97	Early up
5	792	Linear up
6	937	Linear down
7	52	Positive Inparabolic
8	3	Negative parabolic

3.3.2 Comparison of Fertilized and Unfertilized Megagametophytes

In Douglas-fir, the development of the megagametophyte begins prior to fertilization and proceeds up to the formation of mature egg cells. In fertilized ovules, an embryo grows into the corrosion cavity and the megagametophyte begins accumulating significant amounts of storage proteins and lipids (Owens et al., 1993). In the absence of fertilization neither of these processes occur (von Aderkas et al., 2005a). Instead the megagametopyte degenerates two weeks after fertilization would have occurred (Rouault et al., 2004).

Effectors of PCD

Very little is known about the genetics of seed development in conifers or of ovular abortion. PCD is known to occur at other points in the development of viable conifer seed (Vuosku et al., 2009; He and Kermode, 2003a,b; Filonova et al., 2002), but these occurrences have not been characterized genetically. Megagametophyte abortion could involve transcriptomic differences at two levels: regulation and execution. There may be differential expression in genes regulating hormonal response, embryo

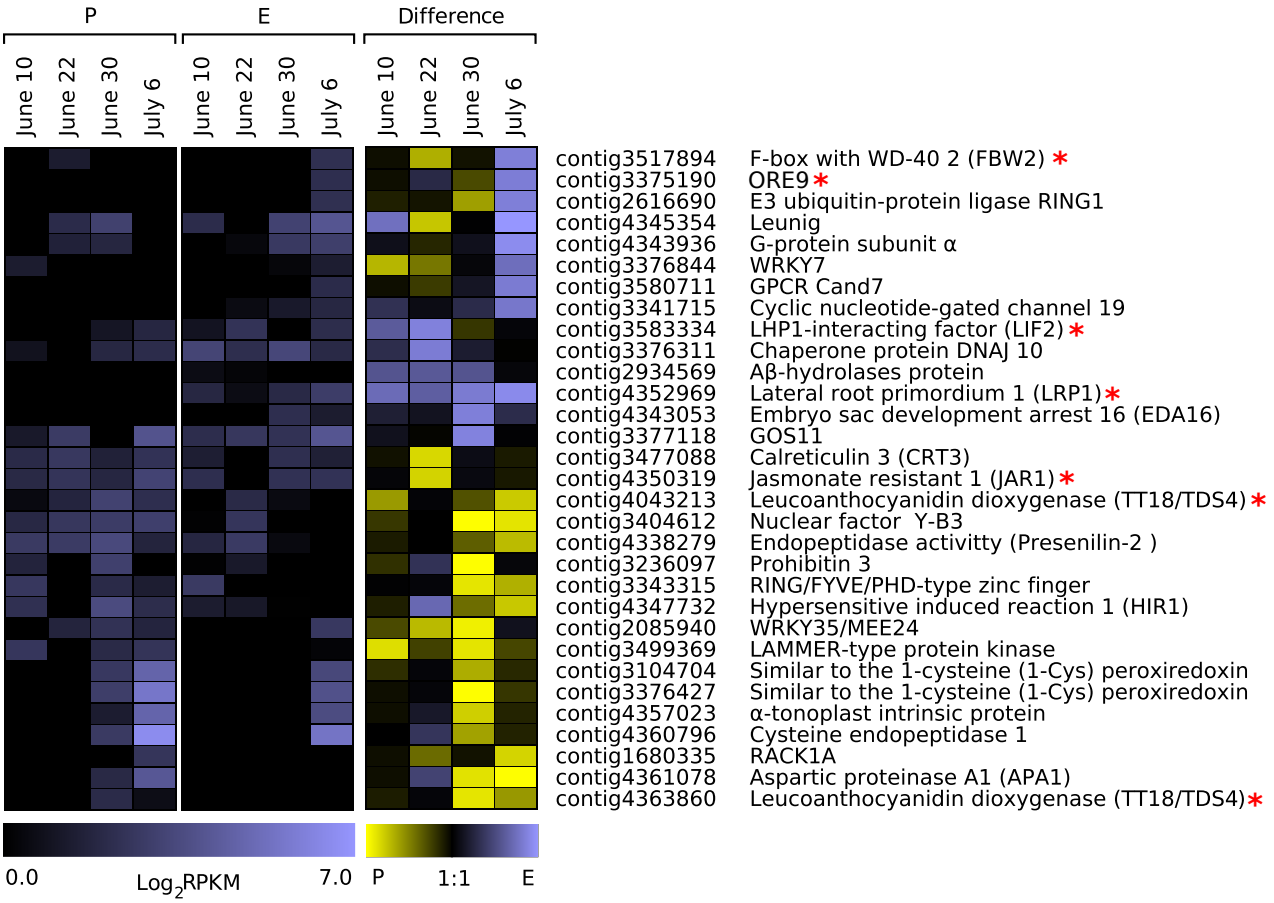


Figure 3.4: Transcripts differentially expressed between pollinated and unpollinated megagametophytes that are potentially involved in megagametophyte abortion or survival. Heat maps show expression independently in pollinated (P) and unpollinated (E) samples. Differences in expression between P and E are shown in the third heat map (maximum differece = 8). Candidate genes were selected based on differential expression between E and P at each harvesting date and BLAST hits against a database of PCD-related genes from *Mus musculus* and *Arabidopsis thaliana*. Discussed transcripts are marked with red asterisks.

development, and cell differentiation and division. The execution of a coordinated PCD program would be expected to rely on effectors such as proteases, endonucleases, membrane transporters, and ROS-related enzymes. Transcripts for these genes were not found among transcripts highly differentially expressed between pollinated and unpollinated megagametophytes during seed development.

Endonuclease and protease activity is present during the post-germinative PCD of *Picea glauca* megagametophytes (He and Kermode, 2003a,b). There are several reasons why these transcripts may not be evident in the expression data. Potential PCD genes were selected based partially on high differential expression between pollinated and unpollinated tissues. However, this may not be necessary to mount a proteolytic cascade. In animals, caspases are constantly and ubiquitously expressed as zymogens called procaspases (Fuchs and Steller, 2011). Apoptosis is initiated by proteolytic conversion of procaspases into active enzymes. Caspases can activate other caspases by proteolysis (Inoue et al., 2009). Caspase-3 is able to precipitate DNA fragmentation by proteolytically released inhibition of CAD endonuclease (Porter and Ja, 1999). The effectors required for PCD during megagametophyte degradation may be more effectively identified using proteomics rather than transcriptomics. Instead, higher level regulators of megagametophyte abortion may be a target for transcriptional regulation and could be discovered by a transcriptomic approach.

Potential Regulators

Several transcripts are differentially expressed leading up to abortion. Some are strong candidates as potential regulators of megagametophyte survival, while others do not obviously fit the biological events occuring during Douglas-fir seed development. To effectively combine my data with the current knowledge of Douglas-fir seed biology, I will focus on putative transcripts that have robust annotations in other

organisms and whose expression correlates with well-studied aspects of conifer seed development.

Jasmonic acid (JA) signalling may have a role in responding to fertilization and embryo formation. A Douglas-fir transcript similar to *Arabidopsis* jasmonate resistant 1 (JAR1) is basally expressed in the megagametophyte (Figure 3.4). Around the time of fertilization and proembryogeny, it becomes heavily repressed in the unpollinated megagametophyte, before resuming basal expression. JAR1 is an adenylating enzyme that is induced by JA. It precipitates JA responses by adenylating JA (Staswick et al., 2002), which allows JA to be further conjugated to isoleucine (Staswick and Tiryaki, 2004). Jasmonic acid and its conjugates have many effects, including regulation of defence responses, oxidative stress responses, and development (Staswick et al., 2002). JA-Ile may have a role in successful embryogenesis and seed maturation. The repression of its synthesis at the normal time of fertilization in unpollinated megagametophytes could be involved in inducing the abortive response to the absence of fertilization.

Two putative leucoanthocyanidin dioxygenase (LDOX) enzymes are expressed in the pollinated samples late in the experiment (Figure 3.4). These enzymes catalyze a major step in proanthocyanidin (PA) and anthocyanin synthesis (Abrahams et al., 2003). The seeds of LDOX-impaired *Arabidopsis* mutants are deficient in PAs that are normally found in specialized PA-containing cells in the seed coat (Abrahams et al., 2003). Seed coat PAs have antimicrobial functions and contribute to the impermeability of the seed coat (Debeaujon et al., 2003).

They also have regulatory roles in *Arabidopsis* seed development. In *Arabidopsis*, fertilization is a prerequisite for anthocyanin deposition in the seed coat (Debeaujon

et al., 2003). Accumulation of PAs in the seed coat, may induce ABA synthesis and dormancy. External application of PAs to imbibed seeds increases the expression of ABA biosynthetic genes and inhibits germination (Jia et al., 2012). Jia et al. (2012) found that several *Arabidopsis* mutants with mutations in PA-biosythetic genes had reduced ABA levels compared to wild-type seeds. The increased expression of putative Douglas-fir LDOX mRNAs begins in the mid to late stages of seed development and may later induce ABA synthesis. A failure to generate PAs in unfertilized Douglas-fir seed may prevent ABA accumulation and dormancy. Gutmann et al. (1996) previously found strong increases in PA deposition in hybrid larch *Larix* × *leptoeuropaea* somatic embryos deprived of ABA. They suggested that PA accumulation may be a stress response to ABA deprivation. Instead accumulation of PAs may favour ABA synthesis in response to low exogenous ABA.

ABA has a major role in seed dormancy and germination. Late in seed development, it can stimulate seed storage protein deposition (Chiwocha and von Aderkas, 2002) and dormancy and suppress germination (Finch-Savage and Leubner-Metzger, 2006). The parasitic wasp *Megastigmus spermotrophus* lays its eggs within Douglas-fir megagametophytes. It lays in both pollinated and unpollinated ovules and prevents the onset of megagametophyte abortion (Rouault et al., 2004). Chiwocha et al. (2006) reported that the presence of a wasp larva in the unpollinated megagametophyte elevates ABA levels to several times those in megagametophytes that are both unpollinated and not infested.

In aborting megagamgetophytes, a transcript is expressed with sequence similarity to FBW2 (Figure 3.4), an *Arabidopsis* regulator of ABA sensitivity. It has been named F-box with WD-40 2 (FBW2), despite the fact that it doesn't contain a WD-40 domain (Earley et al., 2010). The F-box domain suggests that FBW2 functions

by mediating ubiquitination. FBW2 destabilizes Argonaute 1 (AGO1), a protein involved in miRNA silencing that is required for normal ABA response (Earley et al., 2010). Earley et al. (2010) reported that *Arabidopsis* possessing a defective FBW2 were hyposensitive to increasing ABA concentrations, germinating much more quickly than wild-type plants. FBW2's ability to modify the ABA response may have a role in Douglas-fir megagametophyte abortion. ABA levels in pollinated megagametophytes are slightly higher than those in unpollinated megagametophytes, but the difference is minimal (Chiwocha et al., 2006). The physiological effects of ABA can be regulated by increasing or decreasing cellular sensitivity to ABA. While high sensitivity to ABA may be expected to increase seed storage deposition and induce dormancy, this does not occur in the unpollinated megagametophytes. Instead, ABA may be acting to promote PCD in unpollinated megagametophytes.

ABA can induce leaf abscission, suppress shoot branching, and accelerate leaf senescence (Kim et al., 2011; Stirnberg et al., 2007; Woo et al., 2001). In *Arabidopsis*, these processes are partially regulated by ABA-sensitive ORE genes (Kim et al., 2011). An transcript with sequence similarity to ORE9 is expresssed late in the unpollinated treatment (Figure 3.4). ORE9 is an F-box protein that interacts with the SCF ubiquitin conjugation complex (Woo et al., 2001). It is able to promote senescence by degrading promoters of longevity (Woo et al., 2001). More recently Stirnberg et al. (2007) demonstrated that ORE9 can suppress the branching of shoots by inhibiting growth in the axillary buds. ORE9 also has a role in barley seed development. PCD is integral to development of the barley endosperm, which is dead at maturity (Bethke et al., 1999). During PCD, ORE9 is expressed in the barley endosperm in concert with the 26S proteasome (Sreenivasulu et al., 2006).

A Douglas-fir transcript with sequence homology to *Arabidopsis* LATERAL ROOT PRIMORDIA 1 (LRP1) has sustained expression in the unpollinated samples over the course of the experiment, but it was not expressed in the pollinated (Figure 3.4). LRP1 is a SHI (short internode) family member (Kuusk et al., 2006) that was first associated with early growth in *Arabidopsis* lateral primordia (Smith and Fedoroff, 1995). It and other members of the SHI family members also function additively to promote gynoecium development in *Arabidopsis* (Kuusk et al., 2006). The SHI family in angiosperms remains mostly uncharacterized, but they appear to be involved in positively regulating development of different organs.

SHI family members have been found in *Selaginella möllendorfii* Hieron. and yellow cypress (*Chamaecyparis nootkatensis* D. Don.). In yellow cypress, LRP1 is associated with lateral root growth, similarly to *Arabidopsis* (Smith and Fedoroff, 1995). The role of LRP1 in *S.möllendorfii* is more complex. It is able to alternatively promote PCD or growth depending on the tissue type (Eklund et al., 2010).

3.3.3 Prefertilization and Early Embryogenesis

Corrosion Cavity Formation

Several transcripts are upregulated during corrosion cavity formation that may have a role in PCD (Figure 3.5), which is implicated in the formation of the corrosion cavity in conifers (Vuosku et al., 2009). During autolytic PCD, autophagic vacuolation of the cytoplasm appears to occur. It is responsible for large scale degradation of intracellular protein and organelles to generate amino acids (Bassham, 2007). A transcript with sequence similarity to *Arabidopsis* APG8 is upregulated around the time of corrosion cavity formation (Figure 3.5). A component of the *Arabidopsis* autophagy system, APG8, is essential to protein recycling. APG8 is a protein tag

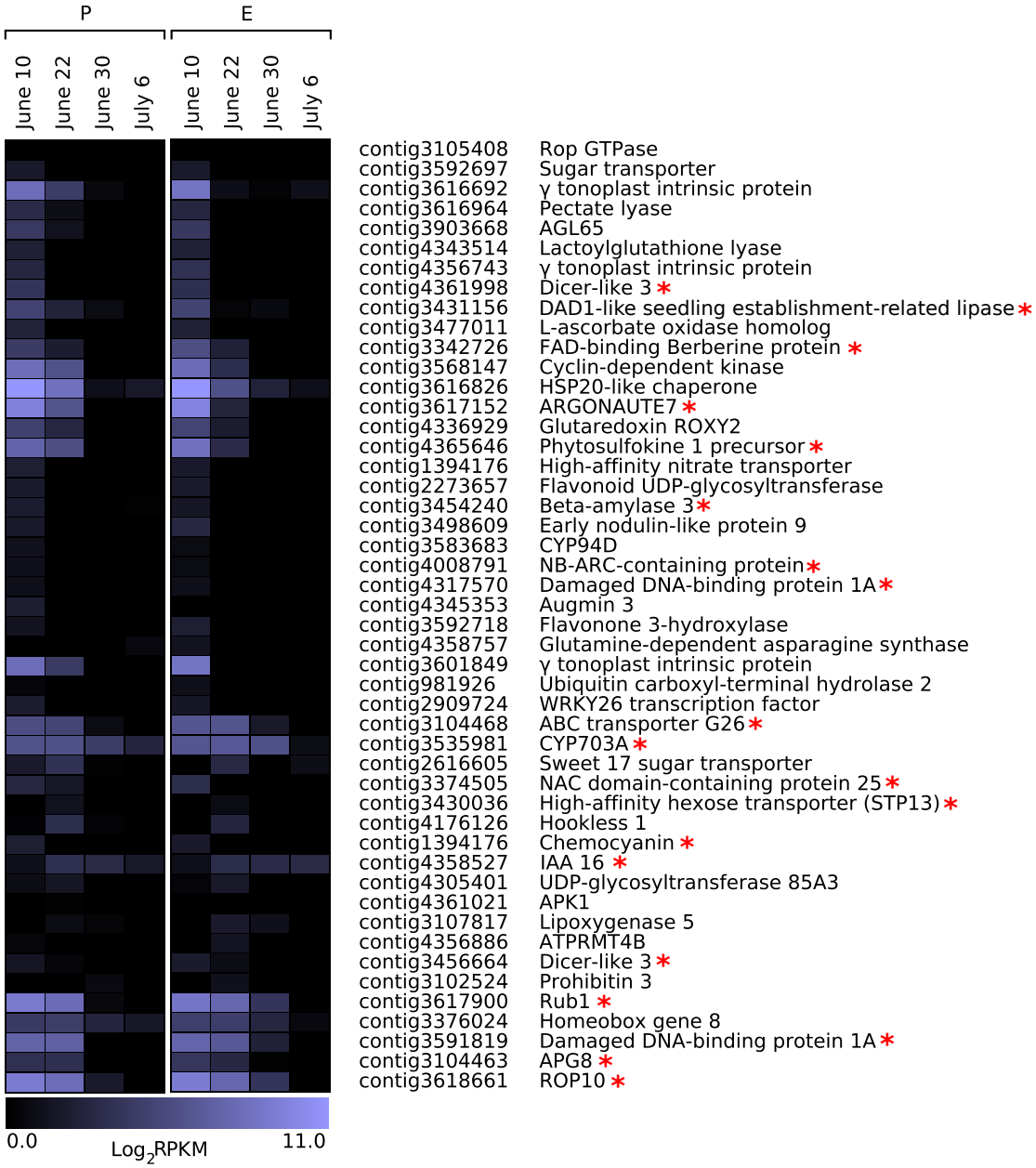


Figure 3.5: Transcripts potentially expressed during prefertilization and megagametophyte development. Heat maps show expression independently in pollinated (P) and unpollinated (E) samples. Quadratic regression models were used to find transcripts that most highly expressed early in experiment that would coincide with perfertilization and megagametophyte maturation. Transcripts in this heat map have similarity to *Arabidopsis* genes with robust annotations. Discussed transcripts are marked with red asterisks.

that is necessary for the formation of autophagosomes. Like ubiquitin, APG8 is first activated by an activating enzyme, APG7 (Doelling et al., 2002). Once activated, APG8 is able to bind phosphatidylethanolamine which leads to the formation of an autophagosome (Yoshimoto et al., 2004). During periods of high autophagy, APG8 accumulates strongly in the central vacuole as autophagosomes arrive at its surface (Yoshimoto et al., 2004). APG8 may serve to support PCD in the corrosion cavity and break down complex proteins into amino acids that accumulate in the corrosion cavity fluid.

Another transcript expressed during corrosion cavity formation bears similarity to an *Arabidopsis* high-affinity hexose transporter, STP13 (Figure 3.5). Norholm et al. (2006) correlated expression of STP13 with PCD. They found that STP13 is upregulated during senescence, in accelerated cell death mutants, and in response to fungal toxins or bacterial infection. As a transporter, this protein is believed to load sugars into sink tissues (Schofield et al., 2009). *Arabidopsis* plants overexpressing STP13 amass much higher levels of sucrose and amino acids under nutrient rich conditions than wild type plants (Schofield et al., 2009), implicating the activity of STP13 in regulating nitrogen uptake. The expression of this transcript in Douglas-fir correlates with the formation of the corrosion cavity and could have a dual role in regulating PCD and promoting sugar and amino acid accumulation.

The expression of a putative FAD-binding berberine protein, and a high-affinity nitrate transporter during corrosion cavity formation (Figure 3.5) suggests that amino acids may be actively transported to the corrosion cavity. The corrosion cavity fluid is filled with amino acids (Carman and Reese, 2005). The combined expression of STP13 and nitrogen transporters could help to enrich the fluid in the corrosion cavity. Searches for further nitrogen and sugar transporters coexpressed with the transcripts

described above would further support this idea.

Starch accumulates in the center of the megagametophyte prior to fertilization (von Aderkas et al., 2005a; Chiwocha and von Aderkas, 2002). The starch is evidently released into the corrosion cavity upon lysis of these cells (Carman and Reese, 2005). A putative β-amylase is highly-expressed in the megagametophyte when the corrosion cavity is forming (Figure 3.5) and is likely required for breaking down starch in the nascent corrosion cavity. Carman and Reese (2005) found that simple sugars such as maltose constitute a major portion of the sugars dissolved in Douglas-fir corrosion cavity fluid. Coordinated PCD of prothallial cells in the corrosion cavity region, breakdown of their lysate, and transport of nutrients to the corrosion cavity result in a nutrient rich fluid in which the embryo develops.

Developmental Signals

Many transcripts are upregulated during fertilization that have significant sequence similarity to *Arabidopsis* developmental regulators. A transcript with sequence similarity to an *Arabidopsis* phytosulfokine is expressed prior to fertilization, extending into early embryogenesis (Figure 3.5). Phytosulfokines are peptides that can induce cell division in many plants (Yang et al., 2001). Recently, they have been found to affect defence and stress responses in zinnia and *Arabidopsis*. Deleterious mutations in the phytosulfokine receptor in *Arabidopsis* reduce cell proliferation while increasing sensitivity to bacterial elicitors (Igarashi et al., 2012). Phytosulfokines downregulate components of the zinnia stress response during the early stages of transdifferentiation of mesophyll cells into tracheary elements (Motose et al., 2009). They can also be used to promote somatic embryogenesis in the conifers *Cryptomeria japonica* (L.f.) and *Larix leptolepsis* (Lamb.) (Igasaki et al., 2003; Umehara et al., 2005). A transcript expressed before and after fertilization bears sequence similarity to *Arabidopsis*

phytosulfokine α precursor. It may be involved in the rapid cell divison during embryogenesis, but could also weaken the defensive responses against the pollen tube and embryo.

A potential inhibitor of ABA-responsive signalling is expressed prior to fertilization (Figure 3.5). This transcript shares homology with Rho-related protein from plants 10 (ROP10), which negatively regulates the ABA-response in *Arabidopsis* (Zheng et al., 2002). In megagametophytes awaiting fertilization, ROP10 may negatively regulate the response to ABA and favour auxin-responsive signalling. ROP10 expression correlates with auxin and ABA concentrations in developing Douglas-fir megagametophytes. During seed development, auxin is present at high-levels prior to and during embryogenesis; ABA levels increase later, just before deposition of seed storage proteins and lipids (Chiwocha and von Aderkas, 2002). Auxin is involved in the development of the megagametophyte and the embryo (Chiwocha and von Aderkas, 2002). ABA is implicated in initiating and maintaining dormancy (Finch-Savage and Leubner-Metzger, 2006) and promoting the synthesis of seed storage proteins. ROP10 is able to weaken dormancy by inhibiting ABA signalling. The expression of ROP10 may help to maintain low levels of ABA during early megagametophyte development (Chiwocha and von Aderkas, 2002). Together, auxin responses and inhibition of ABA responses may prevent seed storage protein accumulation and dormancy.

The auxin response mechanism in plants relies on the degradation of transcriptional repressors of auxin-responsive genes called AUX/IAA genes. A putative auxin repressors, IAA16, is expressed in the megagametophyte during the experiment (Figure 3.5). Its expression correlates with increased endogenous auxin levels (Chiwocha and von Aderkas, 2002), which are high from prefertilization through early embryogenesis (Chiwocha and von Aderkas, 2002; Chiwocha et al., 2006). Auxin signalling

appears to be key to Douglas-fir embryogenesis and megagametophyte development. IAA proteins are continually produced and have very short half-lives *in situ*. This prevents the repressors from accumulating to a large level. Auxin generates physiological responses by inducing the degradation of these inhibitors by ubiquitination via an SCF-type E3 ligase complex (Kepinski and Leyser, 2005). When this occurs AUX/IAA genes become upregulated (Worley et al., 2000; Tiwari et al., 2001) making the cell ready to repress auxin-induced genes when the auxin signal drops. Prior to fertilization and into embryogenesis, the Douglas-fir megagametophyte expresses putative components of the auxin-responsive SCF complex and an IAA gene.

The expression of a putative RUB1 in Douglas-fir (Figure 3.5) also correlates with an increase in auxin levels in the megagametophyte just before fertilization (Chiwocha and von Aderkas, 2002). RUB1 is a ubiquitin-related protein that is required for auxin signalling (del Pozo, 1998). It is conjugated to CULLIN, a component of SCF E3-ubiquitin ligase complexes by two proteins. The conjugation is mediated by two proteins AXR1 and ECR1, which are specifically expressed in auxin-responsive cells (del Pozo and Dharmasiri, 2002). The DNA damaged-binding protein 1a (DDB1a) associates with SCF complex and has recently been shown to be RUB-modified, but the function of this modification is unknown (Hotton et al., 2012). DDB1a is required for embryo development in *Arabidopsis* (Bernhardt et al., 2010) and a putative homologue is coexpressed with RUB in Douglas-fir (Figure 3.5).

RNA-silencing appears to be important in prefertilization and embryogenesis stages of seed development. Three transcripts putatively related to RNA-silencing are expressed prior to fertilization and decrease over the course of development (Figure 3.5). Two of these are similar to *Arabidopsis* Dicer-like 3. Zhang et al. (2012) reported that DCL3-like enzymes were expressed in developing somatic embryos of *Larix lep-*

tolepsis somatic embryos. In the pro-embryogenic mass and in the early embryo, a DCL3 homologue is highly expressed, but its expression drops significantly as the embryo matures. DCL3 proteins cleave dsRNA precursor into interfering small RNAs. These associate with an RNA-induced silencing complex (RISC) and target RNAs for degradation. One component of the RISC is an Argonaute protein. ARGONAUTE7 follows the same expression patterns as the putative dicer-like transcripts. In *Arabidopsis*, it is important for normal development of mature leaves (Carbonell et al., 2012). In Douglas-fir, it is expressed in prefertilization megagametophyte or early embryo, as these tissues were harvested together. Consequently its role is not conserved in *Arabidopsis*.

In both the pollinated and unpollinated samples, genes are expressed prior to fertilization that are similar to genes associated with pollen development in *Arabidopsis* (Figure 3.5). They are highly expressed in pollinated as well as unpollinated Douglas-fir megagametophytes suggesting that they are involved in formation of the megaspore wall that surrounds the megagametophyte. Spores walls in land plants share a common origin and all have sporopollenin-like components in their walls. A putative ATP-binding cassette (ABC) transporter is expressed in Douglas-fir early in the experiment. This transcript shares strong sequence similarity with the *Arabidopsis* ABC transporter G26, which is required for pollen exine formation. It transports sporopollenin precursors to the exine during pollen development (Quilichini et al., 2010). A putative CYP703A follows the same expression pattern in Douglas-fir. In *Arabidopsis* this cytochrome P450 catalyzes the generation of sporopollenin precursors from lauric acid (Morant et al., 2007). These results suggest a high degree of sequence conservation between genes involved in the formation of the angiosperm exine and the conifer megaspore wall. Searching the Douglas-fir expression data with other *Arabidopsis* pollen wall synthetic genes could strenghthen this finding.

Fertilization

Fertilization in Douglas-fir occurs six weeks after pollination. About four or five weeks after pollination, the pollen grain begins to grow. The successful growth of the pollen tube towards the archegonium appears to be the result of a chemotactic process. A transcript with sequence similarity to lily chemocyanin is expressed in both pollinated and unpollinated megagametophytes (Figure 3.5). Its expression peaks at June 22, coinciding with pollen tube entry and fertilization. *Lilium longiflorum* (Thunb.) stigmata release chemocyanin to attract growing pollen tubes (Kim et al., 2003). Chemocyanin is specifically attractive to homogeneric pollen in lily and actually deters pollen tubes of tobacco (Kim et al., 2003). Douglas-fir pollen grains are target archegonia for penetration *in vitro*, suggesting a chemical signal is invovled. Unlike lily, pollen attraction is not strongly homogeneric in Douglas-fir. Douglas-fir megagametophytes attract and are readily penetrated by foreign pollen tubes from *Picea sitchensis* Bong. and *Pinus monticola* Douglas (Dumont-BéBoux et al., 1998).

3.3.4 Regulators of Embryo Developmcent

Many putative transcription factors are expressed after fertilization until the end of the experiment (Figure 3.6). These proteins may be involved in Douglas-fir embryogenesis and warrant further research. Three NAC domain-containing transcription factors are expressed during embryogensis. NAC domains interact with DNA and are found only in plants (Larsson et al., 2012). A transcript with sequence similarity to NAC domain-containing 2 (ATAF1) is expressed during embryogenesis in Douglas-fir. ATAF1 was one of the first characterized members of this protein family (Lu et al., 2007). Inhibition of ATAF1 expression in *Arabidopsis* causes severe developmental defects (Kleinow et al., 2009). It can suppress stress responses during drought (Lu et al., 2007), but is also required to mount some defence responses (Jensen et al.,

2008). This may be due to its influence on ABA-signalling. ATAF1 can inhibit the synthesis of ABA (Jensen et al., 2008), which bolsters resistance of *Arabidopsis* to the fungus *Blumeria gramnis* Speer.(Jensen et al., 2008). Jensen et al. (2008) also found that mutation of ATAF1 inhibited seed dormancy. By regulating ABA-synthesis, ATAF1 is able to regulate biotic and abiotic stress responses and ABA-dependent developmental processes (Ton et al., 2009). In Douglas-fir, this ATAF1 homologue may be essential to successful embryo development and entry into dormancy.

A transcript with similarity to *Arabidopsis* NAC domain-containing 98 is coexpressed along with the putative ATAF1 (Figure 3.6). This protein is also called CUP-SHAPED COTYLEDON (CUC) in *Arabdopsis* due to shoot apical meristem defects and cotyledon fusion in CUC mutants (Aida et al., 1999). PaNAC01, an orthologue of CUC in *Picea abies* is necessary for shoot apical meristem formation and separated cotyledons in somatic embryos (Larsson et al., 2012). It is expressed throughout embryogenesis and its expression is mediated by embryo polar auxin transport (Larsson et al., 2012). This protein likely plays a similar role in Douglas-fir. PaNAC01 is able to complement *Arabidopsis* CUC mutants, suggesting a high level of functional conservation in CUC orthologues.

One transcript expressed later in embryogenesis in Douglas-fir has sequence similarity to another regulator of *Arabidopsis* meristem identity (Figure 3.6). Late meristem identity 2 (LMI2/MYB17) *Arabidopsis* is a MYB transcription factor that helps to determine meristem identity in *Arabidopsis* and is also expressed in imbibed seeds (Zhang et al., 2009). Induction of LMI2 favours a strong meristem identity transition from vegetative to floral (Pastore et al., 2011), which leads to flower formation. This process occurs in mature *Arabidopsis* plants. The function of this gene may not be conserved between *Arabidopsis* and Douglas-fir. However, potential roles for LMI2 in

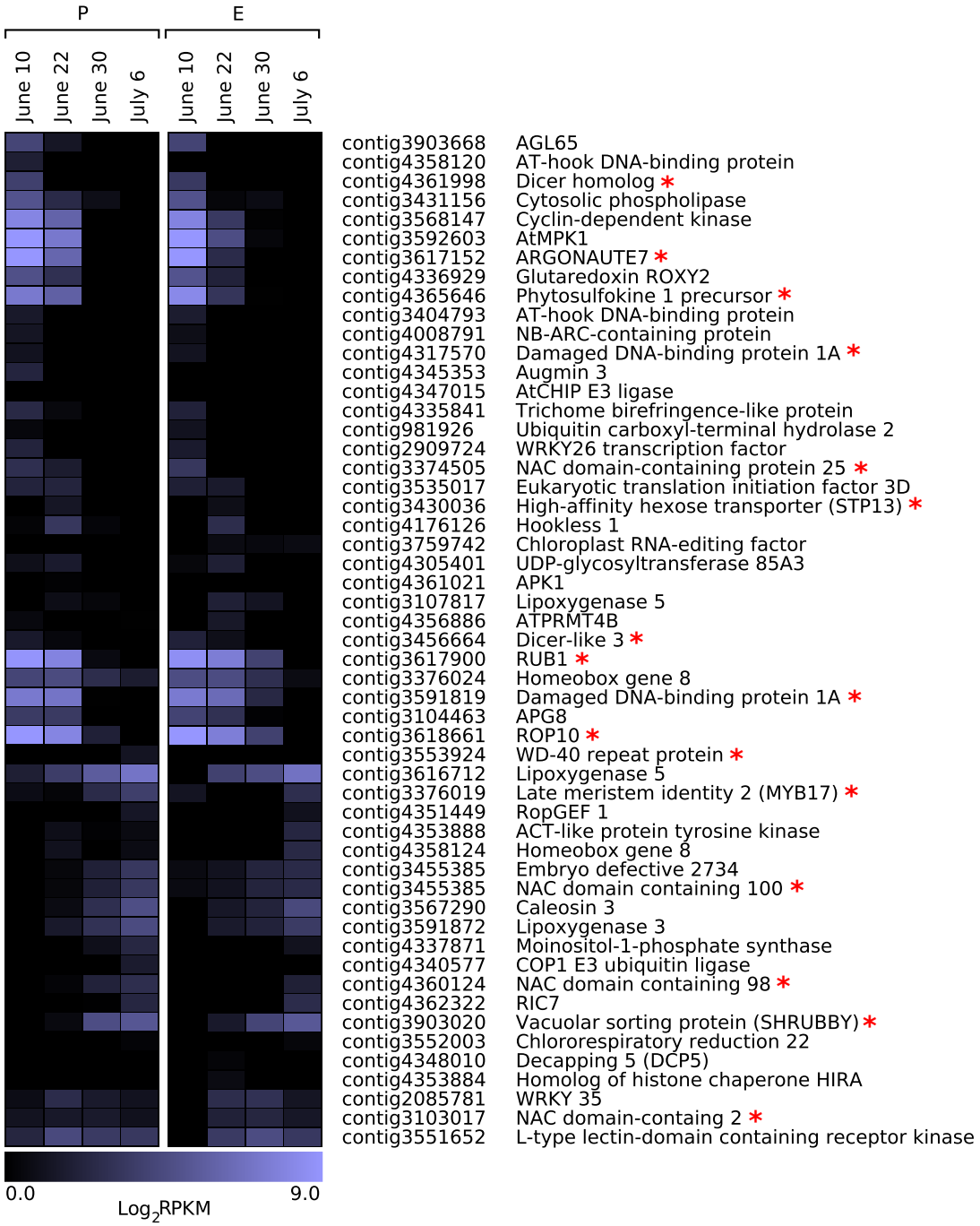


Figure 3.6: Transcripts potentially involved in embryo development. Heat maps show expression independently in pollinated (P) and unpollinated (E) samples. Quadratic regression models were used to find transcripts that were highly expressed after fertilization that may have roles in embryo development. Transcripts were selected from this pool based similarity to *Arabidopsis* genes with strong annotations potentially related to embryogenesis or other developmental processes. Discussed transcripts are marked with red asterisks.

seeds have not been studied. Pastore et al. (2011) suggested that LMI2 likely diverged from other MYB transcription factors early in flowering plants. LMI2's function as a promoter of reproductive meristem formation in *Arabidopsis* is unknown.

Root meristem development is also represented in the Douglas-fir transcripts expressed during embryo development. A transcript with sequence similarity to SHRUBBY is expressed during embryogenesis in Douglas-fir (Figure 3.6). SHRUBBY is a vacuolar sorting protein from *Arabidopsis* that is essential to normal root development and meristem function. Plants with defective SHRUBBY genes develop defective meristems that produce far fewer cells than in wild-type plants (Koizumi and Gallagher, 2013). The mutant plants also produce irregular radial patterning of the roots. It may be involved in root meristem development in the Douglas-fir embryo. This could be further explored by searching the expression data for signalling partners of SHRUBBY such as SCARECROW (SCR) and SHORT-ROOT (SHR) (Koizumi and Gallagher, 2013).

Many of the expressed transcripts that are expected to be involved in embryogenesis are expressed in both pollinated and unpollinated megagametophytes. This suggests that some of the functions ascribed to these genes in angiosperms may not be limited to embryogenesis in conifers, but may also be involved in female gametophyte formation. Another possiblity is that the tissues of the Douglas-fir embryo and megagametophyte share similar transcriptomes. Somatic embryos can be generated from megagametophyte tissue in a number of conifers (Pullman and Bucalo, 2011; Niskanen et al., 2004; von Aderkas et al., 1990). The demonstrated embryonic potential of conifer megagametophytes may be due to similarites in transcriptional programming between the two tissues.

3.3.5 Accumulation of Seed Reserves

Several storage proteins are have increased expression at the end of the experiment (Figure 3.7) that have shared sequence similarity with *Arabidopsis* seed storage proteins including glutelins, 12S and 2S storage proteins, and RmlC cupin-like family members. The RmlC-like cupins have significant sequence similarity to characterized angiosperm seed storage proteins in the Uniprot database including globulins and glutelins. Most of the storage proteins also have hits to seed proteins that have been identified in Douglas-fir and others conifers including species of *Picea* and *Pinus*.

Storage proteins are not evident in histological sections of unpollinated Douglas-fir megagametophytes (von Aderkas et al., 2005a). However, unpollinated megagametophytes transcribe seed storage protein mRNAs. These transcripts accumulate later than in pollinated megagametophytes suggesting that additional regulation of storage protein synthesis occurs after transcription. The post-transcriptional regulation of storage protein accumulation is not well-studied, but several studies have found discordance between transcript levels and protein accumulation. Hajduch et al. (2010) used a systems approach to compare transcription of seed protein mRNAs and synthesis of the actual protein. They found only 56% concurrence between mRNA and protein levels, indicating significant post-transcriptional regulation contributed to final protein levels. In tobacco, post-transcriptional regulation can strongly inhibit storage protein expression even in the presence of high mRNA levels. Schubert et al. (1994) transformed an oat globulin protein into tobacco. In oat, this protein is expressed specifically in the endosperm while another protein functions as a storage reserve in the embryo. Tobacco storage proteins mostly lack specificity to either structure. The oat globulin mRNA was highly expressed in both the embryo and the endosperm, but the protein itself was completely absent in the embryo. Post-

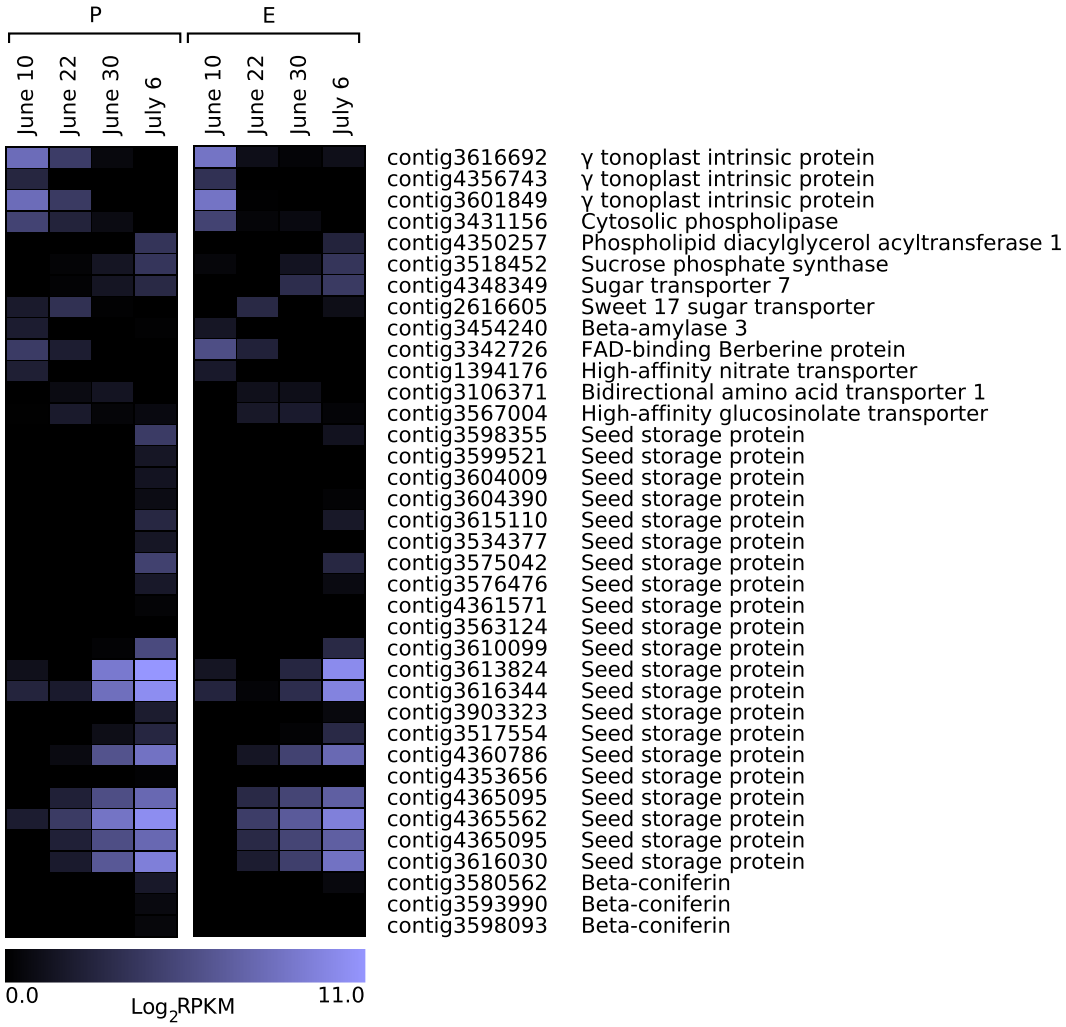


Figure 3.7: Transcripts potentially involved in seed storage. Heat maps show expression independently in pollinated (P) and unpollinated (E) samples. Quadratic regression models were used to find transcripts that followed various expression patterns. These data were used to find candidate seed storage proteins, transporters, and enzymes based on sequence similarity to putative *Arabidopsis* homologues.

transcriptional regulation appears to be entirely responsible for the tissue-specificity of this transgenic globulin.

One transcript has sequence similarity to *Arabidopsis* phospholipid:diacylglycerol acyltransferase (PDAT) (Figure 3.7). This enzyme is involved in seed lipid synthesis in a number of angiosperm species (Li et al., 2010b). Prior to lipid deposition, a putative cytosolic phospholipase is expressed. A second transcript is similar to *Arabidopsis* DAD1-like seedling establishment-related lipase (DSEL). DSEL prevents mobilization of lipid reserves during seedling growth (Kim et al., 2011). Plants overexpressing this gene have high populations of peroxisomes and oil bodies in their seed storage tissues (Kim et al., 2011). This may serve to prevent lipid catabolism in the developing megagametophyte since these reserves are required for post-germinative growth of the seedling.

Transcripts potentially involved in sugar transport and starch synthesis are expressed after fertilization (Figure 3.7). They share sequence similarity to *Arabidopsis* sucrose phosphate synthase 3F (SPS3F) and sugar transporters SWEET17 and sugar transporter 7 (STP7). Their expression coincides with starch accumulation in the embryo and in the chalazal end of the megagametophyte in Douglas-fir (Chiwocha and von Aderkas, 2002). SPS3F is not highly expressed in *Arabidopsis* leaves, where other isoforms synthesize starch from photosynthate (Sun et al., 2011). In both *Arabidopsis* and Douglas-fir SPS3F may be involved in starch synthesis in non-leaf sink tissues such as the megagametophyte. Sugar transporter 7 (STP7) is not characterized but is closely related to other monosaccharide-proton transporters (Yamada et al., 2011). SWEET17 is a vacuolar sugar transporter that regulates leaf fructose levels in leaves and has no influence on starch and sucrose levels (Chardon et al., 2013). Other closely related members of the SWEET family transport glucose and sucrose. Together these

proteins may convert incoming sucrose into the glucose necessary for starch synthesis.

3.3.6 Preparation for Dormancy

A seed entering dormancy must become tolerant of dessication, a condition intrinsic to seed maturity. The seeds of many species must additionally be resistant to high and low temperatures. This tolerance is believed to be mediated by proteins produced prior to dormancy including late embryogenesis abundant (LEA) proteins, heat shock proteins (HSP), and thaumatin-like osmotins.

A putative LEA protein is expressed late in the experiment in both the pollinated and unpollinated samples (Figure 3.8). LEA proteins were initially correlated with late embryogenesis and the beginning of dormancy. They have been found in the seeds of both gymnosperms (Leal and Misra, 1993) and angiosperms and have also been found in vegetative tissues (Tunnacliffe and Wise, 2007). LEA proteins are versatile. They are able to prevent protein aggregation and damage under dessication and freezing conditions (Goyal et al., 2005). The expression of LEA protein in both pollinated and unpollinated megagametophytes suggests that the transcription of LEA protein mRNAs may not inhibited in the absence of fertilization.

Increased heat shock protein expression is also associated with late seed development and dessication tolerance (Wehmeyer and Vierling, 2000; Hökstra et al., 2001). Small heat shock proteins are particularly associated with seed development. They are often expressed in the mid to late stages of seed development and their synthesis is induced by ABA (Wehmeyer and Vierling, 2000). Like LEA proteins, small HSPs prevent protein aggregation (Hökstra et al., 2001). They may also be involved in stabilizing membranes (Hökstra et al., 2001). Several heat shock proteins are highly expressed during Douglas-fir seed development (Figure 3.8), however most of them

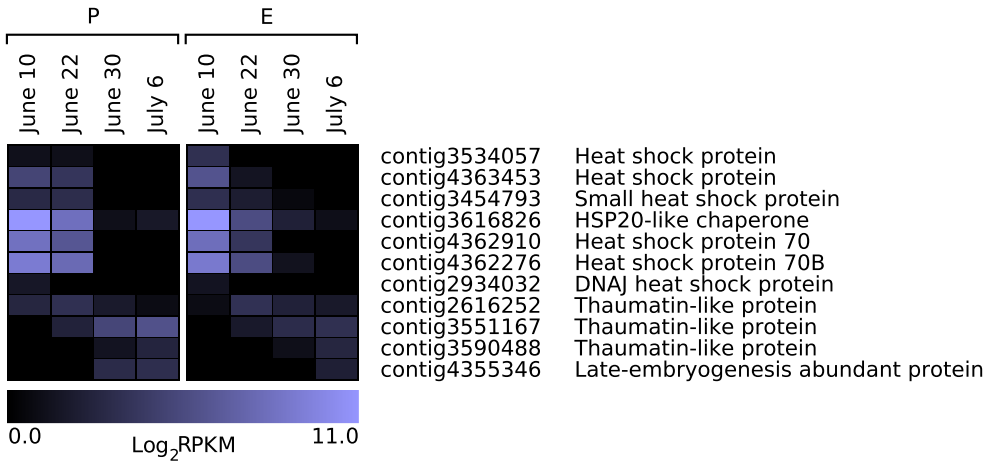


Figure 3.8: Transcripts potentially involved in seed stress tolerance. Heat maps show expression independently in pollinated (P) and unpollinated (E) samples. Candidate genes were selected based on changes in expression over the course of the experiment and sequence similarity to known heat, dessication, and pathogen tolerance genes expressed during seed maturation.

have high expression levels before LEA protein and seed storage protein expression. Only one has significant sequence similarity to a small heat shock protein. Heat shock protein expression may be better characterized in Douglas-fir by targetted querying of the expression data with sHSP sequences from *Arabidopsis*.

Several potential thaumatin-like proteins (TLP) transcripts are upregulated just prior to dormancy (Figure 3.8). TLPs are typically associated with defence responses, but may also be expressed as a consitutive defense against fungal infestation. TLPs are strongly expressed in Douglas-fir roots in response to infestation by the fungus *Phellinus sulphurascens* (Sturrock et al., 2007). Their ability to inhibit fungal growth is thought to result from their disruption the membranes of invading cells (Sturrock et al., 2007). TLPs likely act a constitutive defense against fungal infection in the pollination drops of gymnosperms. TLPs have been identified in the pollination drops of several conifer species including hybrid yew (*Taxus* × *media* Rehder), *Juniperus communis*, *Juniperus oxycedrus*, and *Chamaecyparis lawsoniana* (O'Leary et al., 2007; Wagner et al., 2007). Chitinases have also been detected in conifer pollination drops (Wagner et al., 2007), strengthening their role as a fungal defense. Dormant seeds are able to resist fungal attack for long periods of time on the forest floor. TLPs may serve as a defense against fungal infestation in mature and dormant Douglas-fir seed.

3.3.7 Vegetative and Reproductive Tissues

Vegetative

Transcripts that were highly expressed in the vegetative tissues relative to the megagametophyte tissues fit several functional categories (Figure 3.9). Several of these transcripts had BLAST hits to *Arabidopsis* photosynthesis components. These included photosystem subunits, light-harvesting complexes, an oxygen-evolving com-

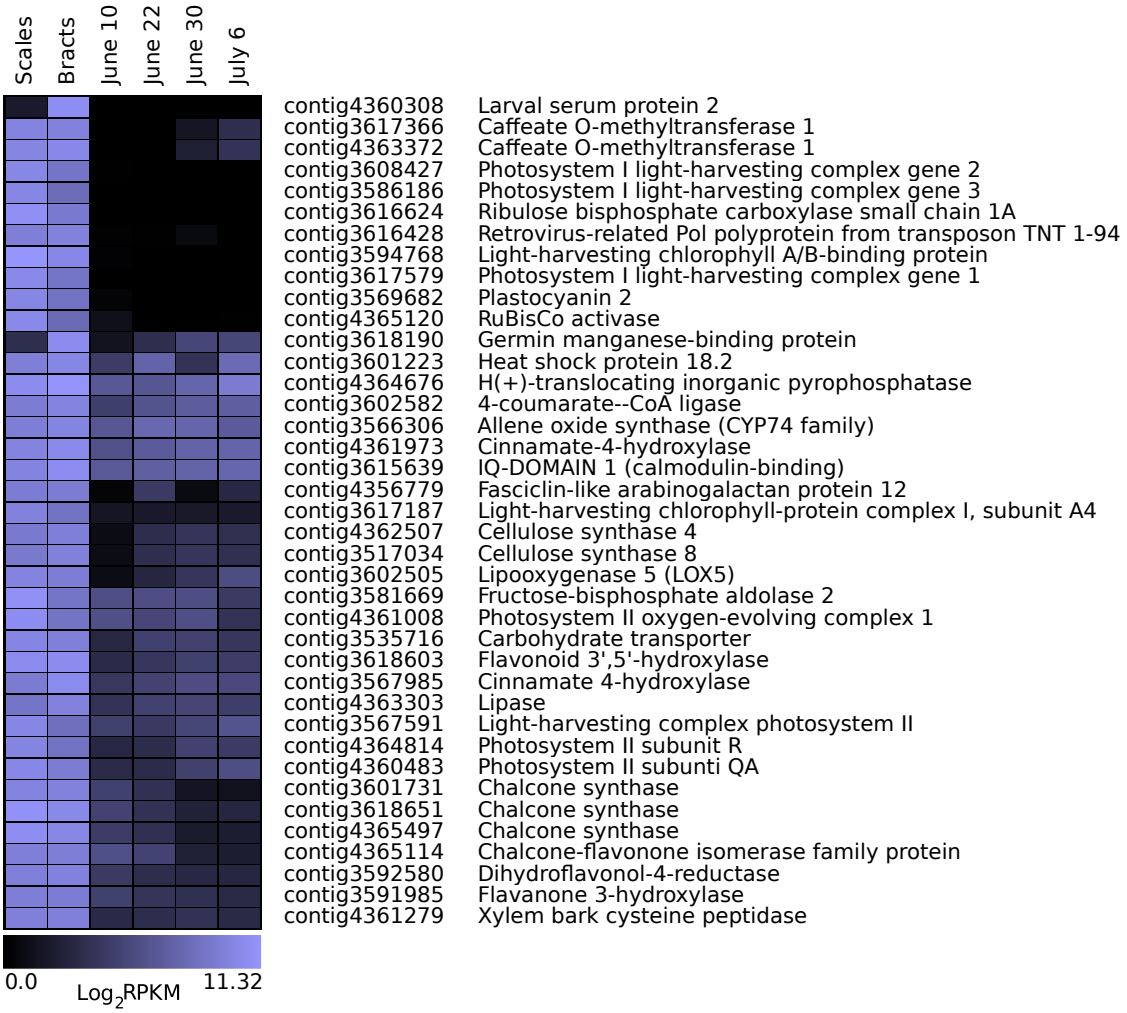


Figure 3.9: Transcripts with high differential expression in bracts and scales versus megagametophytes. Log_2RPKM is averaged for each megagametophyte collection date. The top fifty most differentially expressed transcripts were filtered based on the presence of an annotation in at least one BLAST database.

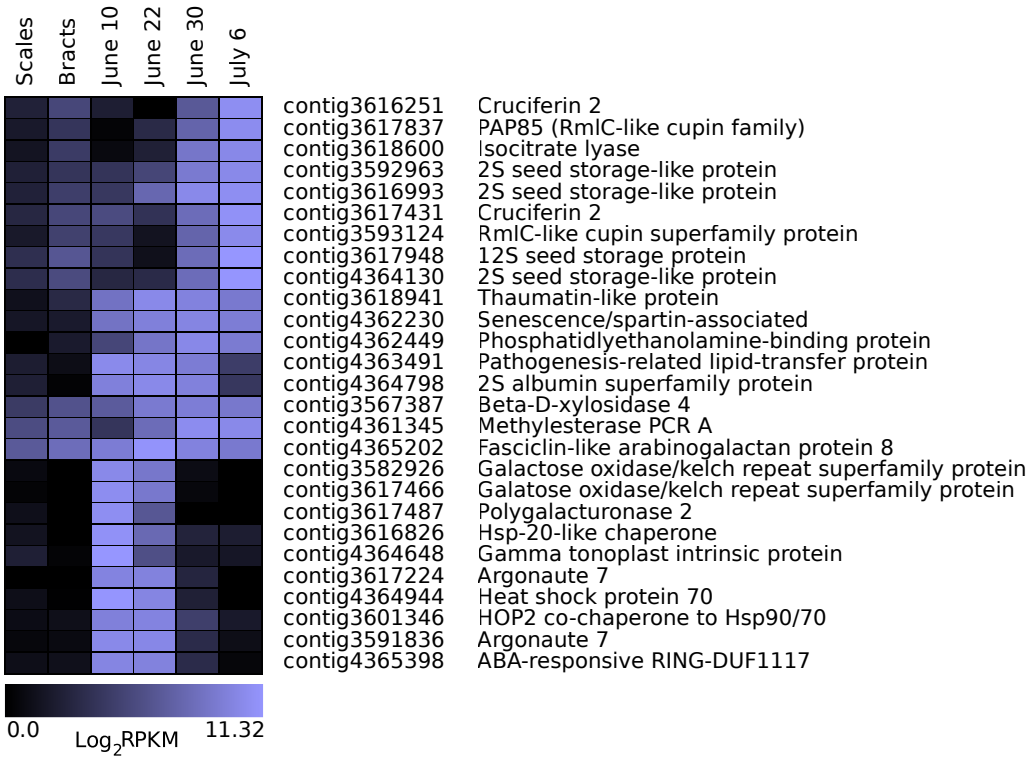


Figure 3.10: Transcripts with high differential expression in megagametophytes versus bracts and scales. Log_2RPKM is averaged for each megagametophyte collection date. The top fifty most differentially expressed transcripts were filtered based on the presence of an annotation in at least one BLAST database.

plex, plastocyanin, and a chloropyll-binding protein. Together, these components form a large part of the light reactions of photosynthesis. The Calvin cycle is centred around the enzyme ribulose-1,5-bisphosphate carboxylase oxygenase (RuBisCo). Its small chain 1A is specifically expressed in the vegetative tissues. Other components of the Calvin cycle such as RuBisCo activase, a catalytic chaperone required for the efficient function of RuBisCo (Portis, 2003) are also expressed highly in vegetative tissues.

Other transcripts that were specifically expressed in vegetative tissues were primarily involved in flavonoid biosynthesis. The first committed step in flavonol synthesis is performed by chalcone synthase (CHS) (Winkel-Shirley, 2001). Three transcripts with sequence similarity to *Arabidopsis* CHS are strongly differentially expressed in vegetative tissues. Numerous enzymes catalyze the subsequent steps and some of these are highly differentially expressed in the vegetative tissue: flavonoid 3′,5′-hydroxylase, chalcone-flavonone isomerase, dihydroflavonol 4-reductase, and flavonone 3-hydroxylase.

Megagametophyte

Some transcripts were specifically expressed in megagametophyte tissues relative to vegetative samples (Figure 3.10). Many of these transcripts demonstrated further specificity to developmental periods. Seed storage proteins, including possible RmlC-like cupins, cruciferins, and 2S and 12S storage proteins are expressed in later megagametophyte samples (Figure 3.10). Putative seed storage transcripts are specifically expressed in both pollinated and unpollinated megagametophytes. Based on histochemical analysis no seed storage proteins are present Douglas-fir megagametophytes destined for abortion (von Aderkas et al., 2005a). This suggests that other mechanisms may regulate the accumulation of storage protein bodies in megagametophytes. Such mechanisms could include regulation of protein synthesis.

Several transcripts potentially involved in cell wall modification are specifically expressed in the megagametophyte including a methylesterase, β-xylosidase, polygalaturonase, and fascilin-like arabinogalactan. These transcripts could be involved in megagametophyte-specific cell wall degradative processes in the formation of the corrosion cavity and softening of the nucellus. They could also be involved in cell wall synthetic processes in embryogeny or megaspore wall thickening.

Transcripts related to photosynthesis flavanoid biosynthesis are not highly expressed in megagametophyte samples compared to vegetative samples. The megagametophyte is encased within the ovule and is not exposed to light. Therefore, neither UV protectant nor photosynthetic genes would not be expected to be highly expressed in this tissue.

3.3.8 Conclusions

Mass expression of PCD effectors does not occur in aborting Douglas-fir megagametophytes. Initation of PCD during abortion of the megagametophyte may instead occur by a proteolytic cascade such as that which occurs during apoptosis. Possible mechanisms for determining the fate of the megagametophyte towards PCD or dormancy are represented in the expression data. The concerted expression of transcripts that regulate ABA-sensitivity or are induced by ABA suggests a key role for this hormone in regulating megagametophyte fate. This concept could be further elaborated by searching the transcriptome data for transcripts that are modifiers of the cellular response to ABA, are potentially regulated by ABA, or are involved in its biosynthesis.

Several potential transcriptional events that could effect megagametophyte fate are evident in the data. The accumulation of anthocyanins in the megagametophyte and

embryo has been previously documented. Two LDOX transcripts were found that are likely involved in this process. PA accumulation may stimulate ABA synthesis, thus inducing seed storage deposition and dormancy. This does not appear to occur in unfertilized megagametophytes and could be confirmed by UV high pressure liquid chromatography. Downstream effectors of ABA as well as modulators of ABA responses are also differentially expressed between pollinated and unpollinated megagametophytes. ABA appears to have a role in regulating megagametophyte fate. Death and survival are likely regulated by a combination of ABA levels and modulation of the ABA response.

Potential embryogenic regulators are consistently expressed in both pollinated and unpollinated samples, suggesting that the transcriptomes of developing megagametophytes and embryos are not disimilar. This may explain the potential of haploid megagametophytes to form somatic embryos *in vitro*. The developmental roles of *Arabidopsis* and Douglas-fir embryogenic regulators appear to be conserved. Further parallels could likely be drawn between conifer and angiosperm seed development using *Arabidopsis* seed development regulators to query the Douglas-fir transcriptome.

The formation of the *Arabidopsis* pollen exine and the Douglas-fir megaspore wall both involve deposition of sporopollenin. Two transcripts likely involved in the synthesis and export of sporopollenin were found in both pollinated and unpollinated Douglas-fir megagametophytes. They share strong sequence similarity to proteins involved in the processes in *Arabidopsis*. This is the first genetic evidence for a common origin for angiosperm microspore walls and gymnosperm megaspore walls.

Vegetative tissues were compared to reproductive tissues to find transcripts that are strongly expressed in only one of these tissue types. Bracts and scales are photo-

synthetic tissues, whereas megagametophytes are non-photosynthetic. Components of photosynthesis such as those in the photosystems and the Calvin cycle would be expected to be expressed at high levels in bracts and scales and at very low levels in megagametophytes. The megagametophyte functions as a nutrient provision and storage tissue throughout embryo development and seed germination. During its development seed storage proteins are produced that constitute a significant portion of the weight of a mature seed (Owens et al., 1993). Strong differential expression of photosynthesis and seed-storage transcripts indicates that the RNA-Seq data can provide an realistic measure of expression levels in the sample tissues.

This work can be extended further by experimentally supporting the ideas put forward in this thesis. Realtime polymerase chain reaction (RT-PCR) can be used to confirm differential expression in other Douglas-fir genetic backgrounds. Further RNASeq could also be used to confirm these results. The roles of putative transcripts identified in Douglas-fir could be supported by complementing *Arabidopsis* mutants with the sequence of the Douglas-fir homologue. Successful complementation would indicate a high degree of functional conservation. The role of PA accumulation in Douglas-fir seed development could be further studied by stain pollinated and unpollinated ovules for PAs and also by studying gross PA levels in viable and aborting seeds using high performance liquid chromatography (HPLC).

Appendix

Table A.1: Multi k-mer assembly results produced by ABySS using incremental odd values of k from 21 to 81. Values were collected with abyss-fac. They include the total number of contigs, the number of contigs exceeding 200bp and the N50 length. The N50 and maximum contig lengths are also shown. Contig counts are in thousands.

k	Contig Count	$>$ 200 bp	$>$ N50	N50 Length (bp)	Max Length (bp)
21	16380	123	35	369	25491
23	14140	149	39	426	82889
25	12700	157	40	445	146547
27	11580	163	41	455	155287
29	10580	167	41	469	155287
31	9688	170	41	483	162433
33	8923	171	41	496	162433
35	8234	173	40	509	162433
37	7605	174	39	525	140100
39	7008	175	38	540	162566
41	6446	175	37	560	162418
43	5883	174	36	581	162567
45	5280	173	35	605	160244
47	4347	169	33	637	269991
49	3902	167	31	666	269794
51	3537	164	30	695	269796
53	3144	161	29	721	269798
55	1436	124	20	923	106637
57	1283	121	20	959	106236
59	1137	118	19	1001	106238
61	1007	115	18	1044	106240
63	888	111	17	1086	106242
65	778	107	17	1142	106244
67	675	102	16	1196	106733
69	587	98	16	1242	104943
71	505	93	16	1292	69508
73	431	89	15	1322	36310
75	353	84	15	1340	30180
77	252	75	14	1382	24334
79	213	69	13	1426	11538
81	174	61	12	1496	11431

Table A.2: The per-database number of contigs with a BLASTx hit against the NCBI non-redundant protein database or Uniprot or a BLASTn hit against a Phytozome genome reference. The maximum e-value cutoff was 1×10^{-5}.

Database	Hits
NCBI Non-redundant protein	319,696
UniProt protein	176,900
Aquilegia caerulea	157,618
Arabidopsis lyrata	154,989
Arabidopsis thaliana (Tair)	154,545
Brachypodium distachyon	170,233
Brassica rapas	165,941
Citrus clementina	162,134
Carica papaya	168,766
Chlamydomonas reinhardtii	86,879
Capsella rubella	157,395
Cucumis sativus	178,396
Citrus sinensis	155,255
Eucalyptus grandis	156,780
Glycine max	180,162
Gossypium raimondii	157,595
Linum usitatissimum	157,611
Malus domestica	163,417
Manihot esculentas	163,703
Mimulus guttatus	159,114
Medicago truncatula	161,386
Oryza sativa	171,079
Physcomitrella patens	146,514
Prunus persica	172,254
Populus trichocarpa	167,073
Panicum virgatum	152,787
Phaseolus vulgaris	160,458
Ricinus communis	167,365
Sorghum bicolor	159,191
Setaria italica	151,900
Selaginella moellendorffii	143,191
Thalophilas	158,448
Volvox carteri	80,391
Vitis vinifera	173,420
Zea mays	149,326

Table A.3: Bowtie paired-end alignment rates for untrimmed reads to a set of sequenced derived from ABySS contigs that had either a BLAST hit or a potential ORF greater than 200 bp in length.

Library	Read Processed	Alignment Rate
Bracts	45.7 million	64.53%
Scales	75.2 million	65.03%
June 10 E	42.0 million	53.13%
June 10 P	30.8 million	51.57%
June 22 E	42.0 million	45.53%
June 22 P	47.8 million	48.94%
June 30 E	68.4 million	56.31%
June 30 P	40.8 million	62.23%
July 6 E	36.9 million	52.14%
July 6 P	44.3 million	54.36%

Table A.4: The number of significantly differentially expressed contigs in each pairwise comparison. The analysis was conducted using NOISeq with a minimum Q cut-off of 0.9

	June 10 E	June 10 P	June 22 E	June 22 P	June 30 E	June 30 P	July 6 E	July 6 P
June 10 E								
June 10 P	309							
June 22 E	1628	996						
June 22 P	485	275	112					
June 30 E	3460	2931	629	1045				
June 30 P	4179	3629	1740	2134	188			
July 6 E	5069	4537	2385	2737	480	233		
July 6 P	5187	4553	2673	3007	846	248	177	

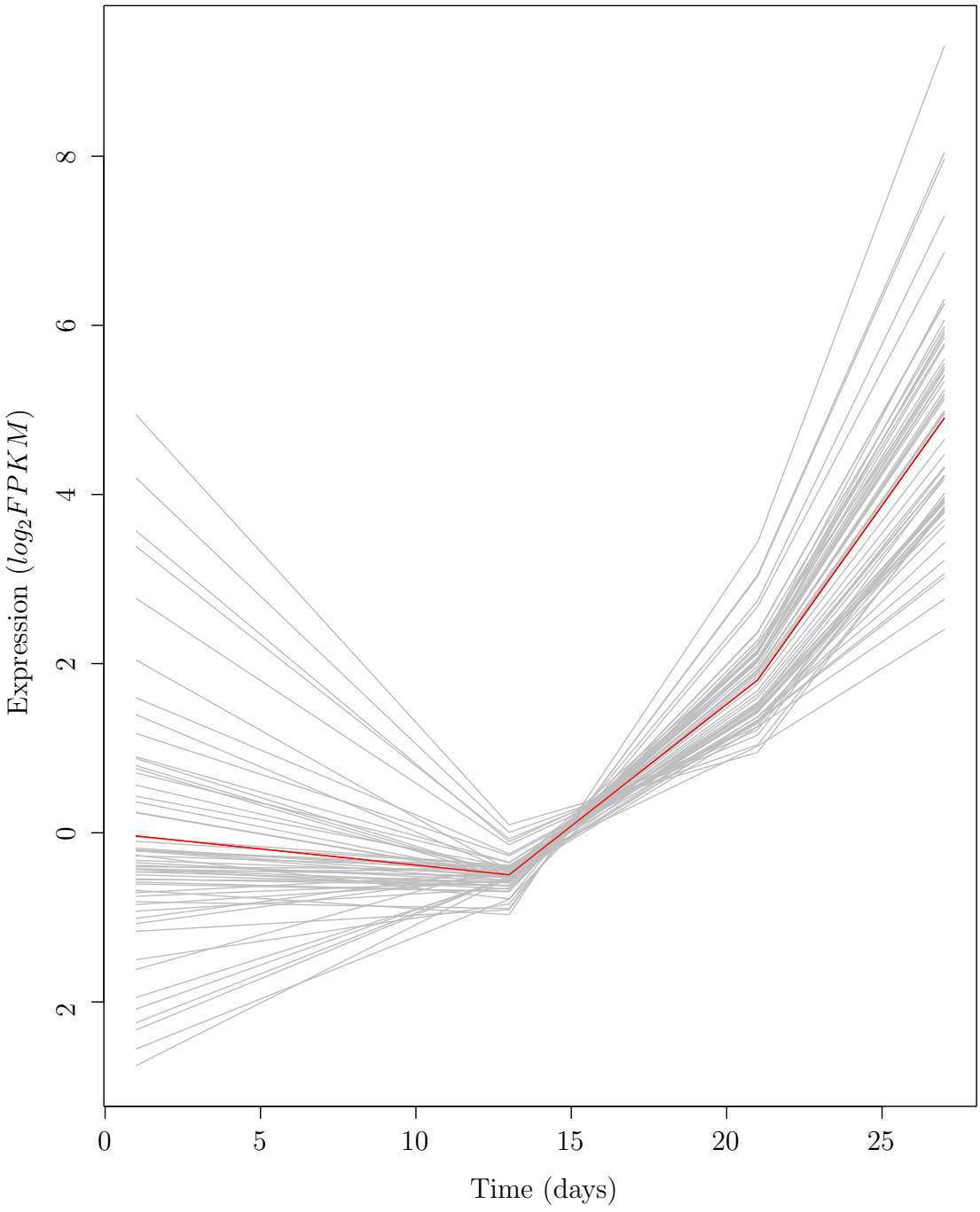


Figure A.1: Transcripts in pollinated megagametophytes that have late increases in expression. They fit quadratic regressions where β_2 and β_1 are positive (Category 1).

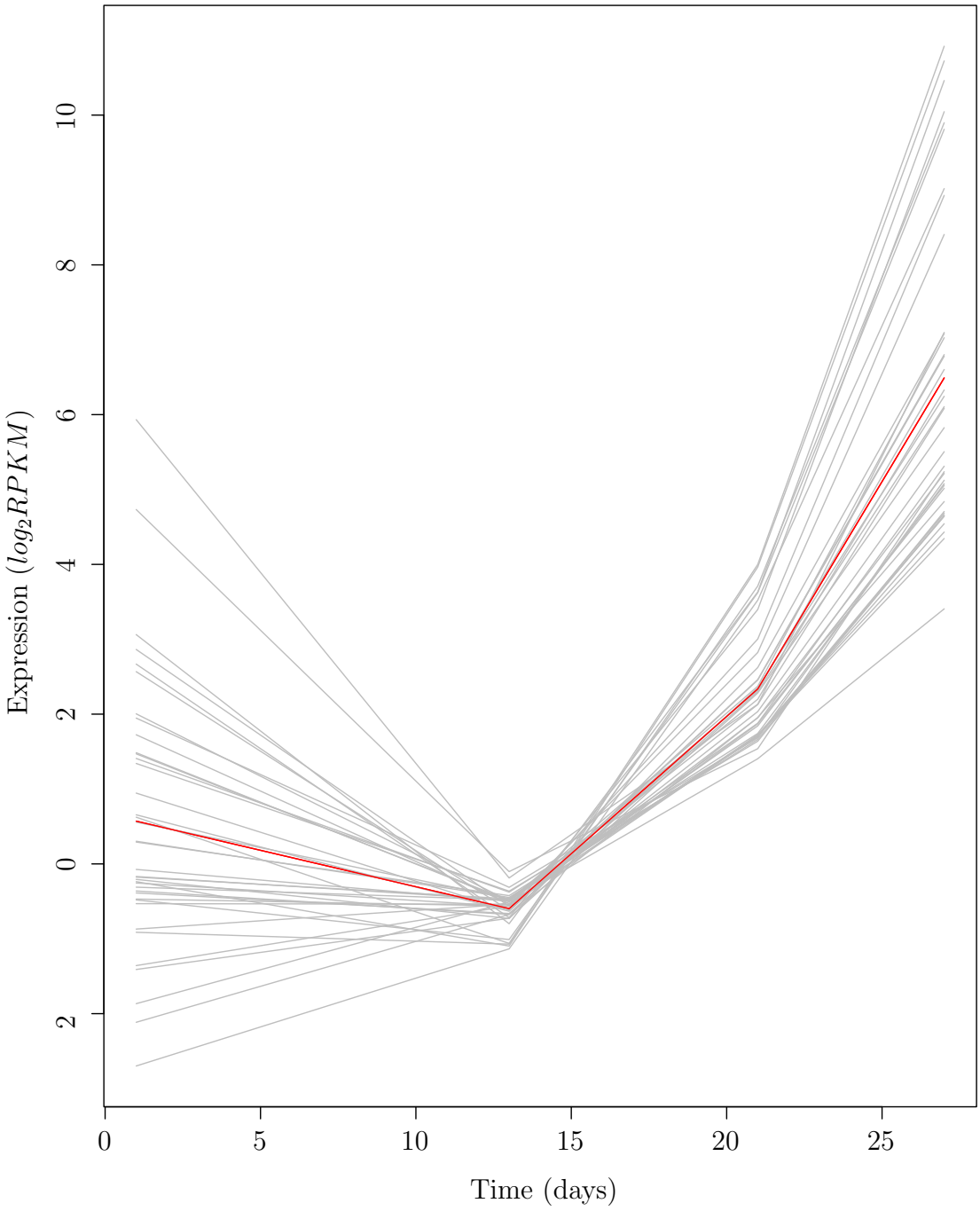


Figure A.2: Transcripts in unpollinated megagametophytes that have late increases in expression. They fit quadratic regressions where β_2 and β_1 are negative (Category 1).

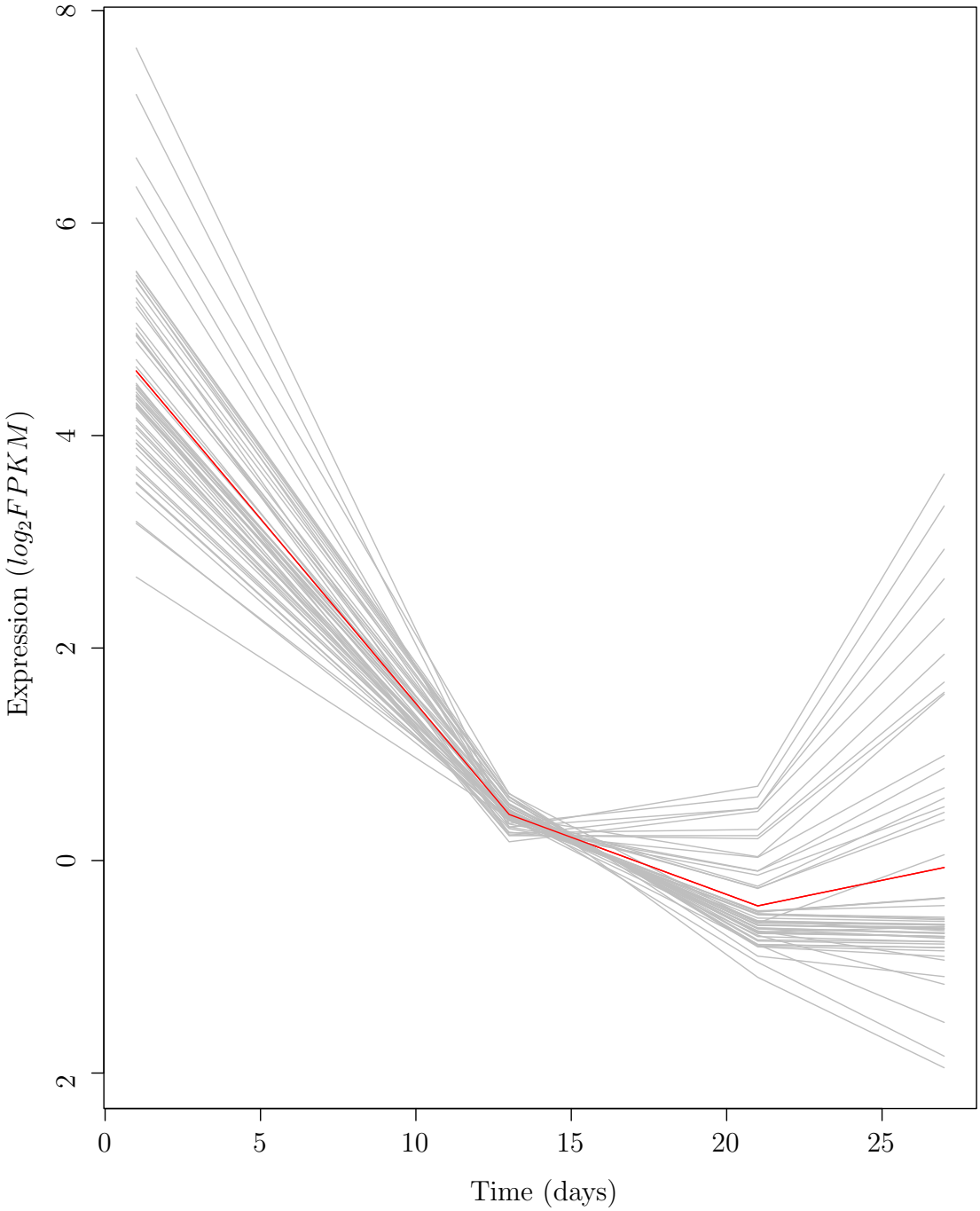


Figure A.3: Transcripts in pollinated megagametophytes that have early decreases in expression. They fit quadratic regressions where β_2 is positive and β_1 is negative (Category 2).

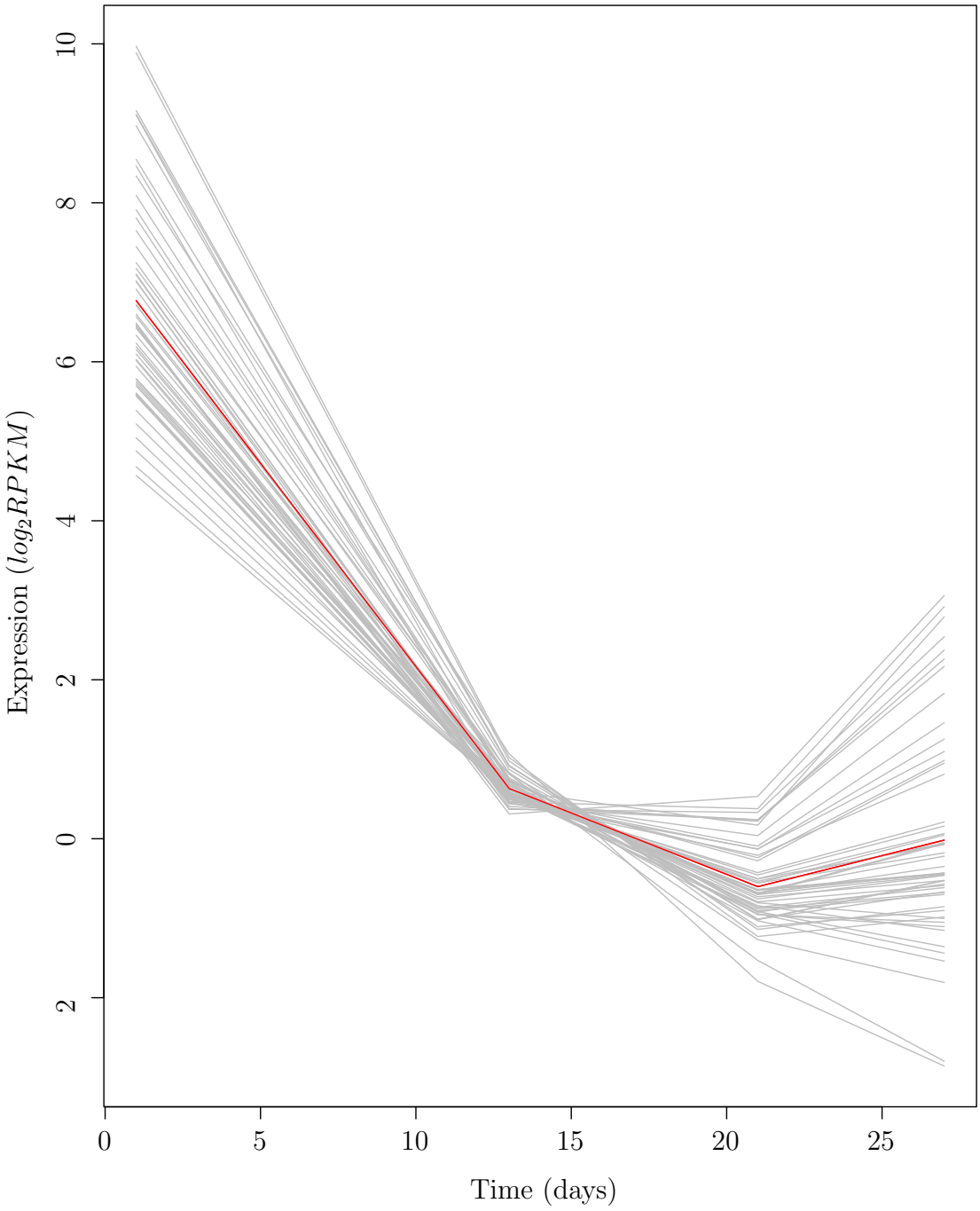


Figure A.4: Transcripts in unpollinated megagametophytes that have early decreases in expression. They fit quadratic regressions where β_2 is positive and β_1 is negative (Category 2).

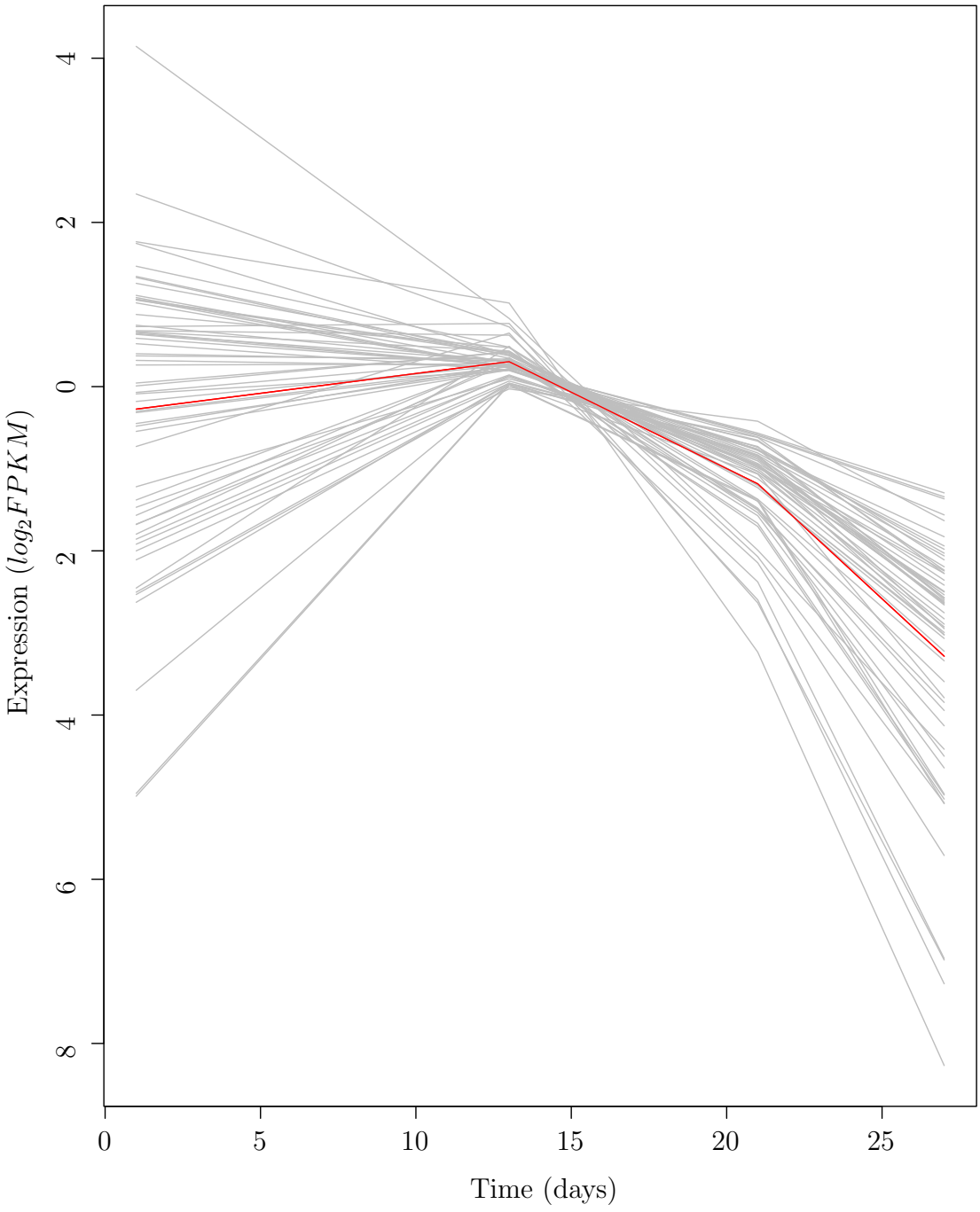


Figure A.5: Transcripts in pollinated megagametophytes that have late decreases in expression. They fit quadratic regressions where β_2 and β_1 are negative (Category 3).

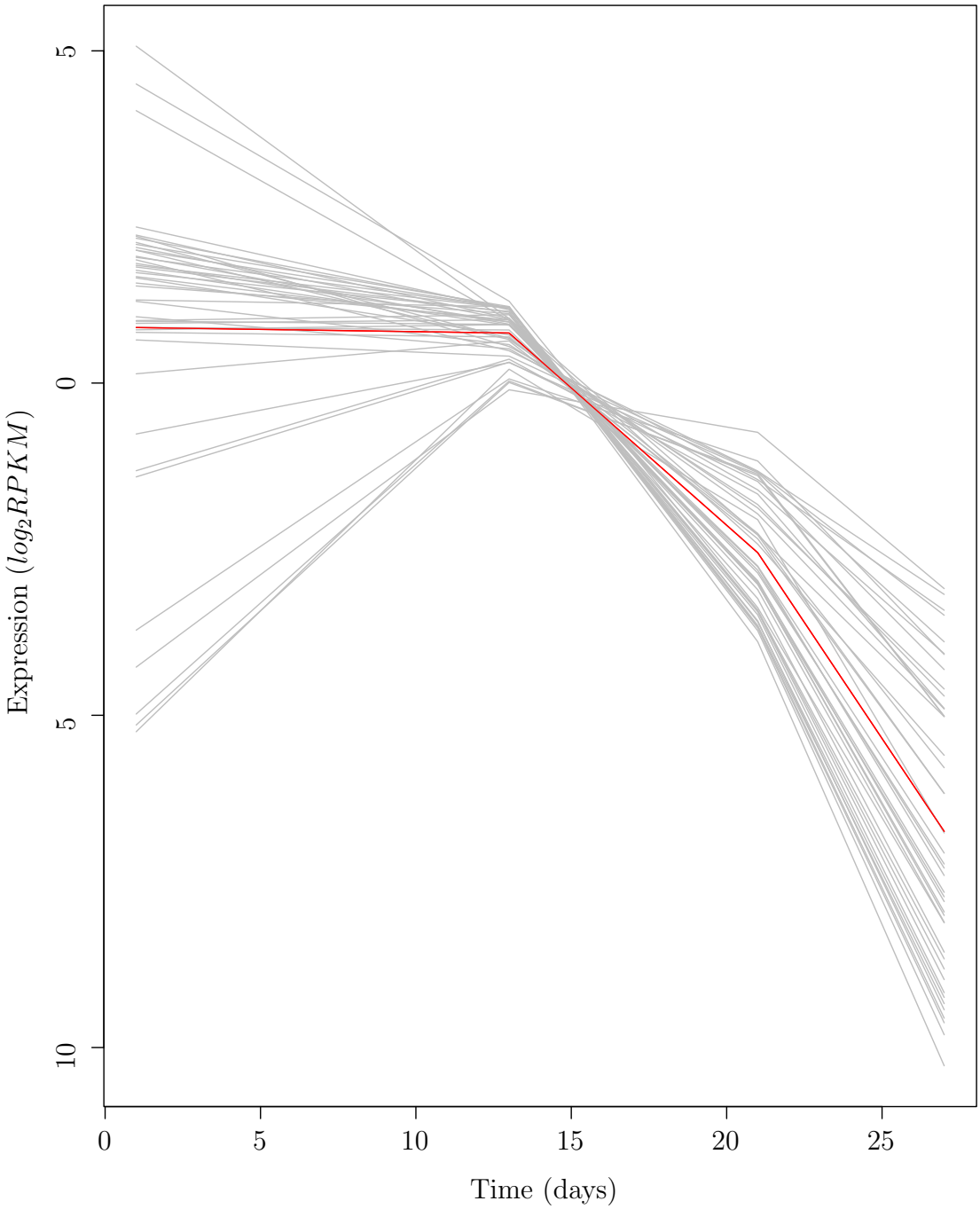


Figure A.6: Transcripts in unpollinated megagametophytes that have late decreases in expression. They fit quadratic regressions where β_2 and β_1 are negative (Category 3).

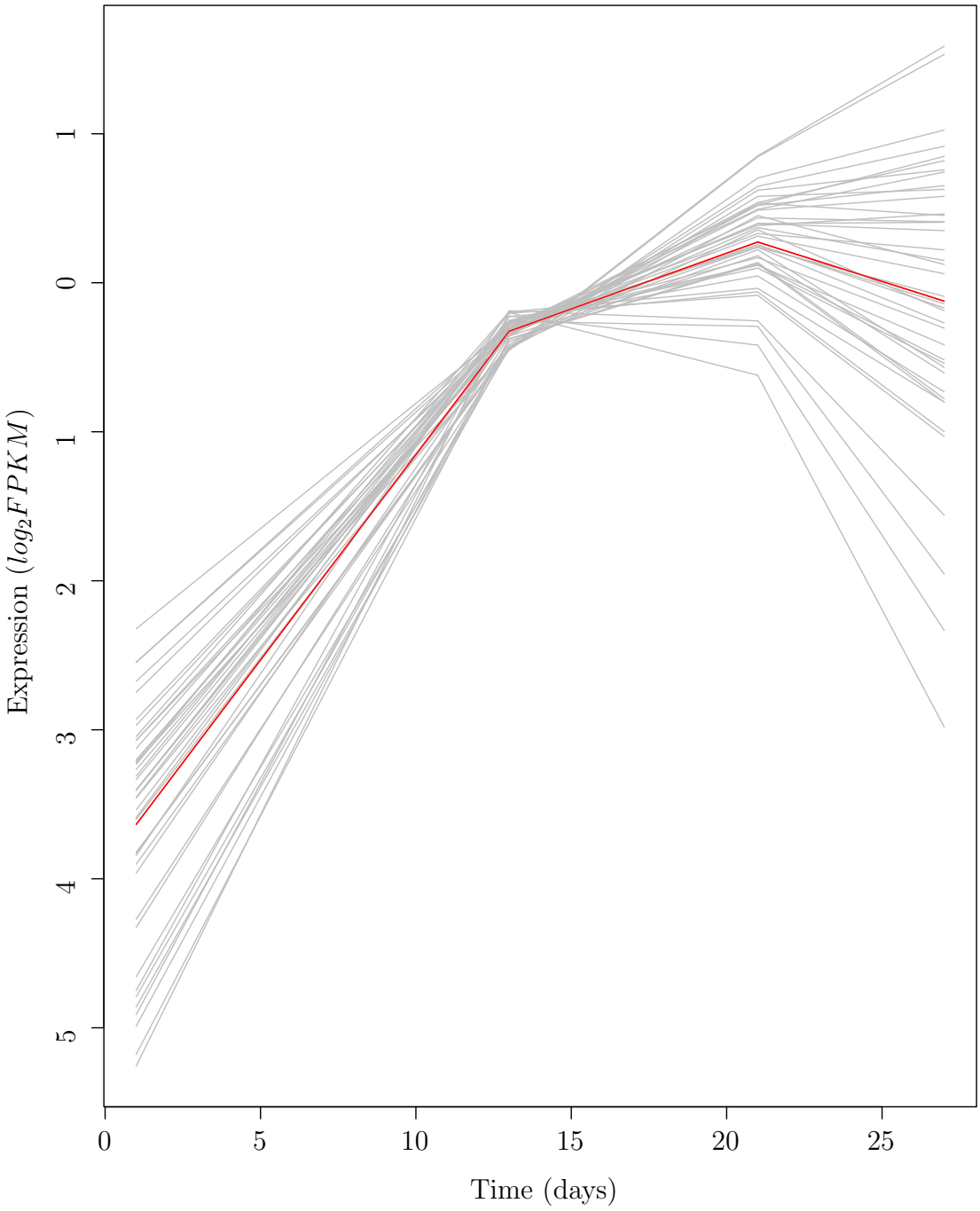


Figure A.7: Transcripts in pollinated megagametophytes that have early increases in expression. They fit quadratic regressions where β_2 is negative and β_1 is positive (Category 4).

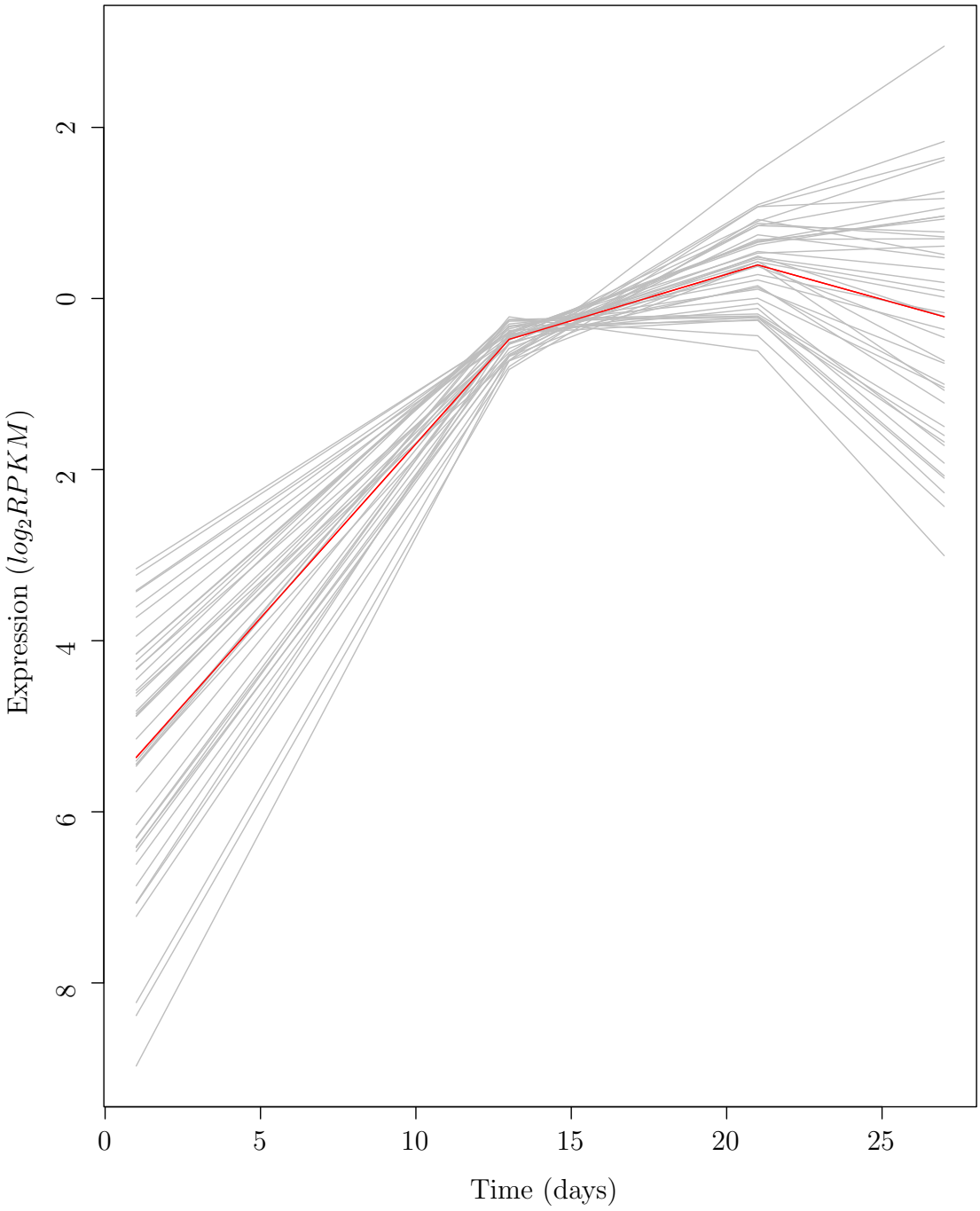


Figure A.8: Transcripts in unpollinated megagametophytes that have early increases in expression. They fit quadratic regressions where β_2 is negative and β_1 is positive (Category 4).

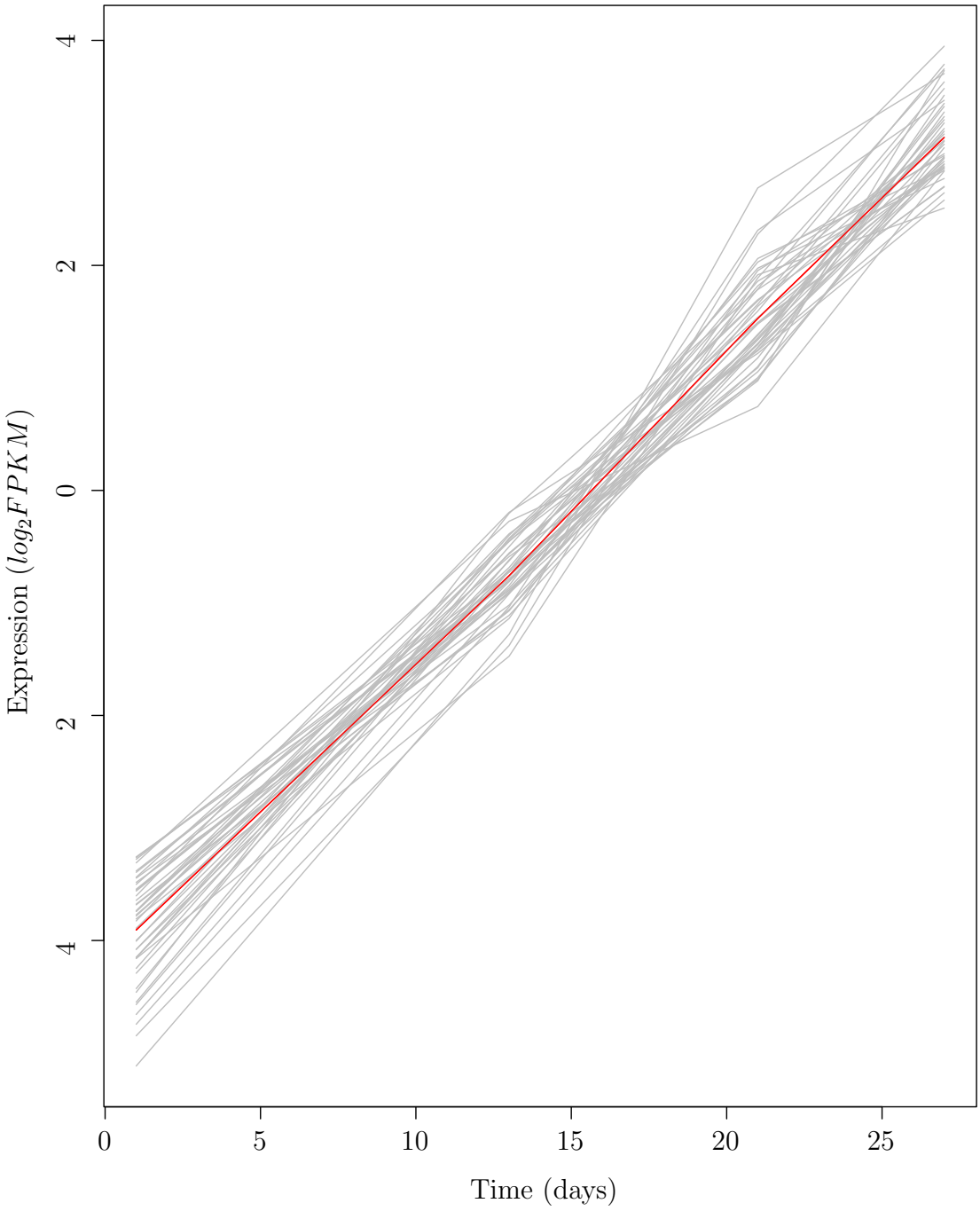


Figure A.9: Transcripts in pollinated megagametophytes that have linear increases expression. They fit quadratic regressions where β_2 is not defined and β_1 is positive (Category 5).

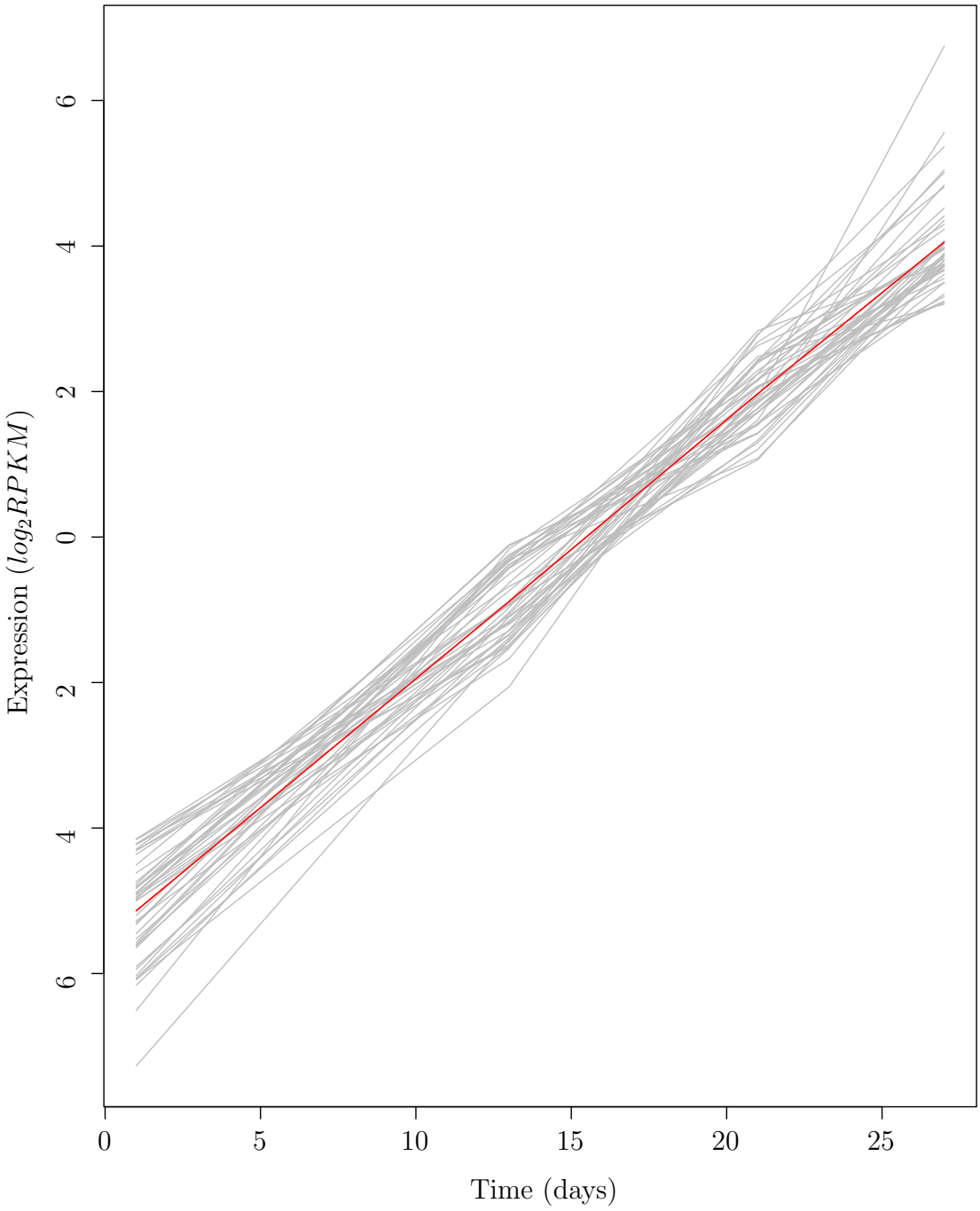


Figure A.10: Transcripts in unpollinated megagametophytes that have linear increases in expression. They fit quadratic regressions where β_2 is not defined and β_1 is positive (Category 5).

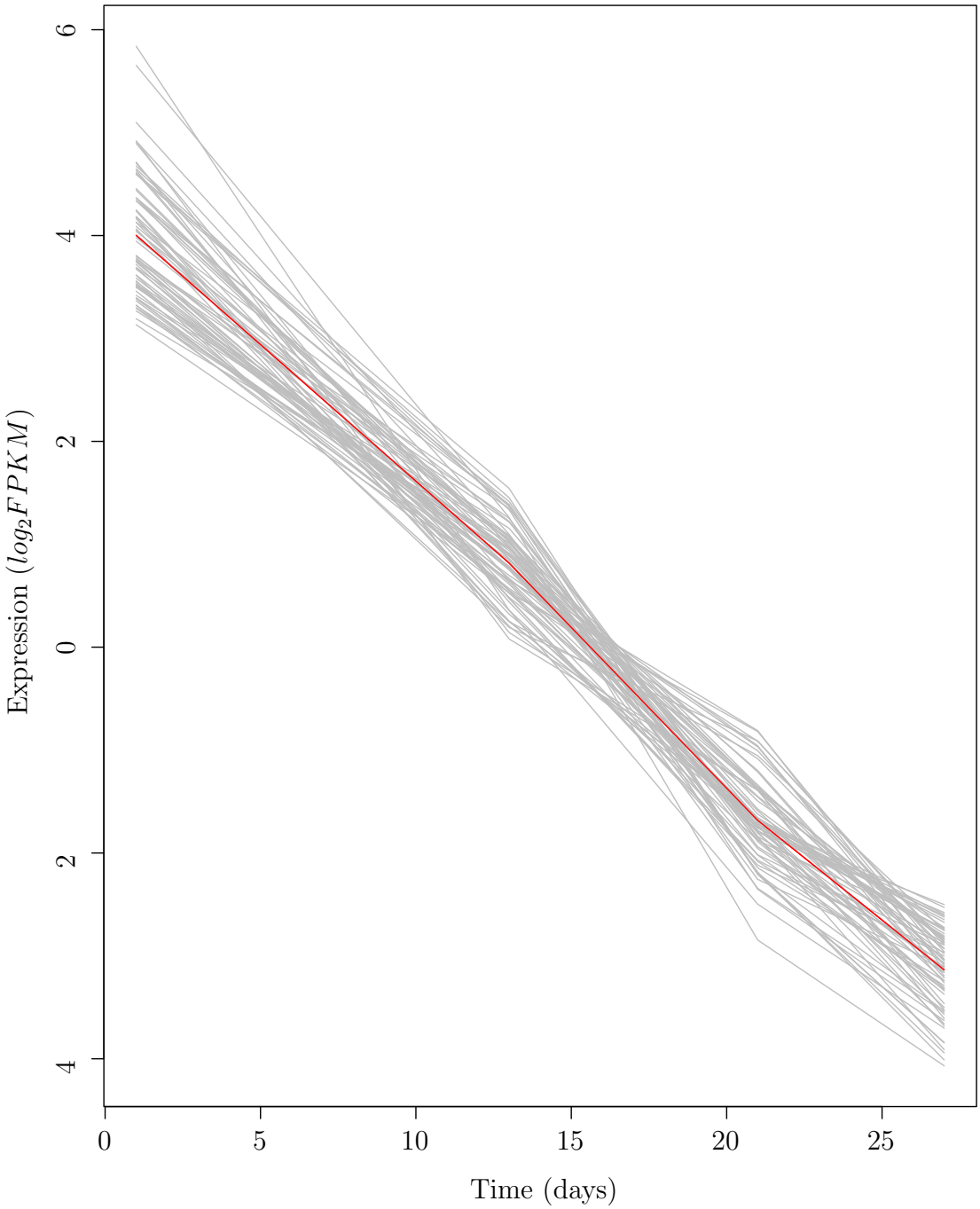


Figure A.11: Transcripts in pollinated megagametophytes that have linear decreases in expression. They fit quadratic regressions where β_2 is not defined and β_1 is negative (Category 6).

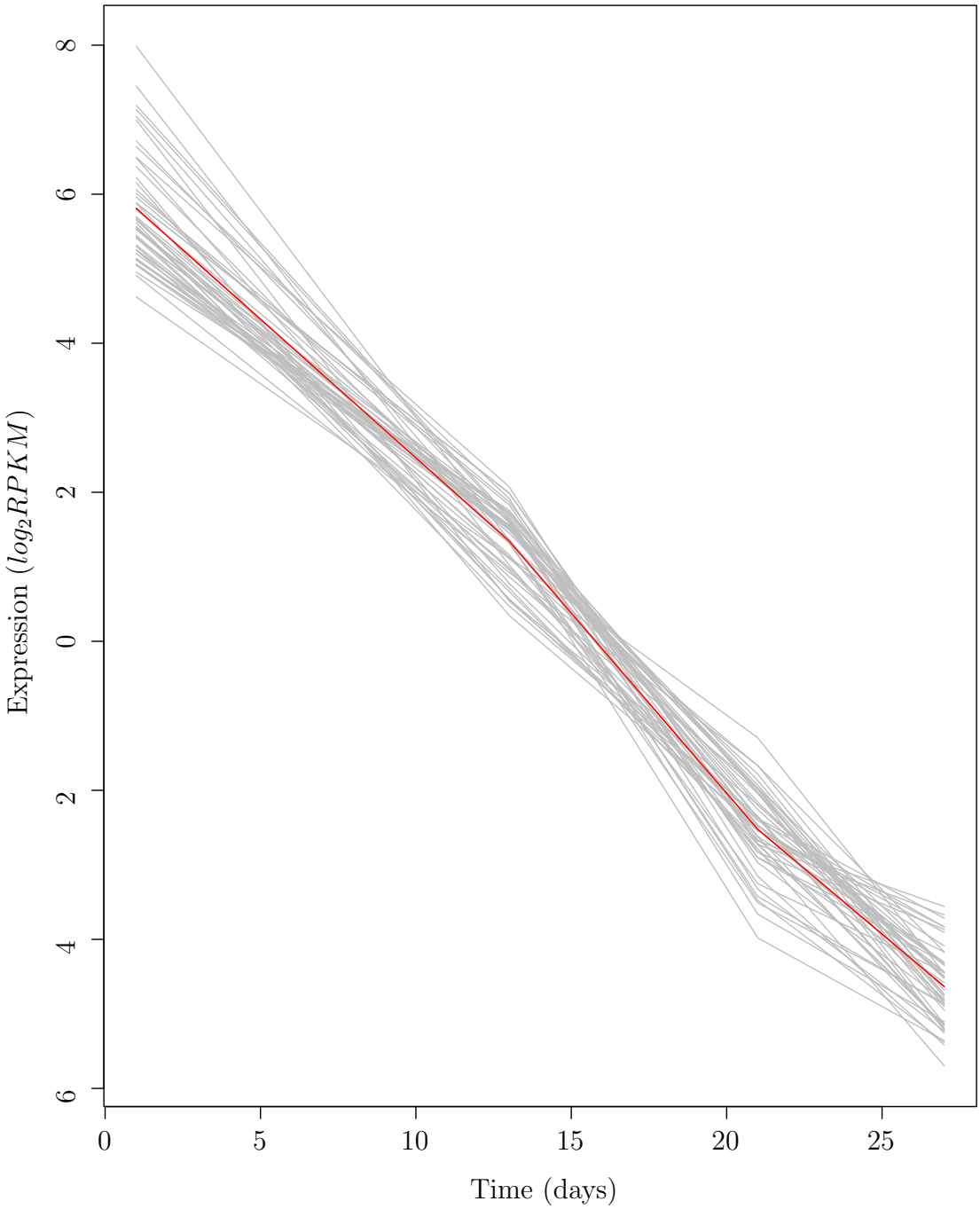


Figure A.12: Transcripts in unpollinated megagametophytes that have linear decreases in expression. They fit quadratic regressions where β_2 is not defined and β_1 is negative (Category 6).

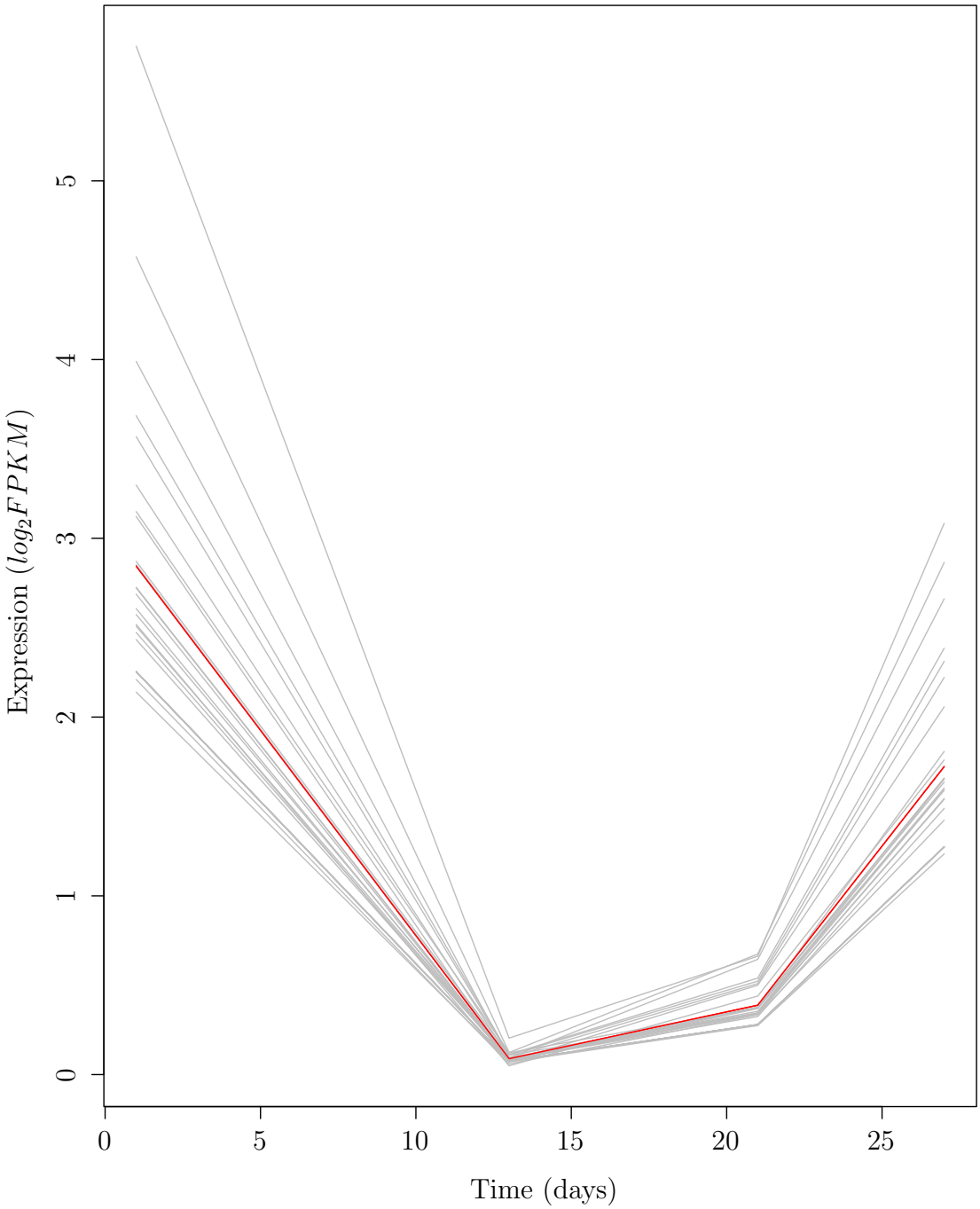


Figure A.13: Transcripts in pollinated megagametophytes that are most highly expressed at the beginning and end of the experiment. They fit quadratic regressions where β_2 is positive and β_1 is not defined (Category 7).

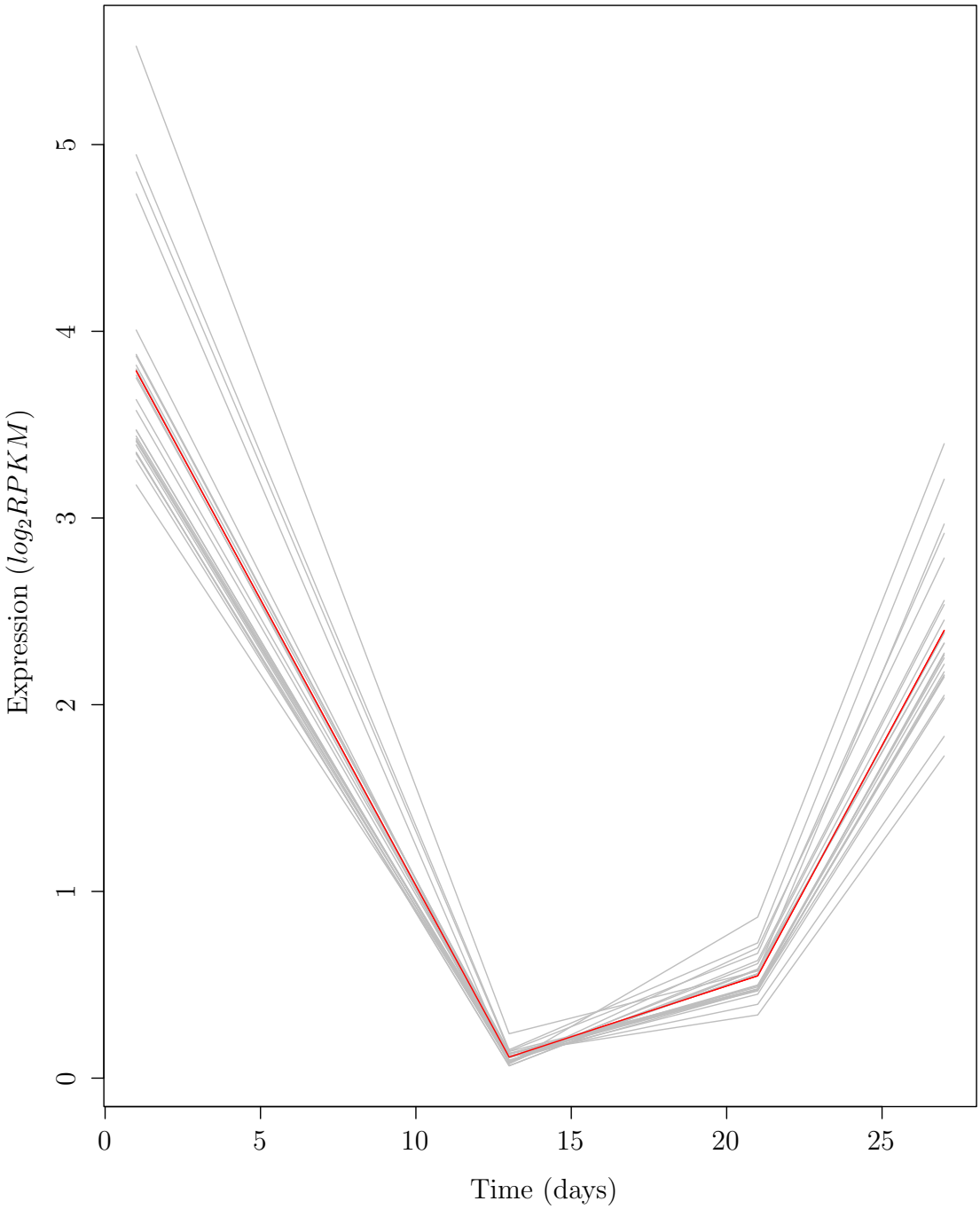


Figure A.14: Transcripts in unpollinated megagametophytes that are most highly expressed at the beginning and end of the experiment. They fit quadratic regressions where β_2 is positive and β_1 is not defined (Category 7).

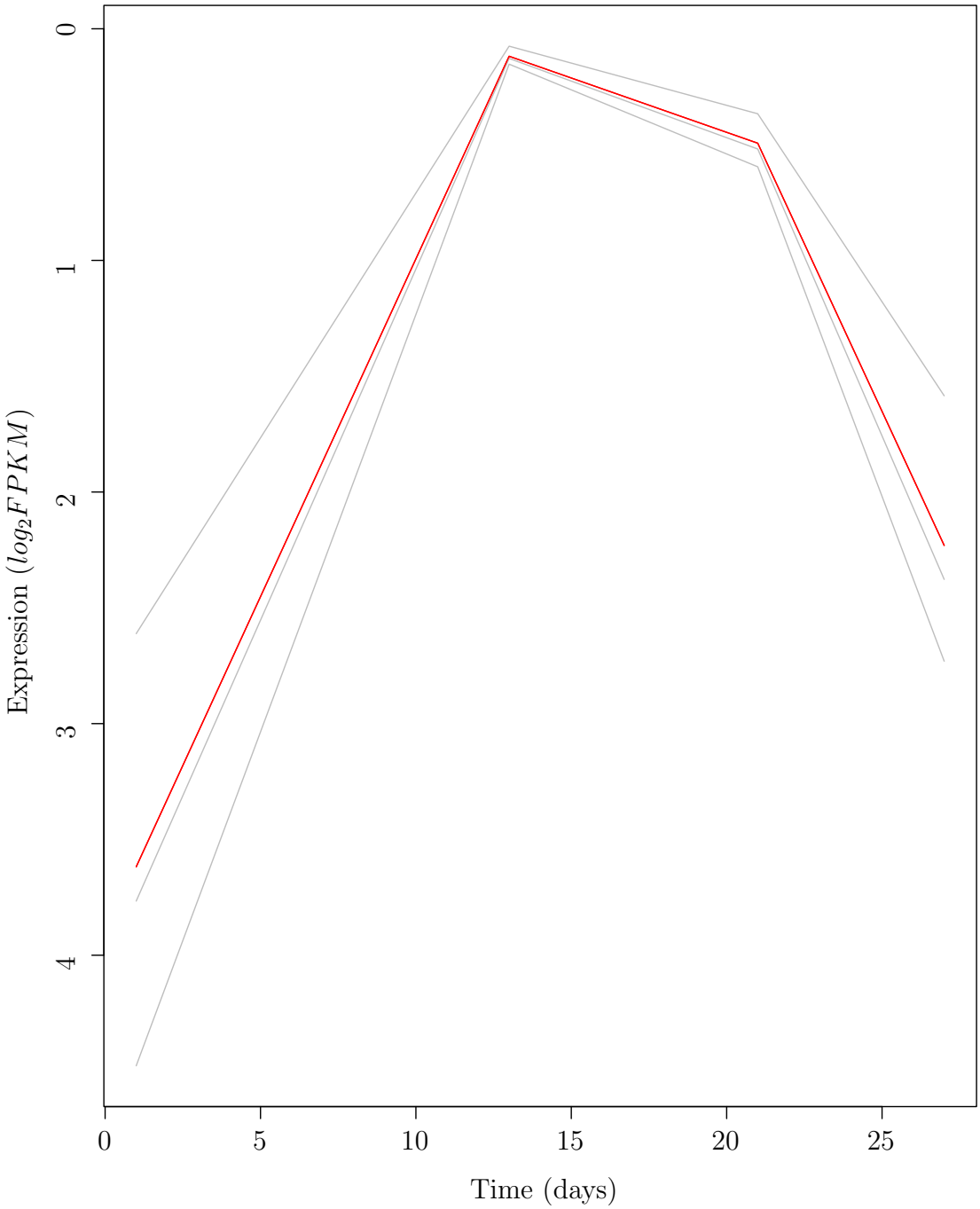


Figure A.15: Transcripts in pollinated megagametophytes that are most highly expressed during the middle timepoints of the experiment. They fit quadratic regressions where β_2 is negative and β_1 is not defined (Category 8).

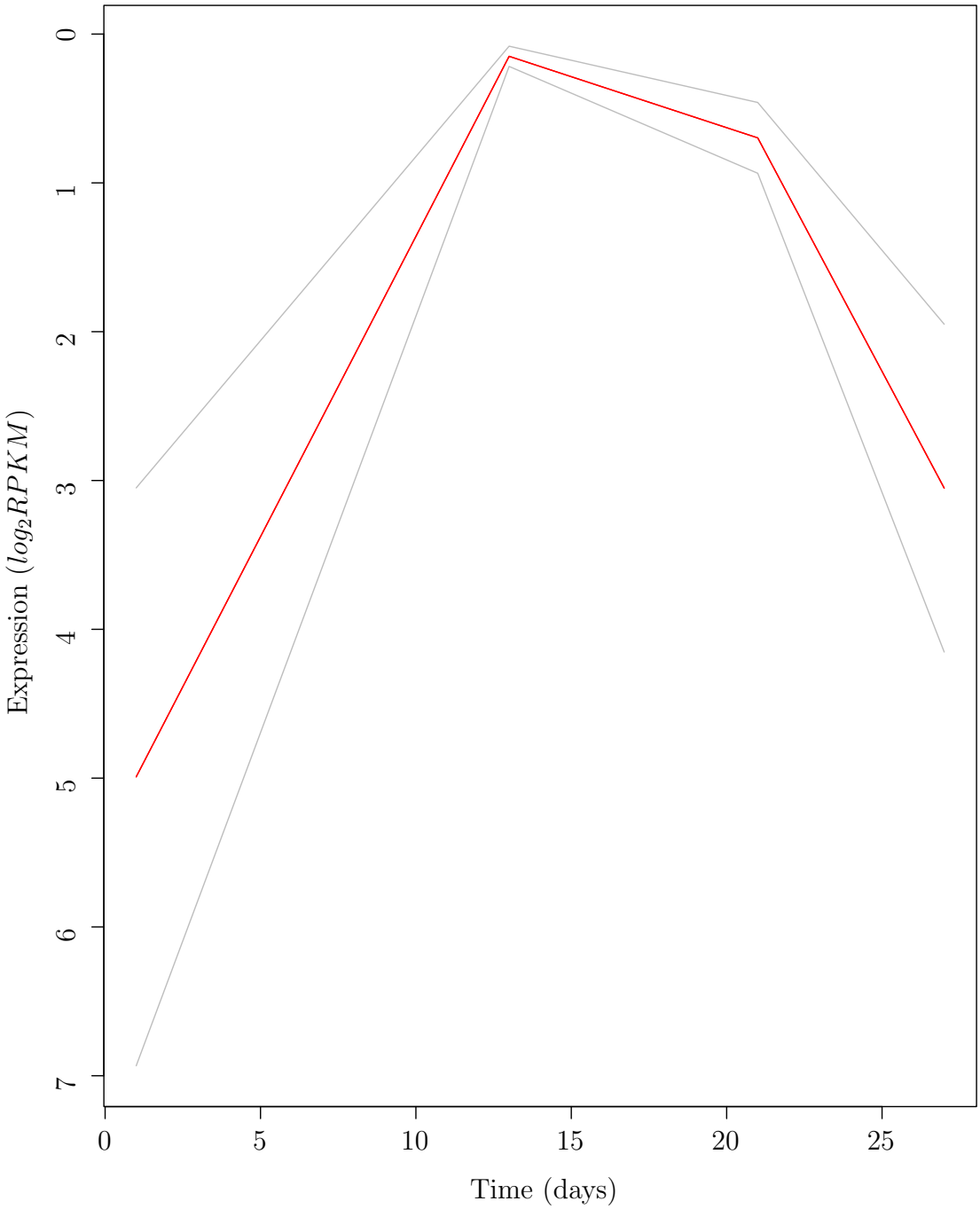


Figure A.16: Transcripts in unpollinated megagametophytes that are most highly expressed during the middle timepoints of the experiment They fit quadratic regressions where β_2 is negative and β_1 is not defined (Category 8).

References

Abrahams S, Lee E, Walker AR, Tanner GJ, Larkin PJ, and Ashton AR. 2003. The *Arabidopsis* TDS4 gene encodes leucoanthocyanidin dioxygenase (LDOX) and is essential for proanthocyanidin synthesis and vacuole development. The Plant Journal 35: 624–636.

Aida M, Ishida T, and Tasaka M. 1999. Shoot apical meristem and cotyledon formation during *Arabidopsis* embryogenesis: interaction among the CUP-SHAPED COTYLEDON and SHOOT MERISTEMLESS genes. Development 126: 1563–70.

Allen GS. 1943. The Embryogeny of *Pseudotsuga taxifolia* (Lamb .) Britt. American Journal of Botany 30: 655–661.

Allen GS, and Owens JN. 1972. *The life history of Douglas-fir*. Information Canada Canada.

Altschul SF, Madden TL, Schäffer aa, Zhang J, Zhang Z, Miller W, and Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389–3402.

Anders S. 2010. HTSeq: Analysing high-throughput sequencing data with python. http://www-huber.embl.de/users/anders/HTSeq.

Anders S, and Huber W. 2010. Differential expression analysis for sequence count data. Genome Biology 11: R106.

Andrews S. 2012. FASTQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Avci U, Petzold HE, Ismail IO, Beers EP, and Haigler CH. 2008. Cysteine proteases XCP1 and XCP2 aid micro-autolysis within the intact central vacuole during xylogenesis in *Arabidopsis* roots. The Plant Journal 56: 303–315.

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale Da, O'Donovan C, Redaschi N, and Yeh LSL. 2005. The Universal Protein Resource (UniProt). Nucleic Acids Research 33: D154–D159.

Bassham DC. 2007. Plant autophagy–more than a starvation response. Current Opinion in Plant Biology 10: 587–593.

Bernhardt A, Mooney S, and Hellmann H. 2010. *Arabidopsis* DDB1a and DDB1b are critical for embryo development. Planta 232: 555–566.

Bethke P, Lonsdale J, Fath A, and Jones R. 1999. Hormonally regulated programmed cell death in barley aleurone cells. The Plant Cell 11: 1033–1046.

Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, Horsman DE, Connors JM, Gascoyne RD, Marra Ma, and Jones SJM. 2009. De novo transcriptome assembly with ABySS. Bioinformatics 25: 2872–2877.

Boisvert S, Laviolette F, and Corbeil J. 2010. Ray : Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies. Journal of Computational Biology 17: 1519–1533.

Bollhöner B, Prestele J, and Tuominen H. 2012. Xylem cell death: emerging understanding of regulation and function. Journal of Experimental Botany 63: 1081–1094.

Bormann BT. 1984. *Douglas-fir: An American Wood*. Forest Service, US Department of Agriculture Washington.

Brenner D, and Mak TW. 2009. Mitochondrial cell death effectors. Current Opinion in Cell Biology 21: 871–877.

Bullard JH, Purdom E, Hansen KD, and Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11: 94.

Carbonell A, Fahlgren N, Garcia-Ruiz H, Gilbert KB, Montgomery Ta, Nguyen T, Cuperus JT, and Carrington JC. 2012. Functional analysis of three *Arabidopsis* ARGONAUTES using slicer-defective mutants. The Plant Cell 24: 3613–3629.

Carman J, and Reese G. 2005. Nutrient and hormone levels in Douglas-fir corrosion cavities, megagametophytes, and embryos during embryony. Canadian Journal of Forest Research 35: 2447–2456.

Chardon F, Bedu M, Calenge F, Klemens PA, Lara S, Clement G, Chietera G, Lëran S, Ferrand M, Lacombe B, Loudet O, Dinant S, Bellini C, Neuhaus E, Daniel-Vedele F, and Krapp A. 2013. Leaf fructose content is controlled by the vacuolar transporter sweet17 in *Arabidopsis*. Current Biology 23: 697–702.

Chipuk JE, Moldoveanu T, Llambi F, Parsons MJ, and Green DR. 2010. The BCL-2 family reunion. Molecular Cell 37: 299–310.

Chiu R, and Nip KM. 2012. Trans-ABySS User Manual. http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss.

Chiwocha S, Rouault G, Abrams S, and Aderkas P. 2006. Parasitism of seed of Douglas fir (*Pseudotsuga menziesii*) by the seed chalcid, *Megastigmus spermotrophus*, and its influence on seed hormone physiology. Sexual Plant Reproduction 20: 19–25.

Chiwocha S, and von Aderkas P. 2002. Endogenous levels of free and conjugated forms of auxin, cytokinins and abscisic acid during seed development in Douglas fir. Plant Growth Regulation 36: 191–200.

Christofferson DE, and Yuan J. 2010. Necroptosis as an alternative form of programmed cell death. Current Opinion in Cell Biology 22: 263–268.

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

Coll NS, Vercammen D, Smidler A, Clover C, Van Breusegem F, Dangl JL, and Epple P. 2010. *Arabidopsis* type I metacaspases control cell death. Science 330: 1393–1397.

Compeau PEC, Pevzner Pa, and Tesler G. 2011. How to apply de Bruijn graphs to genome assembly. Nature Biotechnology 29: 987–991.

Courtois-Moreau CL, Pesquet E, Sjödin A, Muñiz L, Bollhöner B, Kaneda M, Samuels L, Jansson S, and Tuominen H. 2009. A unique program for cell death in xylem fibers of *Populus* stem. The Plant Journal 58: 260–274.

Curtis MJ, and Wolpert TJ. 2004. The victorin-induced mitochondrial permeability transition precedes cell shrinkage and biochemical markers of cell death, and shrinkage occurs without loss of membrane integrity. The Plant Journal 38: 244–259.

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

Degenhardt K, Mathew R, Beaudoin B, Bray K, Anderson D, Chen G, Mukherjee C, Shi Y, Gélinas C, Fan Y, Nelson Da, Jin S, and White E. 2006. Autophagy promotes tumor cell survival and restricts necrosis, inflammation, and tumorigenesis. Cancer Cell 10: 51–64.

del Pozo J, and Dharmasiri S. 2002. AXR1-ECR1dependent conjugation of RUB1 to the *Arabidopsis* cullin AtCUL1 is required for auxin response. The Plant Cell 14: 421–433.

del Pozo JC. 1998. The ubiquitin-related protein RUB1 and auxin response in *Arabidopsis*. Science 280: 1760–1763.

Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, and Jaffrézic F. 2012. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics.

Doelling JH, Walker JM, Friedman EM, Thompson AR, and Vierstra RD. 2002. The APG8/12-activating enzyme APG7 is required for proper nutrient recycling and senescence in *Arabidopsis thaliana*. The Journal of Biological Chemistry 277: 33105–33114.

Domínguez F, Moreno J, and Cejudo FJ. 2001. The nucellus degenerates by a process of programmed cell death during the early stages of wheat grain development. Planta 213: 352–360.

Duan J, Xia C, Zhao G, Jia J, and Kong X. 2012. Optimizing de novo common wheat transcriptome assembly using short-read RNA-Seq data. BMC Genomics 13: 1–12.

Dumont-BéBoux N, Weber M, Ma Y, and P. 1998. Intergeneric pollen - megagametophyte relationships of conifers in vitro. TAG Theoretical and Applied Genetics 97: 881–887.

Dunigan T, Vetter J, and Worley P. 2005. Performance Evaluation of the SGI Altix 3700. 2005 International Conference on Parallel Processing (ICPP'05) pages 231–240.

Earley K, Smith M, Weber R, Gregory B, and Poethig R. 2010. An endogenous F-box protein regulates ARGONAUTE1 in *Arabidopsis thaliana*. Silence 1: 15.

Eckenwalder JE. 2009. *Conifers of the World: The Complete Reference*. Timber Press Portland OR.

Eklund DM, and Edqvist J. 2003. Localization of nonspecific lipid transfer proteins correlate with programmed cell death responses during endosperm degradation in *Euphorbia lagascae* seedlings. Plant Physiology 132: 1249–1259.

Eklund DM, Thelander M, Landberg K, Stå ldal V, Nilsson A, Johansson M, Valsecchi I, Pederson ERa, Kowalczyk M, Ljung K, Ronne H, and Sundberg E. 2010. Homologues of the Arabidopsis thaliana SHI/STY/LRP1 genes control auxin biosynthesis and affect growth and development in the moss *Physcomitrella patens*. Development 137: 1275–84.

Ewing B, and Green P. 1998. Base-Calling of Automated Sequencer Traces Using Phred. II . Error Probabilities. Genome Research 8: 186–194.

Fernando DD, Owens JN, and P. 1998. In vitro fertilization from co-cultured pollen tubes and female gametophytes of Douglas fir (*Pseudotsuga menziesii*). Theoretical and Applied Genetics 96: 1057–1063.

Filonova LH, Bozhkov PV, Brukhin VB, Daniel G, Zhivotovsky B, and von Arnold S. 2000. Two waves of programmed cell death occur during formation and development of somatic embryos in the gymnosperm, Norway spruce. Journal of Cell Science 113: 4399–4411.

Filonova LH, von Arnold S, Daniel G, and Bozhkov PV. 2002. Programmed cell death eliminates all but one embryo in a polyembryonic plant seed. Cell Death and Differentiation 9: 1057–1062.

Finch-Savage WE, and Leubner-Metzger G. 2006. Seed dormancy and the control of germination. The New Phytologist 171: 501–523.

Fuchs Y, and Steller H. 2011. Programmed cell death in animal development and disease. Cell 147: 742–758.

Fukuda H, Watanabe Y, Kuriyama H, Aoyagi S, Sugiyama M, Yamamoto R, Demura T, and Minami A. 1998. Programming of Cell Death during Xylogenesis. Journal of Plant Research 111: 253–256.

Gabriel E, Fagg GE, Bosilca G, Angskun T, Dongarra JJ, Squyres JM, Sahay V, Kambadur P, Barrett B, Lumsdaine A, Castain RH, Daniel DJ, Graham RL, and Woodall TS. 2004. Open MPI: Goals, Concept, and Design of a Next Generation MPI Implementation. Recent Advances in Parallel Virtual Machine and Message Passing Interface pages 353–377.

Galluzzi L, and Krömer G. 2008. Necroptosis: a specialized pathway of programmed necrosis. Cell 135: 1161–1163.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, and Rokhsar DS. 2012. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Research 40: D1178–D1186.

Goossens V, Vos KD, Vercammen D, Steemans M, Vancompernolle K, Fiers W, Vandenabeele P, and Grooten J. 1999. Redox regulation of TNF signaling. Biofactors 10: 145–156.

Goyal K, Walton LJ, and Tunnacliffe A. 2005. LEA proteins prevent protein aggregation due to water stress. The Biochemical Journal 388: 151–157.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson Da, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, and Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29: 644–652.

Grescoe A. 1997. *Giants: the colossal trees of Pacific North America*. Roberts Rinehart Lanham MD.

Groover A, and Jones AM. 1999. Tracheary element differentiation uses a novel mechanism coordinating programmed cell death and secondary cell wall synthesis. Plant Physiology 119: 375–384.

Gutmann M, von aderkas P, Label P, and Marie-Anne L. 1996. Effects of abscisic acid on somatic embryo maturation of hybrid larch. Journal of Experimental Botany 47: 1905–1917.

Häcker G. 2000. The morphology of apoptosis. Cell and Tissue Research 301: 5–17.

Hajduch M, Hearne LB, Miernyk Ja, Casteel JE, Joshi T, Agrawal GK, Song Z, Zhou M, Xu D, and Thelen JJ. 2010. Systems analysis of seed filling in *Arabidopsis*: using general linear modeling to assess concordance of transcript and protein expression. Plant Physiology 152: 2078–2087.

Hansen KD, Brenner SE, and Dudoit S. 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Research 38: e131.

Hansen KD, Irizarry Ra, and Wu Z. 2012. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics 13: 204–216.

Hara-Nishimura I, and Hatsugai N. 2011. The role of vacuole in plant cell death. Cell Death and Differentation 18: 1298–1304.

Hatsugai N, Iwasaki S, Tamura K, Kondo M, Fuji K, Ogasawara K, Nishimura M, and Hara-Nishimura I. 2009. A novel membrane fusion-mediated plant immunity against bacterial pathogens. Genes & Development 23: 2496–2506.

He X, and Kermode AR. 2003a. Nuclease activities and DNA fragmentation during programmed cell death of megagametophyte cells of white spruce (*Picea glauca*) seeds. Plant Molecular Biology 51: 509–521.

He X, and Kermode AR. 2003b. Proteases associated with programmed cell death of megagametophyte cells after germination of white spruce (*Picea glauca*) seeds. Plant Molecular Biology 52: 729–744.

Henschel R, Lieber M, Wu LS, Nista PM, Haas BJ, and LeDuc RD. 2012. Trinity rna-seq assembler performance optimization. In *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond* XSEDE '12 pages 1–8 New York, NY, USA. ACM.

Hermann RK, and Lavender DP. 1999. Douglas-fir planted forests. New Forests 17: 53–70.

Hiratsuka R, Yamada Y, and Terasaka O. 2002. Programmed cell death of *Pinus* nucellus in response to pollen tube penetration. Journal of Plant Research 115: 141–148.

Hökstra Fa, Golovina Ea, and Buitink J. 2001. Mechanisms of plant desiccation tolerance. Trends in Plant Science 6: 431–438.

Holt Ra, and Jones SJM. 2008. The new paradigm of flow cell sequencing. Genome Research 18: 839–846.

Hotton SK, Castro MF, Eigenheer Ra, and Callis J. 2012. Recovery of DDB1a (damaged DNA binding protein1a) in a screen to identify novel RUB-modified proteins in Arabidopsis thaliana. Molecular plant 5: 1163–6.

IFN. 2008. The French forest figures and maps. `http://inventaire-forestier.ign.fr/`.

Igarashi D, Tsuda K, and Katagiri F. 2012. The peptide growth factor, phytosulfokine, attenuates pattern-triggered immunity. The Plant Journal 71: 194–204.

Igasaki T, Akashi N, Ujino-Ihara T, Matsubayashi Y, Sakagami Y, and Shinohara K. 2003. Phytosulfokine stimulates somatic embryogenesis in *Cryptomeria japonica*. Plant & Cell Physiology 44: 1412–1416.

Illumina. 2011. CASAVA v1.8 Changes. `http://support.illumina.com/downloads/casava_18_changes.ilmn`.

Illumina. 2012. Illumina performance specifications. `http://www.illumina.com/systems/hiseq_2500_1500/performance_specifications.ilmn`.

Inoue S, Browne G, Melino G, and Cohen GM. 2009. Ordering of caspases in cells undergoing apoptosis by the intrinsic pathway. Cell Death and Differentation 16: 1053–1061.

Jackson MB, and Armstrong W. 1999. Formation of aerenchyma and the processes of plant ventilation in relation to soil flooding and submergence. Plant Biology 1: 274–287.

Jensen MK, Hagedorn PH, de Torres-Zabala M, Grant MR, Rung JH, Collinge DB, and Lyngkjaer MF. 2008. Transcriptional regulation by an NAC (NAM-ATAF1,2-CUC2) transcription factor attenuates ABA signalling for efficient basal defence towards *Blumeria graminis* f. sp. hordei in *Arabidopsis*. The Plant Journal 56: 867–880.

Jia L, Wu Q, Ye N, Liu R, Shi L, Xu W, Zhi H, Rahman aNMRB, Xia Y, and Zhang J. 2012. Proanthocyanidins inhibit seed germination by maintaining a high level of abscisic acid in *Arabidopsis thaliana*(F). Journal of Integrative Plant Biology 54: 663–673.

Johnston Ja, Ward CL, and Kopito RR. 1998. Aggresomes: a cellular response to misfolded proteins. The Journal of Cell Biology 143: 1883–1898.

Ju ST, Panka DJ, Cui H, Ettinger R, El-Khatib M, Sherr DH, Stanger BZ, and Marshak-Rothstein A. 1995. Fas(CD95)/FasL interactions reqired for programmed cell death after T-cell activation. Nature 373: 444–448.

Kepinski S, and Leyser O. 2005. The *Arabidopsis* F-box protein TIR1 is an auxin receptor. Nature 435: 446–451.

Kim EY, Seo YS, and Kim WT. 2011. AtDSEL, an *Arabidopsis* cytosolic DAD1-like acylhydrolase, is involved in negative regulation of storage oil mobilization during seedling establishment. Journal of Plant Physiology 168: 1705–1709.

Kim S, Mollet JC, Dong J, Zhang K, Park SY, and Lord EM. 2003. Chemocyanin, a small basic protein from the lily stigma, induces pollen tube chemotropism. Proceedings of the National Academy of Sciences of the United States of America 100: 16125–16130.

Kircher M, Heyn P, and Kelso J. 2011. Addressing challenges in the production and analysis of Illumina sequencing data. BMC Genomics 12: 382.

Kirkin V, McEwan DG, Novak I, and Dikic I. 2009. A role for ubiquitin in selective autophagy. Molecular Cell 34: 259–269.

Kleinow T, Himbert S, Krenz B, Jeske H, and Koncz C. 2009. NAC domain transcription factor ATAF1 interacts with SNF1-related kinases and silencing of its subfamily causes severe developmental defects in *Arabidopsis*. Plant Science 177: 360–370.

Klionsky DJ. 2000. Autophagy as a Regulated Pathway of Cellular Degradation. Science 290: 1717–1721.

Koizumi K, and Gallagher KL. 2013. Identification of SHRUBBY, a SHORT-ROOT and SCARECROW interacting protein that controls root growth and radial patterning. Development 140: 1292–1300.

Kolosova N, Miller B, Ralph S, Ellis BE, Douglas C, Ritland K, and Bohlmann J. 2004. Isolation of high-quality RNA from gymnosperm and angiosperm trees. BioTechniques 36: 821–824.

Kong Y. 2011. Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. Genomics 98: 152–153.

Korbecka G, Klinkhamer PGL, and Vrieling K. 2002. Selective embryo abortion hypothesis revisited–A molecular approach. Plant Biology 4: 298–310.

Kothakota S, Azuma T, Reinhard C, Klippel A, Tang J, Chu K, Mcgarry TJ, Kirschner MW, Koths K, Kwiatkowski DJ, and Williams LT. 1997. Caspase-3-generated fragment of gelsolin: Effector of morphological change in apoptosis. Science 294: 3–8.

Kraft C, Peter M, and Hofmann K. 2010. Selective autophagy: ubiquitin-mediated recognition and beyond. Nature Cell Biology 12: 836–841.

Krömer G, and Jäättelä M. 2005. Lysosomes and autophagy in cell death control. Nature Reviews Cancer 5: 886–898.

Krömer G, and Levine B. 2008. Autophagic cell death: the story of a misnomer. Nature Reviews Molecular Cell Biology 9: 1004–1010.

Krueger F. 2013. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries.

Kuusk S, Sohlberg JJ, Magnus Eklund D, and Sundberg E. 2006. Functionally redundant SHI family genes regulate *Arabidopsis* gynoecium development in a dose-dependent manner. The Plant Journal 47: 99–111.

Lakhani SA, Masud A, Kuida K, Porter GA, Booth CJ, Mehal WZ, Inayat I, and Flavell RA. 2006. Caspases 3 and 7: Key mediators of mitochondrial events of apoptosis. Science 311: 847–851.

Langmead B, and Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9: 357–359.

Langmead B, Trapnell C, Pop M, and Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10: R25.

Larsson E, Sundström JF, Sitbon F, and von Arnold S. 2012. Expression of PaNAC01, a *Picea abies* CUP-SHAPED COTYLEDON orthologue, is regulated by polar auxin transport and associated with differentiation of the shoot apical meristem and formation of separated cotyledons. Annals of Botany 110: 923–934.

Leal I, and Misra S. 1993. Developmental gene expression in conifer embryogenesis and germination. III. Analysis of crystalloid protein mRNAs and desiccation protein mRNAs in the developing embryo and megagametophyte of white spruce (*Picea glauca* (Moench) Voss). Plant Science 88: 25–37.

Lee Cy, and Baehrecke EH. 2001. Steroid regulation of autophagic programmed cell death during development. Development 128: 1443–1455.

Levine B, and Klionsky DJ. 2004. Development by Self-Digestion: Molecular Mechanisms and Biological Functions of Autophagy Review. Developmental Cell 6: 463–477.

Levine B, and Yuan J. 2005. Autophagy in cell death: an innocent convict? The Journal of Clinical Investigation 115: 2679–2688.

Li B, and Dewey CN. 2011. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323.

Li B, Ruotti V, Stewart RM, Thomson Ja, and Dewey CN. 2010a. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics 26: 493–500.

Li H, and Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

Li H, and Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589–595.

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

Li J, and Yuan J. 2008. Caspases in apoptosis and beyond. Oncogene 27: 6194–6206.

Li P, Nijhawan D, Budihardjo I, Srinivasula SM, Ahmad M, Alnemri ES, and Wang X. 1997. Cytochrome *c* and dATP-dependent formation of Apaf-1/Caspase-9 Complex initiates an apoptotic protease cascade. Cell 91: 479–489.

Li R, Lan SY, and Xu ZX. 2004. Programmed cell death in wheat during starchy endosperm development. Journal of Plant Physiology and Molecular Biology 30: 183–188.

Li R, Yu K, and Hildebrand DF. 2010b. DGAT1, DGAT2 and PDAT expression in seeds and other tissues of epoxy and hydroxy fatty acid accumulating plants. Lipids 45: 145–157.

Li W, and Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658–1659.

Lim PO, Kim HJ, and Nam HG. 2007. Leaf senescence. Annual Review of Plant Biology 58: 115–136.

Lin H, Ma X, Feng W, and Samatova NF. 2011. Coordinating Computation and I/O in Massively Parallel Sequence Search. IEEE Transactions on Parallel and Distributed Systems 22: 539–542.

Lindgreen S. 2012. AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Research Notes 5: 337.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, and Law M. 2012a. Comparison of next-generation sequencing systems. Journal of Biomedicine & Biotechnology 2012: 251364.

Liu S, Zhang Y, Zhou Z, Waldbieser G, Sun F, Lu J, Zhang J, Jiang Y, Zhang H, Wang X, Rajendran K, Khoo L, Kucuktas H, Peatman E, and Liu Z. 2012b. Efficient assembly and annotation of the transcriptome of catfish by RNA-Seq analysis of a doubled haploid homozygote. BMC Genomics 13: 595.

Liu Y, and Bassham DC. 2010. TOR is a negative regulator of autophagy in *Arabidopsis thaliana*. PloS One 5: e11883.

Liu Y, and Bassham DC. 2012. Autophagy: pathways for self-eating in plant cells. Annual Review of Plant Biology 63: 215–237.

Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, and Usadel B. 2012. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Research 40: W622–W627.

Lu PL, Chen NZ, An R, Su Z, Qi BS, Ren F, Chen J, and Wang XC. 2007. A novel drought-inducible gene, ATAF1, encodes a NAC family protein that negatively regulates the expression of stress-responsive genes in *Arabidopsis*. Plant Molecular Biology 63: 289–305.

Luo X, Budihardjo I, Zou H, Slaughter C, and Wang X. 1998. Bid, a Bcl2 Interacting Protein, mediates cytochrome *c* release from mitochondria in response to activation of cell surface death receptors. Cell 94: 481–490.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben La, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt Ka, Volkmer Ga, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, and Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet Journal 17: 10–12.

Martinou JC, and Youle RJ. 2011. Mitochondria in apoptosis: Bcl-2 family members and mitochondrial dynamics. Developmental Cell 21: 92–101.

Miller JR, Koren S, and Sutton G. 2010. Assembly algorithms for next-generation sequencing data. Genomics 95: 315–327.

Mizushima N. 2005. The pleiotropic role of autophagy: from protein metabolism to bactericide. Cell Death and Differentation 12: 1535–1541.

Morant M, Jørgensen K, Schaller H, Pinot F, Møller BL, Werck-Reichhart D, and Bak S. 2007. CYP703 is an ancient cytochrome P450 in land plants catalyzing in-chain hydroxylation of lauric acid to provide building blocks for sporopollenin synthesis in pollen. The Plant Cell 19: 1473–1487.

Motose H, Iwamoto K, Endo S, Demura T, Sakagami Y, Matsubayashi Y, Moore KL, and Fukuda H. 2009. Involvement of phytosulfokine in the attenuation of stress response during the transdifferentiation of zinnia mesophyll cells into tracheary elements. Plant Physiology 150: 437–447.

Narendra D, Tanaka A, Suen DF, and Youle RJ. 2008. Parkin is recruited selectively to impaired mitochondria and promotes their autophagy. The Journal of Cell Biology 183: 795–803.

Narendra DP, Jin SM, Tanaka A, Suen DF, Gautier Ca, Shen J, Cookson MR, and Youle RJ. 2010. PINK1 is selectively stabilized on impaired mitochondria to activate Parkin. PLoS Biology 8: 1–21.

NCBI. 2009. Blast program selection guide: BLAST database content. `http://blast.ncbi.nlm.nih.gov/BLAST_guide.pdf`.

Niskanen AM, Lu J, Seitz S, Keinonen K, and von Weissenberg, Kim abd Pappinen A. 2004. Effect of parent genotype in somatic embryogenesis in scots pine (*Pinus sylvestris*). Tree Physiology 24: 1259–1265.

Norholm MHH, Nour-Eldin HH, Brodersen P, Mundy J, and Halkier B. 2006. Expression of the *Arabidopsis* high-affinity hexose transporter STP13 correlates with programmed cell death. FEBS letters 580: 2381–2387.

Obara K, Kuriyama H, and Fukuda H. 2001. Direct evidence of active and rapid nuclear degradation triggered by vacuole rupture during programmed cell death in *Zinnia*. Plant Physiology 125: 615–626.

Oehmen C, and Nieplocha J. 2006. ScalaBLAST : A Scalable Implementation of BLAST for High-Performance Data-Intensive Bioinformatics Analysis. IEEE Transactions on Parallel and Distributed Systems 17: 740–749.

Oku M, and Sakai Y. 2010. Peroxisomes as dynamic organelles: autophagic degradation. The FEBS Journal 277: 3289–3294.

O'Leary SJ, Poulis BA, and Von Aderkas P. 2007. Identification of two thaumatin-like proteins (tlps) in the pollination drop of hybrid yew that may play a role in pathogen defence during pollen collection. Tree Physiology 27: 1649–1659.

Oshlack A, Robinson MD, and Young MD. 2010. From RNA-seq reads to differential expression results. Genome Biology 11: 220.

Owens J, Morris S, and Misra S. 1993. The ultrastructural, histochemical, and biochemical development of the post-fertilization megagametophyte and the zygotic embryo of *Pseudotsuga menziesii*. Canadian Journal of Forest Science 23: 816–827.

Owens J, Simpson SJ, and Molder M. 1981. The pollination mechanism and the optimal time of pollination in Douglas-fir (*Pseudotsuga menziesii*). Canadian Journal of Forest Research 11: 36–50.

Owens JN. 1969. The relative importance of initiation and early development on cone production in Douglas fir. Canadian Journal of Botany 47: 1039–1049.

Owens JN, Colangeli AM, and Morris SJ. 1991. Factors affecting seed set in Douglas-fir (*Pseudotsuga menziesii*). Canadian Journal of Botany 69: 229–238.

Owens JN, and Morris SJ. 1990. Cytological basis for cytoplasmic inheritance in *Pseudotsuga menziesii*. I. Pollen tube and archegonial development. American Journal of Botany 77: 433–445.

Owens JN, and Smith FH. 1964. The initiation and early development of the seed cone of Douglas fir. Canadian Journal of Botany 42: 1031–1047.

Pankiv S, Clausen THy, Lamark T, Brech A, Bruun JA, Outzen H, Ø vervatn A, Bjø rkøy G, and Johansen T. 2007. p62/SQSTM1 binds directly to Atg8/LC3 to facilitate degradation of ubiquitinated protein aggregates by autophagy. The Journal of Biological Chemistry 282: 24131–24145.

Pastore JJ, Limpuangthip A, Yamaguchi N, Wu MF, Sang Y, Han SK, Malaspina L, Chavdaroff N, Yamaguchi A, and Wagner D. 2011. LATE MERISTEM IDENTITY2 acts together with LEAFY to activate APETALA1. Development 138: 3189–3198.

Pearson WR, and Lipman DJ. 1988. Improved tools for biological sequence comparison. PNAS 85: 2444–2448.

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, and Quackenbush J. 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19: 651–652.

Pettersson E, Lundeberg J, and Ahmadian A. 2009. Generations of sequencing technologies. Genomics 93: 105–111.

PicardTools. 2009. Overview of Picard command-line tools. `http://picard.sourceforge.net/command-line-overview.shtml`.

Pop C, and Salvesen GS. 2010. Human Caspases: Activation, specificity, and regulation. The Journal of Biological Chemistry 284: 21777–21781.

Porter AG, and Ja RU. 1999. Emerging roles of caspase-3 in apoptosis. Cell Death and Differentiation 6: 99–104.

Portis AR. 2003. Rubisco activase - Rubisco's catalytic chaperone. Photosynthesis Research 75: 11–27.

Pullman GS, and Bucalo K. 2011. Pine somatic embryogenesis using zygotic embryos as explants. Methods in Molecular Biology 710: 267–285.

Quilichini TD, Friedmann MC, Samuels aL, and Douglas CJ. 2010. ATP-binding cassette transporter G26 is required for male fertility and pollen exine formation in *Arabidopsis*. Plant Physiology 154: 678–690.

Quinlan AR, Clark Ra, Sokolova S, Leibowitz ML, Zhang Y, Hurles ME, Mell JC, and Hall IM. 2010. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. Genome Research 20: 623–635.

Quirino BF, Noh YS, Himelblau E, and Amasino RM. 2000. Molecular aspects of leaf senescence. Trends in Plant Science 5: 278–282.

Rabinowitz JD, and White E. 2010. Autophagy and metabolism. Science 330: 1344–1348.

Rice P, Longden I, and Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. Trends in Genetics 16: 2–3.

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

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu AL, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJM, Hoodless PA, and Birol I. 2010. De novo assembly and analysis of RNA-seq data. Nature Methods 7: 909–912.

Robinson MD, McCarthy DJ, and Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140.

Roche. 2011. GS FLX+ System. Technical report Roche Applied Science.

Ronaghi M. 2001. Pyrosequencing Sheds Light on DNA Sequencing. Genome Research 11: 3–11.

Rothberg JM, and Leamon JH. 2008. The development and impact of 454 sequencing. Nature Biotechnology 26: 1117–1124.

Rouault G, Turgeon J, Candau JN, Roques A, and Aderkas P. 2004. Oviposition strategies of conifer seed chalcids in relation to host phenology. Naturwissenschaften 91: 472–480.

Rudel T, and Bokoch GM. 1997. Membrane and morphological changes in apoptotic cells regulated by caspase-mediated activation of PAK2. Science 276: 1571–1574.

Schliesky S, Gowik U, Weber APM, and Bräutigam A. 2012. RNA-Seq assembly - Are we there yet? Frontiers in Plant Science 3: 1–12.

Schofield Ra, Bi YM, Kant S, and Rothstein SJ. 2009. Over-expression of STP13, a hexose transporter, improves plant growth and nitrogen use in *Arabidopsis thaliana* seedlings. Plant, Cell & Environment 32: 271–285.

Schubert R, Panitz R, Manteuffel R, Nagy I, Wobus U, and Bäumlein H. 1994. Tissue-specific expression of an oat 12S seed globulin gene in developing tobacco seeds: differential mRNA and protein accumulation. Plant Molecular Biology 26: 203–210.

Schulz MH, Zerbino DR, Vingron M, and Birney E. 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28: 1086–1092.

Schussler EE, and Longstreth DJ. 2000. Changes in cell structure during the formation of root aerenchyma in *Sagittaria lancifolia* (Alismataceae). American Journal of Botany 87: 12–19.

Serrano I, Irene S, Pelliccione S, Salvatore P, Olmedilla A, and Adela O. 2010. Programmed-cell-death hallmarks in incompatible pollen and papillar stigma cells of *Olea europaea* L. under free pollination. Plant Cell Reports 29: 561–572.

Shendure J, and Ji H. 2008. Next-generation DNA sequencing. Nature Biotechnology 26: 1135–1145.

Shimizu S, Narita M, and Tsujimoto Y. 1999. Bcl-2 family proteins regulate the release of apoptogenic cytochrome *c* by the mitochondrial channel VDAC. Nature 399: 483–487.

Simpson JT, Wong K, Jackman SD, Schein JE, and Jones SJM. 2009. ABySS : A parallel assembler for short read sequence data. Genome Research 19: 1117–1123.

Smith B, and Darr D. 2004. *U.S. Forest Resource Facts and Historical Trends.* Forest Service, US Department of Agriculture Washington.

Smith DL, and Fedoroff NV. 1995. LRP1, a gene expressed in lateral and adventitious root primordia of *Arabidopsis*. The Plant Cell 7: 735–745.

Smith S, and Gilbert J. 2003. National Inventory of Woodland and Trees. `http://www.forestry.gov.uk/forestry/hcou-54pg9u`.

Sreenivasulu N, Radchuk V, Strickert M, Miersch O, Weschke W, and Wobus U. 2006. Gene expression patterns reveal tissue-specific signaling networks controlling programmed cell death and ABA- regulated maturation in developing barley seeds. The Plant Journal 47: 310–327.

Staswick P, Tiryaki I, and Rowe M. 2002. Jasmonate response locus JAR1 and several related *Arabidopsis* genes encode enzymes of the firefly luciferase superfamily that show activity on jasmonic, salicylic, and indole-3-acetic acids in an assay for adenylation. The Plant Cell 14: 1405–1415.

Staswick PE, and Tiryaki I. 2004. The oxylipin signal jasmonic acid is activated by an enzyme that conjugates it to isoleucine in *Arabidopsis*. The Plant Cell 16: 2117–2127.

Stirnberg P, Furner IJ, and O Leyser HM. 2007. MAX2 participates in an SCF complex which acts locally at the node to suppress shoot branching. The Plant Journal 50: 80–94.

Sturn A, Quackenbush J, and Trajanoski Z. 2002. Genesis: cluster analysis of microarray data. Bioinformatics 18: 207–208.

Sturrock R, Islam M, and AKM E. 2007. Hostpathogen interactions in douglas-fir seedlings infected by *Phellinus sulphurascens*. Phytopathology 97: 1406–1414.

Suen DF, Norris KL, and Youle RJ. 2008. Mitochondrial dynamics and apoptosis. Genes & Development 22: 1577–1590.

Sun J, Zhang J, Larue CT, and Huber SC. 2011. Decrease in leaf sucrose synthesis leads to increased leaf starch turnover and decreased rubp regeneration-limited photosynthesis but not rubisco-limited photosynthesis in arabidopsis null mutants of spsa1. Plant, Cell and Environment 35: 592604.

Sun Z, and Zhu Y. 2012. Systematic comparison of RNA-Seq normalization methods using measurement error models. Bioinformatics 28: 2584–2591.

Surget-Groba Y, and Montoya-Burgos JI. 2010. Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Research 20: 1432–1440.

Takeshige K, Baba M, Tsuboi S, Noda T, and Ohsumi Y. 1992. Autophagy in yeast demonstrated with proteinase-deficient mutants and conditions for its induction. The Journal of Cell Biology 119: 301–311.

Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, and Conesa A. 2011. Differential expression in RNA-seq: a matter of depth. Genome Research 21: 2213–2223.

Taylor RC, Cullen SP, and Martin SJ. 2008. Apoptosis: controlled demolition at the cellular level. Nature Reviews Molecular Cell Biology 9: 231–241.

Thompson AR, Doelling JH, Suttangkakul A, and Vierstra RD. 2005. Autophagic nutrient recycling in *Arabidopsis* directed by the ATG8 and ATG12 conjugation pathways. Plant Physiology 138: 2097–2110.

Thompson CB. 1995. Apoptosis in the Treatment and Pathogenesis of Disease. Science 267: 1456–1462.

Thorburn A. 2004. Death receptor-induced cell killing. Cellular Signalling 16: 139–144.

Tiwari S, Wang X, Hagen G, and Guilfoyle TJ. 2001. AUX/IAA proteins are active repressors, and their stability and activity are modulated by auxin. The Plant Cell 13: 2809–2822.

Ton J, Flors V, and Mauch-Mani B. 2009. The multifaceted role of ABA in disease resistance. Trends in Plant Science 14: 310–317.

Trinity Team. 2013. Trinity: Abundance Estimation Using RSEM. `http://trinityrnaseq.sourceforge.net/analysis/abundance_estimation.html`.

Tunnacliffe A, and Wise MJ. 2007. The continuing conundrum of the LEA proteins. Naturwissenschaften 94: 791–812.

Turcatti G, Romieu A, Fedurco M, and Tairi AP. 2008. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. Nucleic Acids Research 36: e25.

Umehara M, Ogita S, Sasamoto H, Eun CH, Matsubayashi Y, Sakagami Y, and Kamada H. 2005. Two stimulatory effects of the peptidyl growth factor phytosulfokine during somatic embryogenesis in Japanese larch (*Larix leptolepis* Gordon). Plant Science 169: 901–907.

van Doorn WG, and Woltering EJ. 2005. Many ways to exit? Cell death categories in plants. Trends in Plant Science 10: 117–22.

Vanden Berghe T, Vanlangenakker N, Parthoens E, Deckers W, Devos M, Festjens N, Guerin CJ, Brunk UT, Declercq W, and Vandenabeele P. 2010. Necroptosis, necrosis and secondary necrosis converge on similar cellular disintegration features. Cell Death and Differentation 17: 922–930.

Vidaković M. 1991. *Conifers: morphology and variation*. Grafički zavod Hrvatske Zagreb 1 edition.

von Aderkas P, Klimaszewska K, and JM B. 1990. Diploid and haploid embryogenesis in *Larix leptolepsis, L. decidua*, and their reciprocal hybrids. Canadian Journal of Forest Research 20: 9–14.

von Aderkas P, Rouault G, Wagner R, Chiwocha S, and Roques A. 2005a. Multinucleate storage cells in Douglas-fir (*Pseudotsuga menziesii* (Mirbel) Franco) and the effect of seed parasitism by the chalcid *Megastigmus spermotrophus* Wachtl. Heredity 94: 616–622.

von Aderkas P, Rouault G, Wagner R, Rohr R, and Roques A. 2005b. Seed parasitism redirects ovule development in Douglas fir. Proceedings of the Royal Society B: Biological Sciences 272: 1491–1496.

Vuosku J, Sutela S, Tillman-Sutela E, Kauppi A, Jokela A, Sarjala T, and Häggman H. 2009. One tissue, two fates: different roles of megagametophyte cells during Scots pine embryogenesis. Plant Signaling & Behavior 4: 928–932.

Wagner RE, Mugnaini S, Sniezko R, Hardie D, Poulis B, Nepi M, Pacini E, and Von Aderkas P. 2007. Proteomic evaluation of gymnosperm pollination drop proteins indicates highly conserved and complex biological functions. Sexual Plant Reproduction 20: 181–189.

Wang X. 2001. The expanding role of mitochondria in apoptosis. Genes & Development 15: 2922–2933.

Wang Z, Gerstein M, and Snyder M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews Genetics 10: 57–63.

Wehmeyer N, and Vierling E. 2000. The expression of small heat shock proteins in seeds responds to discrete developmental signals and suggests a general protective role in desiccation tolerance. Plant Physiology 122: 1099–108.

Wertman J, Lord CE, Dauphinee AN, and Gunawardena AH. 2012. The pathway of cell dismantling during programmed cell death in lace plant (*Aponogeton madagascariensis*) leaves. BMC Plant Biology 12: 1–16.

Wilhelm BT, and Landry JR. 2009. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. Methods 48: 249–257.

Wilson NS, Dixit V, and Ashkenazi A. 2009. Death receptor signal transducers: nodes of coordination in immune signaling networks. Nature Immunology 10: 348–355.

Winkel-Shirley B. 2001. Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. Plant Physiology 126: 485–493.

Woo HR, Chung KM, Park JH, Oh Sa, Ahn T, Hong SH, Jang SK, and Nam HG. 2001. ORE9, an F-box protein that regulates leaf senescence in *Arabidopsis*. The Plant Cell 13: 1779–1790.

Worley CK, Zenser N, Ramos J, Rouse D, Leyser O, Theologis a, and Callis J. 2000. Degradation of Aux/IAA proteins is essential for normal auxin signalling. The Plant Journal 21: 553–562.

Wu CH. 2003. The Protein Information Resource. Nucleic Acids Research 31: 345–347.

Xiong H, Li Y, and Li L. 2006. A unique form of cell death occurring in meristematic root tips of completely submerged maize seedlings. Plant Science 171: 624–631.

Yamada K, Kanai M, Osakabe Y, Ohiraki H, Shinozaki K, and Yamaguchi-Shinozaki K. 2011. Monosaccharide absorption activity of arabidopsis roots depends on expression profiles of transporter genes under high salinity conditions. Journal of Biological Chemistry 286: 4357–43586.

Yamada T, Marubashi W, and Niwa M. 2000. Apoptotic cell death induces temperature-sensitive lethality in hybrid seedlings and calli derived from the cross of *Nicotiana suaveolens* $\times$ *N. tabacum*. Planta 211: 614–622.

Yang H, Matsubayashi Y, Nakamura K, and Sakagami Y. 2001. Diversity of *Arabidopsis* genes encoding precursors for phytosulfokine, a peptide growth factor. Plant Physiology 127: 842–851.

Yang J. 1997. Prevention of aApoptosis by Bcl-2: Release of cytochrome *c* from mitochondria blocked. Science 275: 1129–1132.

Yin Xm, Wang K, Gross A, Zhao Y, Zinkel S, Klocke B, Roth KA, and Korsmeyer SJ. 1999. Bid-deficient mice are resistant to Fas-induced hepatocellular apoptosis. Nature 400: 886–891.

Yorimitsu T, Nair U, Yang Z, and Klionsky DJ. 2006. Endoplasmic reticulum stress triggers autophagy. The Journal of Biological Chemistry 281: 30299–30304.

Yoshimoto K, Hanaoka H, Sato S, Kato T, Tabata S, Noda T, and Ohsumi Y. 2004. Processing of ATG8s, ubiquitin-like proteins, and their deconjugation by ATG4s are essential for plant autophagy. The Plant Cell 16: 2967–2983.

Young TE, and Gallie DR. 2000. Programmed cell death during endosperm development. Plant Molecular Biology 44: 283–301.

Zdolsek JM, and Svensson I. 1993. Effect of reactive oxygen species on lysosomal membrane integrity. Virchows Archiv B: Cell Pathology 64: 401–406.

Zerbino DR, and Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18: 821–829.

Zhang J, Wu T, Li L, Han S, Li X, Zhang S, and Qi L. 2012. Dynamic expression of small RNA populations in larch (*Larix leptolepis*). Planta 237: 89–101.

Zhang Y, Cao G, Qu LJ, and Gu H. 2009. Characterization of *Arabidopsis* MYB transcription factor gene AtMYB17 and its possible regulation by LEAFY and AGL15. Journal of Genetics and Genomics 36: 99–107.

Zheng Z, Nafisi M, Tam A, Li H, Crowell DN, narasimha Chary S, Schroeder JI, Shen J, and Yang Z. 2002. Plasma membrane-associated ROP10 small GTPase is a specific negative regulator of abscisic acid responses in *Arabidopsis*. The Plant Cell 14: 2787–2797.