In depth temporal transcriptome profiling reveals a crucial developmental switch with roles for RNA processing and organelle metabolism that are essential for germination in *Arabidopsis thaliana*

Reena Narsai+, Simon R Law+, Chris Carrie, Lin Xu and James Whelan

ARC Centre of Excellence in Plant Energy Biology, Bayliss Building M316 University of Western Australia, and Centre for Computational Systems Biology, 35 Stirling Highway, Crawley 6009, Western Australia, Australia.

+ = these authors contributed equally to this work

*Author for correspondence responsible for distribution of materials integral to the findings presented in this article is:

Reena Narsai  
Centre for Computational Systems Biology  
ARC Centre of Excellence in Plant Energy Biology  
Bayliss Building M316  
University of Western Australia,  
35 Stirling Highway,  
Crawley 6009,  
Western Australia, Australia  
Fax: 61-8-64884401  
Email: reena@cyllene.uwa.edu.au

**Running Title:** Arabidopsis germination transcriptome and organellar proteins
Abstract

Germination represents a rapid transition from dormancy to a high level of metabolic activity. In-depth transcriptomic profiling at 10 time points in Arabidopsis thaliana, including fresh seed, ripened seed, during stratification, germination and post-germination per se, revealed specific temporal expression patterns that have not previously been identified. Over 10,000 transcripts were differentially expressed during cold stratification, with sub-equal numbers up-regulated as down-regulated, revealing an active period in preparing seeds for germination, where transcription and RNA degradation both play important roles in regulating the molecular sequence of events. A previously unidentified transient expression pattern was observed for a group of genes, whereby a significant rise in expression was observed at the end of stratification and significantly lower expression observed 6 hours later. These genes were further defined as germination specific, as they were most highly expressed at this time in germination, in comparison to all developmental tissues in the AtGenExpress dataset. Functional analysis of these genes using genetic inactivation revealed that they displayed a significant enrichment for embryo defective or arrested phenotype. This group was enriched in genes encoding mitochondrial and nuclear RNA-processing proteins, including >45% of all pentatricopeptide domain-containing proteins expressed during germination. The presence of mitochondrial DNA replication factors (mTERF), and RNA-processing functions (PPR proteins) in this germination specific subset, represents the earliest events in organelle biogenesis, preceding any changes associated with energy metabolism. GFP analysis also confirmed organellar localisation for 65 proteins, largely showing germination specific expression. These results suggest that mitochondrial biogenesis involves a two-step process to produce energetically active organelles, an initial phase at the end of stratification involving mtDNA synthesis and RNA processing, and a later phase for building the better-known energetic functions. This also suggests that signals with a mitochondrial origin, retrograde signals, may be crucial for successful germination.
Introduction

Seeds represent a crucial stage in the plant life cycle as they are essential for the propagation of the species and allow dispersal to new locations. Seeds also allow plants to optimise survival strategy, as seeds display dormancy, which can be ‘broken’ via a variety of environmental factors, thus allowing plants to optimise growth with reference to environmental conditions. As the seed is often the primary product utilised by humans, it is not surprising that seed germination is an intensively studied topic. There are a plethora of excellent articles and reviews characterising seed dormancy and germination with loss-of-function and gain-of-function mutants analysed, across multiple species and utilising many “omics” technologies (Gallardo et al., 2001; Fu et al., 2005; Nakabayashi et al., 2005; Cadman et al., 2006; Holdsworth et al., 2008; Sreenivasulu et al., 2008; Howell et al., 2009). It has been observed that seed dormancy occurs in most plant species and provides seeds with a mechanism to survive extended periods of debilitating conditions prior to germination. In this way germination can occur under favourable conditions, ensuring the greatest chance at seedling establishment (Baskin and Baskin, 2004). A range of factors have been identified in relation to dormancy initiation, maintenance and alleviation; including temperature, moisture content, day length, light quality and mineral nutrition (Allen et al., 2007). These external triggers are perceived by the seed and elicit a series of signal-transduction pathways, leading to the modulation of the phytohormones abscisic acid (ABA) and gibberellic acid (GA) (Allen et al., 2007). Studies have shown that it is the antagonistic interaction between these two hormones that is at the core of dormancy maintenance (linked to ABA activity), dormancy release and germination initiation (linked to GA activity) in plants (Allen et al., 2007; Holdsworth et al., 2008; Sreenivasulu et al., 2008).

In order to dissect the molecular mechanisms that exist downstream of the hormonal signals regulating germination, transcriptomic and metabolomic profiling studies have been carried out in a range of species including Arabidopsis thaliana (Arabidopsis) (Gallardo et al., 2001; Fu et al., 2005; Nakabayashi et al., 2005; Fait et al., 2006), Hordeum vulgare (barley) (Sreenivasulu et al., 2008) and Oryza sativa (rice) (Howell et al., 2009). These studies have revealed some common properties of transcriptomic changes that occur across these plant species. Firstly, a large number of mRNA species (12,000-17,000) are present in the dry seeds or embryos, secondly there appears to be a tightly regulated, temporally controlled transition through germination characterised by phasic changes in transcript abundances (Nakabayashi et al., 2005; Sreenivasulu et al., 2008; Howell et al., 2009). For Arabidopsis, transcriptomic
changes were analysed with a focus on profiling stored mRNAs in dry seed, to better differentiate the transition from late embryogenesis to seed germination (Nakabayashi et al., 2005). This profiling of dry seed (0h) and seeds at 6 h, 12 h and 24 h post imbibition under continuous light, focussed on ABA regulation and revealed transcriptomic differences between germination in wild-type (WT) and abi5 mutants (Nakabayashi et al., 2005). However, the maturation of freshly harvested seeds (prior to dark desiccation) and the breaking of dormancy over stratification were not examined. Likewise in studies with rice and barley germination of ripened seed, either at 0, 1, 3, 12 and 24 h for rice (Howell et al., 2009) or 0 and 24 h (and 48 h and 72 h) for barley (Sreenivasulu et al., 2008), will not identify important steps from seed maturation (ripe seed) through desiccation and stratification. A recent systems level approach, analysing regulators of germination in flowering plants, utilised various publicly available transcriptome datasets and found that specific co-ordinated transcriptional regulation occurs separating the transition from dormancy to germination in flowering plants and that dormancy may have evolved by adjusting existing cellular phase transition and abiotic stress response related genetic pathways (Bassel et al., 2011).

In Arabidopsis, dormancy is broken via the cold (4°C) imbibition of seeds in darkness. In most experiments involving the growth of Arabidopsis, seeds are typically sown and placed at 4°C in the dark for at least 48 h. This process is referred to as stratification and studies examining the germination rate of stratified and non-stratified seed alike reveal a significant increase in germination rates when seeds are subjected to stratification (Yamauchi et al., 2004; Dave et al., 2011). A role for GA induced signalling during stratification has been shown using 8K Arabidopsis microarrays (Yamauchi et al., 2004). While the essential role for GA has been demonstrated for dormancy release, promoting embryonic expansion, inducing mobilisation of storage reserves and mediating the weakening of tissues that envelope the embryo (Brady and McCourt, 2003; Yamauchi et al., 2004; Feurtado and Kermode, 2007; Holdsworth et al., 2008), the detailed molecular networks and signalling that results in this increased germination rate and synchrony during and after stratification, are not yet elucidated. While the identity of some components responsive to these hormonal cues have been identified, such as the role of Della proteins in maintaining dormancy under ABA control (Tyler et al., 2004; Cao et al., 2006; Dohmann et al., 2010), a detailed molecular sequence of events, underpinning the transition from a dormant seed to the young seedling is lacking; especially with respect to temporal resolution.
Successful germination requires the mobilisation of energy reserves to power germination until photosynthesis is established. Studies of mitochondrial biogenesis during germination in rice revealed that poorly differentiated mitochondria, lacking cristae and matrix structure were present in dry seed, and that the peak in transcript abundance for components encoding the machinery of oxidative phosphorylation was 24 h after imbibition (Howell et al., 2006). In maize a similar study at a protein level showed that respiratory chain components did not peak until 48 h post-imbibition (Logan and Leaver, 2000). However, other measures such as analysis of cristae structure, increase in respiration and ability to import proteins reveal that mitochondrial biogenesis and activity are activated earlier than the peak in transcript abundance encoding components involved in oxidative phosphorylation (Howell et al., 2006; Howell et al., 2007). In addition, protein import complexes were able to import protein just 30 minutes after imbibition (Howell et al., 2006; Howell et al., 2007). Complementing this, a transcriptomic study analysing germination in rice revealed a surge in transcript abundance for genes encoding transport functions at 3 hours after imbibition (Howell et al., 2009). This suggests that signals (and responses) affecting mitochondrial function are taking place earlier in germination.

To gain a greater temporal dissection of the processes occurring during germination in Arabidopsis, a detailed time course providing an expansive view of the process; before and after seed desiccation, over the course of stratification, to germination and post-germination was analysed. Analysis of this extensive time course enabled the identification of novel, stage specific and transient patterns of expression. The identification of these tight expression patterns provided the basis for determining the relationship between co-expression, co-localisation and function of encoded proteins. Functional analysis also revealed a link between function and the pattern of gene expression. GFP tagging was carried out to verify organellar localisation for a large number of proteins, many of which were annotated as having “unknown functions”. This in depth temporal analysis revealed a transient peak in expression for transcripts associated with ethylene metabolism, novel organelle proteins and a role for RNA processing and mRNA decay at the earliest stages of germination in Arabidopsis. The greater dissection of these processes at a temporal level uncovers processes that have gone unnoticed and thus represent a mechanistic gap in the understanding of transition from dormancy to germination.
Results

Overview of transcriptomic changes – from seed to stratification and germination

In order to gain a comprehensive insight into Arabidopsis germination, 10 time points were selected including freshly harvested Col-0 seed (before desiccation, directly upon removal from silique – “H”), seed desiccated for 15 days in darkness (0 h), seeds stratified at 4 °C in the dark for; 1 h (1 h S), 12 h (12 h S) and 48 h (48 h S). Stratified seeds were then transferred into continuous light and further collected at 1, 6, 12, 24 and 48 h into the light (1 h SL, 6 h SL, 12 h SL, 24 h SL and 48 h SL respectively). As germination is generally defined as concluding when a part of the embryo emerges from the testa (generally around 24 h after imbibition in the light), the final time point (48 h SL) is considered a post-germination time point. During the time course analysed, 15,789 genes were found to be expressed at one or more time points with more than 95% of these genes significantly up/down regulated (following false discovery rate (FDR) correction) during this crucial developmental stage, reflecting the extensive regulation occurring at the transcript level (Supplementary Table 1). Whilst a previous study has examined the transcriptomic responses for less than 8,000 genes during stratification (Yamauchi et al., 2004), to date, there has been no global (22,000 genes) transcriptomic analysis carried out during stratification, which may reflect the general assumption that few significant processes apart from an increase in water content occurs during stratification.

The inclusion of three time points during stratification revealed that greater than 10,000 genes are differentially expressed over the 48 h during stratification (S), with the greatest number of differentially expressed genes (DEGs) changing in transcript abundance between 12 h S and 48 h S (Supplementary Table 1, Figure 1A). Notably, many of the changes observed were increases in transcript abundance, with a total of 7,517 unique transcripts increasing in transcript abundance over 48 h of stratification compared to 9,801 transcripts up-regulated during the first 48 h after transfer to light at 22 °C (Figure 1A). Thus, the observed differential expression is not only a reflection of the clearing out of stored transcripts upon imbibition, but also the transcriptional up-regulation occurring as part of the germination process, which is supported by the observed functional categorisation of the proteins encoded by these transcripts. The functional categories over-represented in the earliest sub-sets of up-regulated DEGs during stratification included the oxidative pentose phosphate pathway (OPP) and nitrogen and hormone metabolism (1 h S v 12 h S; Figure 1B). Upon closer examination, it is seen that genes encoding proteins involved in ethylene signalling are first induced in the early
hours of stratification (1 h S v 12 h S; Supplementary Table 2). This is in agreement with the role of ethylene, which has been shown to augment germination completion (Kecpczyński and Kecpczyńska, 1997; Beaudoin et al., 2000). The largest number of DEGs were observed between 12 h S and 48 h S, including an induction and over-representation of genes encoding nucleotide metabolism, RNA processing and protein synthesis functions (12 h S v 24 h S; Figure 1B). Closer analysis of the subsets in the protein category (12 h S v 48 h S; Figure 1B) revealed an induction of genes encoding cytoplasmic and organelar ribosomal proteins (Supplementary Table 2). In contrast, protein modification and degradation functions were under-represented in these subsets (12 h S v 24 h S; Figure 1B).

Once stratified, seeds were transferred into continuous light, and an induction and over-representation of genes encoding photosynthesis related functions as well as lipid, hormone and secondary metabolism functions were observed (Figure 1). Interestingly, several RNA and protein related functions were seen to be significantly under-represented in the subsets of DEGs up-regulated between 1 h SL and 48 h SL, possibly due to the earlier induction of these genes during stratification (Figure 1B). As may be expected after transfer into light, an over-representation of lipid metabolism and developmental function-related genes; particularly storage proteins and late embryogenesis abundant proteins were observed in the down-regulated subsets of DEGs (1 SL v 6 SL; Figure 1A and B; Supplementary Table 2). Closer examination of these reveals an initial down-regulation of genes encoding lipid transfer functions between 0 h and 1 h SL, followed by a repeated over-representation of tri-acyl-glycerol (TAG) synthesis functions in the genes down-regulated between 1 h SL and 48 h SL (Figure 1; Supplementary Table 2). These changes complement the known breakdown of oil storage reserves that occurs during germination in oil seeds (Graham, 2008). Thus, there appears to be distinct processes occurring between stratification (0 h to 48 S) and germination (48 S to 48 SL).

Correlation between co-expression, localisation and function, in transcripts and proteins

To visualise the expression profiles across stratification and germination (as defined in Figure 2A), the normalised expression values for all 15,789 genes expressed over germination were made relative to the maximum expression over the time course and hierarchically clustered (materials and methods), revealing 4 distinct clusters (Figure 2B). Cluster 1 in Figure 2B represents ~30% of all genes expressed during germination and is characterised by low expression from dry seed to stratification and even up to 6 h SL, followed by significant up-regulation after 6 h SL (Figure 2B). As expected, GO over-representation analysis (materials
and methods) showed that this cluster was enriched in several GO categories including genes encoding proteins targeted to the plastid, chloroplast and ribosome (Figure 2C). Genes encoding proteins with transferase, hydrolase and transporter functions as well as structural molecular activity were also enriched in this cluster (Cluster 1; Figure 2C), corresponding with the increase in energy demand and significant morphological changes that are observed after 12 h SL, as the seed transcends to a seedling (Figure 2A). The morphological changes observed during germination in this study (Figure 2A), comply with previous observations during germination (Fu et al., 2005). Furthermore, this up-regulated expression pattern can be readily observed in the previous Arabidopsis germination study that analysed global transcriptomic changes without stratification of seeds, (Cluster 1 - Supplementary Figure 1; (Nakabayashi et al., 2005)) and is also evidenced during germination in other species, such as rice and barley (Sreenivasulu et al., 2008; Howell et al., 2009). In contrast, Cluster 2 represents the stored mRNAs present in dry seed that remain at a high level of expression up to 12 h S followed by a distinct down-regulation (Figure 2C). Genes in this cluster represents 27% of the total genes expressed during germination and were enriched in nuclear targeted proteins including those showing transcription factor activity (Cluster 2, Figure 2C). Given that germination can still occur in the absence of transcription (Rajjou et al., 2004), the presence of these transcripts in dry seed, early in the time course likely represents the crucial genes necessary for immediate response to imbibition (Figure 2C). Again, the presence of these stored transcripts has also been observed in rice and barley (Sreenivasulu et al., 2008; Howell et al., 2009).

By analysing this extensive time course, it was revealed that a group of genes (in Cluster 3) are expressed at a very low level up to 12 h S, then dramatically increase in abundance between 48 h S and 6 h SL, before decreasing back to significantly lower levels by 48 h SL, the majority decreasing by 12 h SL (Figure 2B). This transient expression pattern for a significant number of genes is not apparent during germination when seeds are not stratified (Supplementary Figure 1; (Nakabayashi et al., 2005)). Examination of over/under-represented GO categories for these transiently expressed genes revealed significant over-representation of genes encoding proteins targeted to the mitochondria, nucleus and ribosomal proteins, corresponding with the observed over-representation of DNA, RNA binding and transcription functions (red boxes; Figure 2C). The over-representation of transcription factor functions in Cluster 2, and DNA, RNA binding functions in Cluster 3, suggests a two-step regulation of genes encoding these regulatory factors, with the genes in Cluster 2 likely encoding the regulatory proteins responsive to imbibition and required for the early stages of germination,
while genes in Cluster 3 are likely to encode the regulatory proteins required for normal germination progression and later plant development. In contrast, the genes in Cluster 4 were enriched in proteins of unknown cellular localisation and biological processes and slowly increased, decreased or largely remained unchanging in abundance across the germination time course (Figure 2B and C). This set of genes may represent genes required for basic cellular functions, not highly responsive under germination conditions.

Previous studies have shown that whilst germination progresses to the point of radicle emergence in the absence of transcription, seedling establishment is prevented (Rajjou et al., 2004). Therefore, during this crucial developmental process, the translation of transcripts expressed during germination is clearly essential for further development. In order to identify any correlation between transcript and protein abundance, the transcript abundance profiles in this study were compared to previous studies examining protein abundance during germination (Gallardo et al., 2001, 2002; Fu et al., 2005; Chibani et al., 2006). Specifically, one study identified over 400 proteins expressed in dry seeds; seeds after 3 days of cold stratification and then 30 h, 48 h, 72 h and 96 h into a 16/8 light/dark cycle (Fu et al., 2005). This study categorised a large number of the protein abundance profiles as present = 1/absent = 0 for each time point, e.g. if a protein was present in dry seed and after stratification, but not following transfer into the light/dark cycle, it was categorised as 11000 (Fu et al., 2005). In this way, present/absent profiles were observed for 117 unique proteins over this time course (Fu et al., 2005), for which parallel expression information is also available in the present study (Supplementary Table 5). A number of the proteins identified by Fu and colleagues have also previously been identified in other germination studies and these have been annotated in Supplementary Table 5 (Gallardo et al., 2001; Fu et al., 2005; Chibani et al., 2006). Using this present/absent (1/0) profiling of protein abundance, these proteins were matched to the corresponding transcript profiles from this study (Figure 2D). Remarkably, it was seen that 81% of the up-regulated proteins over time showed comparable transcript expression profiles (16% slowly up in Cluster 4 + 65% up in Cluster 1; Figure 2D; Supplementary Table 5). Similarly, of the 20 proteins highly expressed in seeds and not detected following stratification or transfer into the light/dark cycle, 15 proteins showed similar transcript expression profiles i.e. transcripts in Cluster 2; highly expressed in freshly harvested and dry seed and decreasing in abundance over the germination time course (Figure 2D). Interestingly, 3 genes encoding DNA/RNA binding functions including; a glycine rich protein 7 - AT2G21660, a heterogeneous nuclear ribonucleoprotein - AT4G14300 and a I sy1-like splicing domain containing protein -
AT3G18790; were seen to have both a transient transcript expression profile (i.e. were in Cluster 3 in this study) and also have a transient protein abundance profile ((Fu et al., 2005); Supplementary Table 5).

Identification of a set of germination specific (GS) genes

To further analyse the genes that showed a transient expression during germination (Cluster 3; Figure 2B), publically available microarrays carried out on a wide variety of tissues in the AtGenExpress developmental set (Supplementary Table 3; (Schmid et al., 2005)) were downloaded, normalised and analysed together with the 30 arrays in this study (materials and methods). Expression levels of the 15,789 genes expressed during germination and development were visualised in the same row (and cluster) order as shown in Figure 2B and were hierarchically clustered by tissue samples (columns) to determine in what other tissues/organs these genes were also expressed (Supplementary Figure 2). Examination of these revealed that the genes in Cluster 1 and 4 (as shown in Figure 2B) consist of genes that are highly expressed or unchanging in expression in most other developmental tissues as well as late germination (Supplementary Figure 2). As expected the genes in Cluster 2 (Figure 2B) that were highly expressed in dry seeds were also highly expressed across the microarrays analysing developing seeds (Supplementary Figure 2). In contrast, it was observed that the genes in Cluster 3 were most highly expressed between 48 h S and 12 h SL, even in comparison to all other developmental tissues (blue box, Supplementary Figure 2). To filter these genes further for primarily germination specific (GS) expression, only those genes that had a relative expression level greater than 0.5 (relative to maximum expression level) between 1 h S and 24 h SL across all tissues were visualised (as these time points strictly represent germination). In this way, 775 unique genes were identified as showing the highest expression during germination (i.e. germination specific; Figure 3A).

Intriguingly, analysis of over/under-represented GO categories for these 775 germination specific genes revealed approximately double the expected percentage of genes in this set encoding proteins targeted to the mitochondria, nucleus and ribosomes; corresponding with the observed over-representation of DNA and RNA binding functions (Figure 3B). Genes encoding transcription factors (TFs) represented only 53 of the 775 genes, which was not significantly greater/less than the expected percentage in the genome. This indicates that the genes encoding mitochondrial and nuclear localised proteins were binding DNA or RNA, but performing functions other than transcriptional regulation. To discover the nature of these other
encoded protein functions, the genes in these mitochondrial and nuclear sub-sets from the 775 GS genes were viewed based on the sub-functional groups within these sets revealing a significant enrichment of RNA processing functions in both sets (Mitochondrial and Nuclear; Figure 3Ci and ii). Additionally, helicase and ribonucleoprotein/RNA binding functions were also seen to be enriched in the nuclear set (Figure 3Cii). For the mitochondrial set from the 775 GS genes, protein fate functions were observed to be enriched (18% vs. 8% in the whole mitochondrial set), despite the majority of genes in the mitochondria encoding metabolism and energy functions (which were under-represented in this set; Figure 3Cii). Closer examination of the genes encoding RNA processing functions revealed a significant (p<0.001) over-representation of pentatricopeptide repeat domain (PPR) containing genes in this germination specific subset, with nearly 10% (75 genes) of the 775 genes encoding a PPR domain containing protein, while PPR domain containing genes only make up less than 2% of all genes in the genome. To confirm the observed PPR gene expression pattern was limited to the transiently expressed genes during germination, the percentage of PPR genes in each cluster was examined (Figure 3Di). Interestingly, it was seen that 45% (137) of all PPRs were in Cluster 3, showing transient expression during germination and of these, 75 genes showed germination specific expression (Figure 3D).

Despite the finding that this transient expression pattern cannot be readily identified when seeds are not stratified (Supplementary Figure 1; (Nakabayashi et al., 2005)), when the 75 PPR genes showing germination specific expression were examined together with the microarrays from this study, it was observed that the transient expression pattern during germination was somewhat maintained (Figure 3Dii), however the temporal resolution was lost or not observed. This suggests that the stage specific expression of these genes is a characteristic of germination, independent of whether seeds undergo stratification (Figure 3Dii). To determine whether these transiently expressed genes also displayed transient expression during germination in other species, rice orthologues of the 775 Arabidopsis germination specific genes were visualised, showing their expression across germination and developmental tissues in rice (details of arrays used shown in Supplementary Table 4; Supplementary Figure 3A). 768 rice orthologues of the 775 Arabidopsis germination specific genes could be identified using InParanoid (version 7.0) (Remm et al., 2001). Only 383 (~50%) of the 768 rice orthologues displayed maximum expression during rice germination (yellow box; Supplementary Figure 3A; (Howell et al., 2009; Narsai et al., 2009), however, these did not display the transient pattern, as observed in Arabidopsis. Instead, the expression pattern observed for the rice
orthologues was more similar to the patterns observed for Cluster 1, 2 and 4 in Arabidopsis. As PPR genes display high levels of orthology between plant species (O'Toole et al., 2008), the subset of 68 rice PPR genes orthologous to the 75 germination specific genes in Arabidopsis (green box; Figure 3Di and ii) were isolated and the expression levels were hierarchically clustered (Figure 3Diii). Expression of the 68 orthologous genes in rice revealed no transient expression pattern (Figure 3Diii). In addition, visualisation of these 68 rice PPR genes across germination and other developmental tissues (Supplementary Figure 3B) further confirmed the divergence in the transcriptomic response of PPR gene expression between monocots and dicots, despite orthology.

*Loss-of-function of germination specific genes results in embryo lethality*

A recent study defined 481 genes as seed essential, where a loss of function mutation in these genes was found to result in a seed related phenotype, mostly embryo lethal. These genes are indicated in the SeedGenes database (Meinke et al., 2008). The list of SeedGenes characterised as showing a seed related phenotype (e.g. seed lethal) was matched against the genes expressed during germination. Of the 481 genes in this database (referred to as “seed-genes”), expression of 422 genes could be detected during this germination time course. A significant enrichment of seed-genes was seen in Cluster 1 (35% v 30% in the genome) and Cluster 3 (22% in Cluster 3 v 15% in the genome; Supplementary Figure 4) with 35 seed-genes observed in the set of 775 germination specific genes (Figure 3A). Interestingly, the rice orthologues for these 35 genes did not show a transient expression pattern during rice germination (Supplementary Figure 4B and C); suggesting that despite orthology, the controlled expression of these genes during germination may be specific to Arabidopsis germination.

Given the enrichment of seed-lethal genes in the 775 germination specific genes identified in Figure 3A, a search was carried out on the 775 to determine whether these genes encode crucial protein functions necessary for seed/seedling or even plant development. To do this, large scale reverse genetic studies identifying phenotypes for knocked-out/silenced genes were matched to the 775 germination specific genes. The studies/databases examined are outlined in Supplementary Table 6. In addition to the large-scale studies/databases, it was observed that 110 of the 775 germination specific genes encode proteins experimentally and/or predicted to be localised to the mitochondria (Figure 2B), therefore all genes encoding these proteins were individually searched for known phenotypes in previous publications. In this way, 114 genes were identified as having altered developmental phenotypes (Figure 4A). In order to
ensure that this process of searching for phenotypes did not have any particular bias, sets of 775 randomly selected genes were generated and examined for phenotypes exactly as carried out for the 775 germination specific set, i.e. all studies/databases in Supplementary Table 6, and individual searching for genes encoding mitochondrial proteins. It was seen that overall, the 775 germination specific set consisted of significantly more genes with known phenotypes compared to the average number across the random gene-sets (p<0.01; 114 genes vs 79 genes expected). Moreover, the types of phenotypes also significantly differed, with more than 3 times the number of genes with seed-lethal/embryo arrested phenotypes seen in the germination specific set (51 genes) compared to the 775 random gene-sets (Figure 4A and B; (Pagnussat et al., 2005; Meinke et al., 2008)). Interestingly, closer examination of the 114 genes with known phenotypes also reveals an obvious enrichment of 40 genes encoding RNA binding/processing functions, compared to only 16 in the random gene-sets (Figure 4A and B). Examples of these genes encoding proteins involved in RNA binding functions included; RNA helicases (e.g. At5g08610), RNA binding proteins (e.g. At4g32720; At1g49400), various ribonucleases (e.g. At1g01040, At2g17510) and proteins containing an RNA recognition motif (e.g. At4g24280; Supplementary Table 6). The identification of these functions as being most highly expressed during germination (Figure 3A) combined with observations that the silencing/loss-of-function of a significant number of these genes results in seed lethal/embryo arrested phenotypes (Figure 4; Supplementary Table 6) reveals the crucial requirement for the expression and function of RNA binding/processing during early germination and development in Arabidopsis.

Confirming the organellar location of proteins with germination specific expression

The transcriptomic results presented strongly suggest a clear link between localisation and co-expression, with genes encoding proteins annotated as plastid/chloroplast localised, being over-represented in Cluster 1 (Figure 2B), whilst genes encoding mitochondrial/nuclear localised proteins being over-represented in Cluster 3 (Figure 2B). Considering these correlations, it was hypothesized that a selection of genes, with hitherto unconfirmed protein localisations, could encode proteins localised to the mitochondria, plastids and/or peroxisomes, based on their expression patterns and predicted localisations as annotated in Arabidopsis SUBA localisation database (Heazlewood et al., 2007). A range of 65 genes (Supplementary Table 7), largely showing germination specific expression (as in Figure 3A) were analysed by GFP targeting to determine protein localisation. Fusion proteins were constructed and transiently transformed into Arabidopsis cell culture using biolistic transformation. Organelle targeting was verified using AOX-RFP as a mitochondrial control, SSU-RFP as a plastid control
and RFP-SRL as a control for peroxisomal targeting (materials and methods). Protein accumulation was characterised as either to the mitochondria, plastid, dual-targeted to the mitochondria and the plastid, peroxisome, cytoplasm, endoplasmic reticulum (ER), golgi or the nucleus. Examples of fluorescent micrographs for proteins targeted to the mitochondria, plastids and peroxisomes are shown in Figure 4A, and the complete set of targeting results shown in Supplementary Figure 5.

Most of the genes for which localisation was determined exhibited low expression in dry seed and 48 h SL i.e. were highly expressed specifically during germination (genes with germination specific expression are indicated by `; Figure 5B). Analysis of the localisation of these confirmed the predicted localisation for most genes, with some exceptions including a mitochondrial transcription termination factor (mTERF; At5g06810) and a PPR containing protein At4g21170 that were predicted to be mitochondrial but appear to be dual-targeted to both mitochondria and plastids (Figure 5B; Supplementary Figure 5). It can be seen that several genes displayed in Figure 5B encode DNA/RNA binding functions including a mitochondrial intron maturase, an mTERF protein, many mitochondrial PPR encoding proteins as well as a plastid targeted chaperone DNAJ protein, PHD finger DNA binding transcription factor and a tRNA methyltransferase (Figure 5B). The transient expression of these genes encoding regulatory functions, during germination and the confirmation of their mitochondrial and plastid localisation reveals a specific step of transcriptional regulation that occurs for organellar proteins during this crucial stage of development. Further supporting the essential role of these genes during the earliest stages of germination, it was revealed that 3 of the 65 genes selected for localisation studies encode proteins for which a loss-of-function results in a seed lethal phenotype (denoted (ED); as determined in (Ding et al., 2006; Meinke et al., 2008; Hammani et al., 2011). Each of these genes; At1g79490 (Figure 5A), At1g10270, At5g60960 and At5g39710 encoded proteins predicted to be mitochondrial, which was confirmed by GFP analysis (Figure 5B; Supplementary Figure 5). Notably, a number of recent studies have also confirmed the observed localisation for 4 of these proteins - AT5G46920 (Keren et al., 2009), AT1G80270 (Doniwa et al., 2010), AT5G60960 (Hammani et al., 2011), AT4G36040 (Chen et al., 2010).

Furthermore, including the 2 seed-lethal genes annotated as “Unknown function”, it can be seen that localisation was determined for 13 other genes of unknown function (Figure 5B). These analyses reveal clues about the possible function of these genes, inferred from their organellar localisation and germination specific expression; forming the basis of further functional studies of these genes (Figure 5B). In addition, it was observed that 2 small auxin
responsive RNA-like (SAUR) encoding genes were predicted to encode mitochondrial targeted proteins, however the localisation of these were determined to be nuclear/cytoplasmic (Figure 5B). The transient expression of these was particularly interesting as it is known that these genes are specifically regulated at the level of mRNA decay, allowing tight control of mRNA levels (Newman et al., 1993). This sub-set of tightly controlled, transiently expressed nuclear, mitochondrial and/or plastid targeted genes (Figure 5B) may represent crucial control factors responsible for normal regulation of gene expression during germination.

**Factors affecting transcript abundance during Arabidopsis germination**

It is known that upon imbibition during germination, stored mRNAs are degraded as in vivo transcription begins, this pattern of decrease in abundance of stored transcripts appears to be conserved; independently of whether seeds are stratified or not (Cluster 2; Figure 6Aii). Similarly, the up-regulation of specific transcripts over the germination time course is also conserved (Cluster 1; Figure 6Ai). Although these expression patterns are relatively conserved during germination with/without stratification, the temporal development sequence is different due to different experimental designs. Two distinct phases of RNA degradation were evidenced i.e. Cluster 2 and 3. A comparison of these genes with the mRNA half-lives of ~13,000 genes that have been previously determined in Arabidopsis (Narsai et al., 2007), allows insight into the regulatory processes that may affect transcript stability or degradation. Given that transcripts encoding core cellular functions, such as those involved in energy e.g. photosynthesis; have relatively long mRNA half-lives (Narsai et al., 2007), it was not surprising to observe a significant (p<0.05) enrichment of transcripts with longer half-lives in Cluster 1 (Figure 6B). Similarly, given the sharp decrease that occurs after the transient expression seen in Cluster 3, it was also as expected to find this cluster was enriched in transcripts with relatively short half-lives (<6 h) (Figure 6B). In contrast, it was surprising to see that there was not an enrichment of transcripts with shorter half-lives in Cluster 2 (Figure 6B), given that this group is characterised by a significant decrease in transcript abundance (Cluster 2; Figure 6A and B). Although this decrease is seen to occur much slower for stratified seeds (this study), it does appears that when seeds are imbibed under continuous light, with no stratification (Nakabayashi et al., 2005), these stored transcripts decrease to about 50% of their dry seed levels within 6 h, indicating a relatively rapid rate of decrease in abundance (Cluster 1; Figure 6A). These findings indicate that the rate of degradation is controlled in a developmental stage specific manner (Figure 6A and B).
The other factor controlling transcript abundance is transcription; therefore, without directly measuring transcription, focus was shifted to how the transcripts encoding these regulatory factors are responding during germination. Thus, the Arabidopsis transcription factor (TF) database was queried (Riano-Pachon et al., 2007), resulting in the observation that Cluster 2 was significantly enriched in genes encoding TFs (35% of all TFs are in Cluster 2; Figure 6Ci). This is particularly interesting as it further supports the important role of mRNA stability during germination, given that genes encoding TFs are known to have short mRNA half-lives (Narsai et al., 2007) and yet are seen to remain stably abundant up to 12 h S (Figure 6A). A closer look at the families of TFs in each Cluster reveals specific over/under-representation patterns, including the significant over-representation of the MYB, LOB, Orphans, AUX/IAA and bHLH families in Cluster 1 (Figure 6Cii). Interestingly, the AUX/IAA and bHLH families were also seen to be over-represented in the group of genes showing the same up-regulation pattern over germination in rice (black boxes; Figure 6Cii) (Howell et al., 2009). Several studies have examined the role of AUX/IAA proteins and have revealed that these proteins have repressor functions crucial for normal development (Ulmasov et al., 1997; Tiwari et al., 2004). Given their conserved expression pattern, it is possible that their role in regulating germination and development is conserved across monocots and dicots. Similarly, Cluster 2 was significantly enriched in the HB, Trihelix and HSF families, with the latter also seen to be over-represented in the parallel cluster during rice germination (Figure 6Cii) (Howell et al., 2009). Previous studies have implicated a role for HSFs for normal development in plants and other eukaryotic species (Kotak et al., 2007), with one study even showing a build-up of both transcript and protein abundance for HSFs over the course of seed development (Kotak et al., 2007). Therefore, it was not surprising to see the conserved expression pattern of these genes across rice and Arabidopsis (Figure 6Cii).

Although Cluster 3 was not enriched in TFs, it was seen that of the TFs present, there was an enrichment of C3H, HB, C2C2 GATA and the Mitochondrial transcription TERmination Factors (mTERFs); with 15 of the 23 mTERFs expressed during germination, present in Cluster 3 (Figure 6Cii). It has been shown that HB TFs have a role in cell differentiation and growth, thus their over-representation in Cluster 3 may reflect an increase in demand for TFs controlling these vital processes during germination (Kappen, 2000). A previous study has shown the crucial role these factors play, not only in mitochondrial transcription termination but also in the initiation of transcription and the control of mtDNA replication (Roberti et al., 2009). The role for CH3 was also characterised in Arabidopsis embryos and were seen to be expressed from
globular to late cotyledon stage (Li and Thomas, 1998). Thus, it was not surprising to see the over-representation of these in Cluster 3 (Figure 6Cii).

In order to analyse putative cis-elements that may be involved in regulating transcript abundance during germination, genes in each cluster and the 775 GS set identified in Figure 3 were examined for over-represented putative 6-mers in the 1kb upstream region of the transcriptional start site. The 775 GS set was specifically chosen as the abundance of these transcripts increases and decreases in a relatively short time period, and thus they are likely to be co-regulated at a transcriptional level; accounting for the observed increase and possibly be actively degraded; to produce the decrease observed. Overall, the 1kb upstream promoter elements displayed significant enrichment of 6-mers, with some elements found in up to 49% of the genes that showed germination specific expression (Table 1). All promoter elements were matched against known cis-element binding sites within the AGRIS database (Davuluri et al., 2003; Palaniswamy et al., 2006; Yilmaz et al., 2011) and studies that have characterised specific binding sites (Kosugi et al., 1995; Schoffl et al., 1998). Any elements showing a significant over-representation (p<0.05) for genes in Clusters 1-4 and the GS subset are shown in Table 1. Promoter analysis of stored mRNAs was carried out previously and a significant over-representation of ABRE binding sites were observed for these (Nakabayashi et al., 2005); a feature confirmed in this study, with ABRE elements also observed in genes of Cluster 2 (Figure 6A; Table 1). Interestingly, there were an overwhelmingly large number of known motifs seen in the GS subset, suggesting numerous factors may be involved in the controlled expression pattern observed for these genes. An example of these included an over-representation of Telobox and Site II elements for the genes in Cluster 3 and the GS subset (Table 1), complying with the role of these genes in the control of genes encoding organellar proteins during development and the circadian regulation of genes encoding mitochondrial proteins (Giraud et al., 2010). Additionally, it can be seen that genes encoding HSF transcription factors were enriched in Cluster 2 (Figure 6C). It is possible that these HSFs have a role in the control of genes in the GS subset, as several heat shock element (HSE) binding site motifs were seen to be enriched in the GS subset (Table 1), implying that the stored HSF encoding transcripts in dry seed (Cluster 2) may be translated upon imbibition and have a downstream role in the control of GS transcript abundance.

Discussion
Profiling analysis has been carried out at different levels during germination in Arabidopsis; from transcriptomic (Nakabayashi et al., 2005) to proteomic (Gallardo et al., 2001; Rajjou et al., 2004; Fu et al., 2005; Chibani et al., 2006) and metabolite analysis (Fait et al., 2006). These analyses have contributed to the understanding of the various processes that occur during germination. This study set out to observe if a greater temporal dissection of germination in Arabidopsis would reveal additional time or stage-specific molecular processes that enable the transition from dormancy to active metabolism and also gain a deeper understanding of the role of stratification in the process of germination. While the above studies have uncovered many novel aspects of germination and given insights into various regulatory processes, the results in this study revealed specific features of germination that have previously gone undetected. Specifically, the number of changes in transcript abundance during 48 h of stratification almost equalled that observed during 48 h of germination in continuous light, also the identification of a specific set of genes that appear to be predominantly expressed during the earliest stages of germination on the transition from stratification to germination, that are enriched in functions required for germination, and lastly, the environmental effects on the rate of RNA degradation during germination.

Stratification specific regulation during germination
The process of germination is strongly affected by environmental cues, which can significantly affect the rate and success of germination. Exposing seeds to cold temperatures in order to assist breaking dormancy, or seed stratification, is an established practice that is routinely utilised to maximise germination potential (Russell et al., 2000; Yamauchi et al., 2004; Dave et al., 2011). However, the molecular mechanisms underpinning the role of seed stratification have not been explored in-depth, with only one previous study examining the effect of stratification on the transcriptome, utilising 8 k DNA microarrays and focussing on the role of GA during stratification (Yamauchi et al., 2004). Thus, the present study, utilising the 22k Arabidopsis genome microarrays represents the most in-depth transcriptomic analysis during stratification, to date. The first overall effects of stratification is evidenced in the number of DEGs, where there was in fact a greater rate of induction of genes upon exposure to light (Cluster 1; Figure 5A) and a significantly slower rate of decrease in abundance for stored mRNAs (Cluster 2; Figure 5A), for seeds that have been stratified vs. not stratified, supporting the positive effect of stratification on germination rate. Also, comparing the DEGs in this study (including stratification) to germination without stratification (Nakabayashi et al., 2005) reveals that the up-regulation of transcripts encoding protein synthesis and hormone metabolism functions occurs between 12 h
In this study, revealing for the first time, that the up-regulation of these genes occurs almost exclusively during stratification in the germination process (Figure 1; Supplementary Figure 6). It has been shown that although germination can succeed when transcription is inhibited, seedling establishment is prevented (Rajjou et al., 2004) revealing a crucial role for transcription post-germination. Taken together with the finding that germination is prevented when translation is inhibited (Rajjou et al., 2004), it is likely that these genes encoding protein synthesis functions represent the next wave of translational machinery required for germination progression.

In addition to protein synthesis functions, hormone metabolism was also seen to be affected during stratification (Figure 1; Supplementary Figure 6). Specifically, a significant up-regulation of genes encoding Ethylene Response Factor family members (ERFs; e.g. ERF1-AT3G23240, ERF2-AT5G47220, ERF5; AT5G47230) was observed within the first 12 h of stratification (Figure 1; Supplementary Table 2). Notably, this early up-regulation of ERFs cannot be distinguished from genes highly expressed in the dry seed or other hormone responsive genes when seeds are not stratified (Supplementary Figure 6A v B; Supplementary Figure 7; (Nakabayashi et al., 2005)). It has been shown that ethylene counteracts the inhibitory actions of ABA (Kecpczyński and Kecpczyńska, 1997; Beaudoin et al., 2000; Linkies et al., 2009), and promotes endosperm cap weakening and endosperm rupture in *Lepidium sativum* and Arabidopsis (Linkies et al., 2009). Thus, the rapid induction of these genes during stratification, as revealed in the present study, not only confirms these roles for ethylene, but reveals that this regulation is possibly one of the earliest contributing factors to the greater germination rates and synchrony generally observed after stratification. Taken together, the findings in this study provided novel insight into why stratification can lead to greater germination rates, i.e. hormonal signalling is activated at this stage before the other changes occur that drive germination and growth. In contrast, when seeds are not stratified, these changes occur simultaneously and thus are not as efficient at ensuring successful germination. This demonstrates that the specific temporal sequence of events that occur during germination are important for developmental progression.

**Identification of a novel transient expression pattern during germination**

Upon examination of expression profiles during germination, it was revealed that ~14% of genes show transient expression, mostly peaking in abundance between 48 h S and 6 h SL (Cluster 3; Figure 2B). Extensive comparisons of these genes across other developmental
tissues then revealed a subset of these genes are specifically expressed during germination (GS set of 775 genes; Figure 3A). Given that this transient expression pattern is not observed for a significant number of genes during germination without stratification (Supplementary Figure 1), this suggests that during stratification, there is a time and temperature specific regulation that occurs, which allows these genes to peak in expression before the observed synchronous decrease in abundance occurs. Thus, this expression pattern for these genes cannot be identified in un-stratified seed as they are grouped with genes in Cluster 2, which decline in abundance during germination. Examination of the genes in this germination specific set reveals an enrichment of genes encoding RNA binding functions; nuclear and mitochondrial proteins, particularly those encoding PPR proteins (Figure 3B and C). This study revealed a tightly controlled stage of regulation for PPR encoding genes over the course of germination (Figure 3D). PPR proteins are defined by a repeating 35 amino acid motif, predicted to form an α-helix and to be targeted to mitochondria and plastids (Schmitz-Linneweber and Small, 2008). To date, PPR proteins have been shown to have roles in RNA splicing, cleavage, editing, stability and translation (Schmitz-Linneweber and Small, 2008). A recent study has even shown that AtPPR2 binds to 23S rRNA and plays a role in mitotic division and cell proliferation during embryogenesis (Lu et al., 2011). Another study showed that a point mutation in the PPR domain of a chloroplast PPR protein delayed chloroplast development (Cao et al., 2011). In addition to this, many of the genes identified as showing embryo lethal phenotypes under loss-of-function conditions (Meinke et al., 2008) encode PPRs, and a significant number of these showed this transient expression pattern in this study, suggesting a finely controlled, stage specific requirement for RNA processing during germination. Overall, a significant proportion of the genes defined as displaying a germination specific profile in this study were observed to display altered phenotypes associated with embryo development or male or female gametophyte development (Figure 4). This shows that the functions of many of the proteins, largely with RNA-binding functions, encoded by these genes are essential for germination, and thus gives insight into some of the earliest processes that occur during germination.

The role of RNA-binding proteins has been examined in a number of studies, from those specifically focused on plastid RNA-binding proteins (Wang et al., 2006), to analysis of families of proteins such as the glycine-rich RNA binding proteins (Kim et al., 2007; Kwak et al., 2011). Specifically, transcriptomic analysis revealed that two genes; cp29A and cp29B were highly expressed in germinating seedlings and therefore were selected for further analysis (Wang et al., 2006). Interestingly, a correlation between the transcript and proteins abundance for these genes was observed, and it was observed that new isoforms of these proteins were generated
following post-translational modification of these proteins during seedling development (Wang et al., 2006). Similarly, studies examining glycine-rich RNA binding proteins have also indicated a crucial role for these proteins during Arabidopsis germination and seedling development (Kim et al., 2007), as well as in the regulation of gene expression at the post-transcriptional level during abiotic stress (Schmidt et al., 2010). A recent study analysing the protein structure of GRP4 and GRP7 during cold acclimation revealed the crucial role of specific sequence domains necessary for correct RNA chaperone activity of these proteins (Kwak et al., 2011). Interestingly, GFP localisation in this study has revealed, for the first time, to our knowledge, that GRP4 is localised to the mitochondria (Figure 5). Also, notably, comparison of the protein abundance data during germination and the transcriptomic analysis from this study has also revealed that GRP7 is transiently expressed both at the transcript (this study) and protein levels (Fu et al., 2005) during germination and seedling establishment. Collectively, these findings suggest a finely-controlled, but crucial role of RNA-binding GRPs during germination.

A recent study used publicly available microarray data to examine the phase transitions from dormancy to germination and generated a condition-dependent network model of transcriptional interactions in Arabidopsis, called SeedNet (Bassel et al., 2011; http://vseed.nottingham.ac.uk). This network comprises of 8,261 nodes and demonstrates two major regions of interactions enriched in transcripts identified by Significance Analysis Microarrays; region 1 which is associated with non-germination, region 2, which represents a transition between non-germination and germination, and region 3, which is associated with germination (Bassel et al., 2011). Closer examination reveals that region 2 is significantly enriched in transcripts encoding RNA metabolism functions and represents a unique cluster of interactions that bridge these two regions, suggesting a mediating role for these genes in the transition from non-germinating to germinating states (Bassel et al., 2011). Interestingly, when the 775 transcripts comprising the germination specific set (identified in Figure 3A) from this study were queried in SeedNet, the majority localised to region 2 and region 3, with only a small proportion being identified in region 1. The large number of GS genes seen in region 2, together with the association of these genes with RNA metabolism, as well and the enrichment of RNA binding/processing functions in this suggest that the transient up-regulation of these gene represent a crucial regulatory switch necessary for germination progression.

In addition to RNA processing and PPR protein encoding genes, closer examination of the genes in the GS set (775 genes; Figure 3A) rapidly reveals other genes encoding
mitochondrial proteins that also show this tight, transient expression pattern. GFP localisation for a selection of the proteins encoded by these genes (Figure 4), confirmed their mitochondrial localisation and supported the co-expression and co-localisation pattern observed for many of these genes. The crucial role of mitochondria during germination has been examined before and it has been suggested that the biogenesis of new mitochondria is more important in oil seeds (Morohashi et al., 1981; Weitbrecht et al., 2011). Notably, this burst in the expression of genes encoding mitochondrial proteins does not represent the building of the bio-energetic functions required to power germination, but rather a specific phase of DNA replication, RNA synthesis and processing that occurs during germination. The genes encoding mitochondrial proteins that were most highly expressed (transiently) at this time included a significant percentage of genes encoding mTERFs (15 out of the 23 genes encoding mTERFs expressed during germination; Figure 5C) and PPR proteins (137 of the 309 genes encoding PPRs expressed during germination; Figure 3Di). It has been shown that mTERFs not only function in mitochondrial transcription termination but also in the initiation of transcription and the control of mtDNA replication (Roberti et al., 2009). Taken together, these findings reveal, a possible pin-point in time during germination when mitochondrial biogenesis is activated. This activation of mitochondrial DNA replication and transcription may play a crucial role in signalling the later expression of other bioenergetics components required for germination. It has been previously reported that a dual targeted protein that plays a role in translation in plastids and mitochondria is required for organelle retrograde signalling (Pesaresi et al., 2006). Thus, these finding during germination suggests the possibility that mitochondrial retrograde signalling may play a crucial role in the process of germination.

Interestingly, the transcriptomic analysis of rice germination also identified a subset of transiently expressed genes (Howell et al., 2009), however this occurred later in germination and the functions of genes in these sets differed. While the transiently expressed genes during rice germination were enriched in genes encoding transcription factors (Howell et al., 2009), genes transiently expressed in Arabidopsis germination were enriched in RNA processing functions (Figure 3A). However, it is possible that the controlled expression of these RNA processing functions also occurs in rice, but is missed due to the rice seed maturity, as rice seeds do not undergo stratification, thus dormancy is broken in an alternate manner. Another possibility is that these differences represent divergences between monocots and dicots or starch seeds and oil seeds. However, more intensive profiling studies in rice or other cereals; from fresh seed, during ripening and during dormancy alleviation may identify a similar process.
Crucial role of mRNA decay during germination

One of the first processes seen to occur upon imbibition is the clearing out of stored transcripts (Cluster 2; Figure 5A). For these, it was seen that even transcripts with mRNA half-lives longer than 6 h decrease to <50% of their starting abundance in under 6 h upon imbibition, in optimal light and temperature conditions (No stratification; Cluster 2; Figure 5Aii and Bii), while transcripts in this cluster decrease in abundance, at a much slower rate during stratification (this study; Cluster 2; Figure 5A). Finding the genes in Cluster 2 i.e. the stored mRNAs, to be enriched in transcription factors was somewhat surprising, given that transcripts encoding transcription factors have been shown to have short half-lives, allowing these mRNAs to be rapidly degraded (Narsai et al., 2007). Thus, these finding suggest that the stored seed transcripts are somehow highly stabilised in the dry seed and that controlled degradation of these specific transcripts occurs upon imbibition, even beginning during stratification. The importance of the role of mRNA decay during germination has been displayed in studies that have shown that normal germination and development suffers when mRNA degradation is affected (Delseny et al., 1977; Nishimura et al., 2005; Hirayama and Shinozaki, 2007). Given that the expression profiles compared for stratified vs. non-stratified seeds revealed differences in the rate of decrease for transcripts in Cluster 2, this indicates that the germination “clock” is highly controlled, with gene expression being tightly regulated and controlled by environmental conditions. A perfect example of this controlled regulation during germination is for genes in Cluster 3 (Figure 2B), which show a strong induction followed by an equivalently strong reduction of these genes, suggesting that tight regulation occurs, most likely occurs both at the transcriptional and degradation levels. A previous study has shown a role for controlled/active mRNA decay in yeast, in response to anaerobic conditions (Dagsgaard et al., 2001). Thus, the tightly controlled regulation of transcript abundance observed in Cluster 2 and 3, as well as the differences in the rates of decrease in abundance under stratified or un-stratified conditions, presents an argument suggesting a possible role for active mRNA decay that occurs in response to the imbibition or light during germination.

Conclusion

The greater temporal dissection of germination combined with functional analysis reveals molecular mechanisms occurring during germination that have previously gone undetected. Identification of these processes provides a molecular explanation for the greater rates of germination that occur during germination after stratification, and also provide greater
resolution of the processes that occur during germination. If such processes also occur in monocots, it may provide new targets to prevent precocious germination in cereals.
Materials and methods

Arabidopsis tissue collection and microarrays
In order to analyse a range of time points before and during Arabidopsis (Col-0) germination, 10 time points were analysed including freshly harvested seed (Har or H, which were collected from a single batch of WT plants that were that were exactly the same age), then the seeds following 15 days of ripening (0 h), and after; 1 h of stratification (1 h S), 12 h of stratification (12 h S), 48 h of stratification (48 h S), plates were then transferred to continuous light and collected 1 hour into the light (1 h SL), 6 hours into the light (6 h SL), 12 hours into the light (12 h SL), 24 hours into the light (24 h SL) and 48 hours into the light (48 h SL). 80 mg of WT Arabidopsis seeds were plated onto MS media (containing 3% sucrose) for each individual time point collection in the time course. Collections were repeated 3 times for 3 biological replicates.

RNA isolation, microarray and differential expression analyses
For all samples collected, the Ambion Plant RNA isolation aid and RNAqeous RNA isolation kit were used for effective isolation of RNA. 400 ng of total RNA was used as the starting amount of RNA for the ATH1 Arabidopsis genome expression array. Using the IVT express kit, microarrays were carried out according to manufacturers’ instructions. Before beginning the microarray experiments as well as during the course of carrying out the microarrays, the Agilent Bioanaylyser was used to ensure high quality starting RNA, generated aRNA and effective fragmentation, prior to hybridisation to the microarray. All raw intensity CEL files were imported into Avadis 4.3 (Strand Genomics, India) and the standard MAS5.0 normalisation was first carried out in order to determine present/absent/marginal calls for each probeset. All probesets that encoded hybridisation controls, bacterial genes and more than one single gene were excluded, leaving a global Arabidopsis expression set consisting of 15,789 probesets. Probesets that were called present in two or more replicates were considered to be expressed and was then used for further analysis. GC-RMA normalisation was carried out for all 30 microarrays and the resulting normalised intensities were used as the input for the differential expression analysis. This was carried out using the Cyber-T method, which implements a Bayesian method (Baldi and Long, 2001) for determination of probesets showing significant changes in transcript abundance. The PPDE method within Cyber-T was used for false discovery rate calculation (Choe et al., 2005). All input criteria were set according to Cyber-T recommendations applicable for each experimental set. A probeset was defined as significantly changing at p<0.05, with a PPDE of >0.96 (false discovery rate). These cut-offs and this
Bayesian method of differential expression has been verified and has been used in previous microarray studies (Narsai et al., 2010). In this way, step-wise differential expression analysis was carried out. All original microarray data files have been deposited to the Gene Expression Omnibus at NCBI; under accession GSE30223.

Hierarchical clustering and z-score analyses of GO categories

In order to view the profiles of expression changes over the time course, all GC-RMA was made relative to the maximum intensity over the time course. This data was then hierarchically clustered using average linkage (based on Euclidean distance) and 4 cluster profiles were drawn using Partek Genomics Suite (version 6.5). To determine if there was a statistically significant over or under-representation of a particular sub-category of gene ontology (i.e. GO cellular component, GO molecular function and GO biological process), a z-score analysis was carried out to compare the two proportions (subset vs. genome).

\[
 z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}
\]

A cumulative standard normal table was used to match the z-score and based on this, the P-values were determined.

PageMan analysis

In order to analyse the functional representation of the genes differentially expressed during the time course, differential PageMan analyses were carried out using the unique set of probesets representing the differentially expressed genes (Usadel et al., 2006). Fisher’s test for ORA (over-representation analysis) analysis (1.0 cut-off, Benjamini-Hochberg (Benjamini and Hochberg, 1995) FDR correction) was carried out in PageMan to determine statistically significant over/under representation of genes classified into specific BINS.

Publically available Arabidopsis and rice microarrays

To compile the publically available Affymetrix Arabidopsis and rice microarrays, all experiments containing CEL files were downloaded from the Gene Expression Omnibus within the National Centre for Biotechnology Information database or from the MIAME ArrayExpress database (http://www.ebi.ac.uk/arrayexpress/). The GSE or EXP numbers for the respective studies are shown in Supplementary table 3. The Arabidopsis AtGenExpress developmental dataset was
downloaded as CEL files (E-AFMX-9). These CEL files in addition to the 30 CEL files from this study were imported and quantile normalised together to enable comparability across these arrays using Partek Genomics Suite version 6.5 (St. Louis, Missouri, USA). To carry out parallel analysis for developmental conditions in rice, 41 developmental tissues for rice (Supplementary table 4) were also downloaded and analysed in the same manner as described for the Arabidopsis developmental set.

**Analysis of orthologues**
The InParanoid: Eukaryotic Ortholog Groups database (version 7.0; (Remm et al., 2001)) was utilised to analyse all orthologues between Arabidopsis and rice. The orthologous group file containing the *Oryza Sativa* vs. *Arabidopsis thaliana* set was downloaded for the whole-genome comparison. This generated information for orthologues identified by AGIs for Arabidopsis and TIGR identifiers for rice.

**Analysis of phenotypes for 775 GS genes and random gene-sets**
For this analysis, only the list of 21,192 Affymetrix probesets that matched to individual AGIs was isolated and used. From those, 3 sets of 775 randomly selected gene-sets were generated using Partek Genomic Suite version 6.5 (St. Louis, Missouri, USA). These sets were analysed and scrutinised for phenotypes in exactly the same manner as was carried out for the 775 GS set identified in Figure 3A. Firstly, large scale forward genetics studies were mined and matched to any genes in these sets, in addition, all genes encoding proteins localised to the mitochondria were individually searched for publications indicating phenotypes and these were collated (numbers for each set shown in Supplementary Table 6B). Full references for all phenotypes identified in the 775 GS set are shown in Supplementary Table 6C).

**Construction of GFP fusion proteins to confirm organelle targeting**
Given that the expression profiles showed correlation between co-expression and co-localisation e.g. the genes showing a transient expression pattern (Cluster 3) were enriched in genes encoding mitochondrial proteins, whilst Cluster 1 was enriched in genes encoding plastid proteins (Figure 2C), a selection of 65 genes, mostly encoding proteins predicted to be mitochondrial were selected for GFP localisation analysis. For each gene, the region of the protein encoding the targeting signals was cloned in frame with GFP by Gateway® cloning (Invitrogen), as previously described (Carrie et al 2009). For proteins predicted to target to the mitochondria or plastid, the first 100 amino acids (AA) were fused to N-terminus of GFP. For proteins predicted to target to
the peroxisome, the last 100 AA were fused to the C-terminus of GFP. Three RFP fusion proteins were utilised as controls for sub-cellular localisation. To control for mitochondrial targeting, the 42 amino acid mitochondrial targeting signal of alternative oxidase (AOX) was fused to RFP (Murcha et al., 2007). To control for plastid targeting, the full-length cDNA of the plastid targeted small subunit of rubisco bisphosphatecarboxylase/oxygenase (Rubisco, SSU) was fused to RFP (Murcha et al., 2007). To control for peroxisomal targeting, the peroxisomal targeting signal 1 (PTS1) was fused to the C-terminus of RFP, which has been used effectively previously (Carrie et al., 2007).

**Confirmation of organelle targeting of the GFP fusion proteins**

Biolistic transformation of the GFP constructs of interest, with their associated RFP controls was carried out on Arabidopsis cell suspension as reported previously (Carrie et al., 2007). Approximately 5 µg of both the GFP fusion protein and the RFP control, were co-precipitated onto gold particles and transiently transformed using the biolistic PDS-1000/He system (Bio-Rad, http://bio-rad.com/). Gold particles were bombarded into 2 ml of Arabidopsis cell suspension spread on filter paper placed on osmoticum plates, followed by incubation at 22 °C for 24 h in the dark. Visualisation of the fluorescent proteins was carried out using an Olympus BX61 fluorescence microscope (http://www.olympusmicro.com) with excitation wavelengths of 460/480 nm for GFP and 535/555 nm for RFP; while emission wavelengths were measured at 495-540 nm for GFP and 570-625 nm for RFP. Micrographs were captured and processed using Cell® imaging software as previously described (Carrie et al., 2007). Localisations as determined by GFP analysis: AT5G60730 - Mitochondrial, AT5G40660 - Mitochondrial, AT1G14620 - Mitochondrial, AT4G14790 - Mitochondrial, AT3G23830 - Mitochondrial, AT1G10270 - Mitochondrial, AT5G39960 - Mitochondrial, AT5G46920 - Mitochondrial, AT1G67620 - Mitochondrial, AT1G61990 - Mitochondrial, AT1G09680 - Mitochondrial, AT1G20300 - Mitochondrial, AT1G80270 - Mitochondrial, AT2G15630 - Mitochondrial, AT2G40240 - Mitochondrial, AT4G02820 - Mitochondrial, AT5G60960 - Mitochondrial, AT5G61370 - Mitochondrial, AT3G15140 - Mitochondrial, AT3G19440 - Mitochondrial, AT5G54580 - Mitochondrial, AT4G29540 - Mitochondrial, AT1G71850 - Mitochondrial, AT2G39120 - Mitochondrial, AT1G08220 - Mitochondrial, AT1G28395 - Mitochondrial, AT1G79490 - Mitochondrial, AT5G39710 - Mitochondrial, AT5G47455 - Mitochondrial, AT5G03455 - Mitochondrial, AT2G16570 - Plastid, AT1G30270 - Plastid, AT4G36040 - Plastid, AT2G27510 - Plastid, AT1G09940 - Plastid, AT4G08790 - Plastid, AT3G07780 - Plastid, AT1G78050 - Plastid, AT5G66950 - Plastid, AT3G56330 - Plastid, AT1G51080 - Plastid, AT1G52720 - Plastid, AT3G19680 - Plastid, AT3G24506 - Plastid, AT3G56360 - Plastid, AT5G06810 - Mitochondrial/Plastid, AT4G21170 - Mitochondrial/Plastid, AT1G06270 - Plastid membrane/ER, AT3G60680 - Peroxisome/Cytoskeleton, AT1G69270 - Peroxisome, AT4G14680 - Nucleus, AT2G44510 - Nucleus, AT2G45210 - Nucleus, AT1G02870 - Nucleus, AT1G07220 - Nucleus, AT1G18680 - Cytoplasmic, AT1G64185 - Cytoplasmic, AT5G51770 - Cytoplasmic, AT5G50760 - Cytoplasmic, AT5G65450 - Cytoplasmic, AT1G31940 - Cytoplasmic, AT1G74680 - Endoplasmic reticulum, AT3G21140 - Endoplasmic reticulum, AT1G04900 - Endoplasmic reticulum, AT4G00750 - Golgi.
Abbreviations

AGI  Arabidopsis Gene Identifier  
CAREs  Cis-Acting Regulatory Element(s)  
DB  Database  
DEG  Differentially Expressed Genes  
FDR  False Discovery Rate  
gDNA  Genomic DNA  
GC-RMA  GC content based Robust Multi-array Average  
GFP  Green Fluorescent Protein  
GO  Gene Ontology  
GS  Germination Specific  
mTERF  Mitochondrial transcription Termination Factor  
PPDE  Posterior Probability of Differential Expression  
RFP  Red Fluorescent Protein  
TAIR  The Arabidopsis Information Resource  
TF  Transcription Factor  
HB  Homeobox  
HSF  Heat Shock Factor  
HSE  Heat Shock factor binding Element  
WT  Wild Type  
H  Hours  
S  Stratified  
SL  (exposed to) Light (after 48 h Stratification)  
TAG  tri-acyl-glycerol

Acknowledgements

This work was supported by an Australian Research Council Centre of Excellence Grant CEO561495 and the Western Australian State Government Centres of Excellence scheme.
References

Allen PS, Benech-Arnold RL, Batlla D, Bradford KJ (2007) Modeling of seed dormancy. In KJ Bradford, H Nonogaki, eds, Seed Development, Dormancy and Germination, Vol 27. Blackwell Publishing

Baldi P, Long AD (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics 17: 509-519

Baskin CC, Baskin JM (2004) A classification system for seed dormancy. Seed Sci Res 14: 1-16

Bassel GW, Lan H, Glaab E, Gibbs DJ, Gerjets T, Krasnogor N, Bonner AJ, Holdsworth MJ, Pro Bart NJ (2011) From the Cover: Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions. Proceedings of the National Academy of Sciences of the United States of America 108: 9709-9714

Beaudoin N, Serizet C, Gosti F, Giraudat J (2000) Interactions between abscisic acid and ethylene signaling cascades. The Plant cell 12: 1103-1115

Benjamini Y, Hochberg Y (1995) Controlling false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B 57: 289-300

Boavida LC, Shuai B, Yu HJ, Pagnussat GC, Sundaresan V, McCormick S (2009) A collection of Ds insertional mutants associated with defects in male gametophyte development and function in Arabidopsis thaliana. Genetics 181: 1369-1385

Brady SM, McCourt P (2003) Hormone cross-talk in seed dormancy. Plant Growth Regul 22: 25-31

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

Cao D, Cheng H, Wu W, Soo HM, Peng J (2006) Gibberellin mobilizes distinct DELLA-dependent transcriptomes to regulate seed germination and floral development in Arabidopsis. Plant physiology 142: 509-525

Cao ZL, Yu QB, Sun Y, Lu Y, Cui YL, Yang ZN (2011) A point mutation in the pentatricopeptide repeat motif of the AtECB2 protein causes delayed chloroplast development. Journal of integrative plant biology 53: 258-269

Carrie C, Murcha MW, Millar AH, Smith SM, Whelan J (2007) Nine 3-ketoacyl-CoA thiolases (KATs) and acetoacetyl-CoA thiolases (ACATs) encoded by five genes in Arabidopsis thaliana are targeted either to peroxisomes or cytosol but not to mitochondria. Plant molecular biology 63: 97-108

Chen KM, Holmstrom M, Raksajit W, Suorsa M, Piippo M, Aro EM (2010) Small chloroplast-targeted DnaJ proteins are involved in optimization of photosynthetic reactions in Arabidopsis thaliana. Plant Mol Biol 10: 43

Chibani K, Ali-Rachedi S, Job C, Job D, Jullien M, Grappin P (2006) Proteomic analysis of seed dormancy in Arabidopsis. Plant Physiol 142: 1493-1510
Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. Genome biology 6: R16

Dagsgaard C, Taylor LE, O’Brien KM, Poyton RO (2001) Effects of anoxia and the mitochondrion on expression of aerobic nuclear COX genes in yeast: evidence for a signaling pathway from the mitochondrial genome to the nucleus. The Journal of biological chemistry 276: 7593-7601

Dave A, Hernandez ML, He Z, Andriotis VM, Vaistij FE, Larson TR, Graham IA (2011) 12-oxo-phytodienoic acid accumulation during seed development represses seed germination in Arabidopsis. The Plant cell 23: 583-599

Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold E (2003) AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. BMC bioinformatics 4: 25

Delseny M, Aspart L, Guitton Y (1977) Effect of the protein synthesis inhibitor cycloheximide on RNA synthesis in radish seedlings. Biochimie 59: 51-57

Ding YH, Liu NY, Tang ZS, Liu J, Yang WC (2006) Arabidopsis GLUTAMINE-RICH PROTEIN23 is essential for early embryogenesis and encodes a novel nuclear PPR motif protein that interacts with RNA polymerase II subunit III. The Plant cell 18: 815-830

Dohmann EM, Nill C, Schwechheimer C (2010) DELLA proteins restrain germination and elongation growth in Arabidopsis thaliana COP9 signalosome mutants. European journal of cell biology 89: 163-168

Doniwa Y, Ueda M, Ueta M, Wada A, Kadowaki K, Tsutsuji N (2010) The involvement of a PPR protein of the P subfamily in partial RNA editing of an Arabidopsis mitochondrial transcript. Gene 454: 39-46

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

Feurtado JA, Kermode AR (2007) A merging of paths: abscisic acid and hormonal cross-talk in the control of seed dormancy maintenance and alleviation. In KJ Bradford, H Nonogaki, eds, Seed Development, Dormancy and Germination, Vol 27. Blackwell Publishing

Fu Q, Wang BC, Jin X, Li HB, Han P, Wei KH, Zhang XM, Zhu YX (2005) Proteomic analysis and extensive protein identification from dry, germinating Arabidopsis seeds and young seedlings. J Biochem Mol Biol 38: 650-660

Gallardo K, Job C, Groot SP, Puype M, Demol H, Vandenekervhove J, Job D (2001) Proteomic analysis of arabidopsis seed germination and priming. Plant Physiol 126: 835-848

Gallardo K, Job C, Groot SP, Puype M, Demol H, Vandenekervhove J, Job D (2002) Proteomics of Arabidopsis seed germination. A comparative study of wild-type and gibberellin-deficient seeds. Plant Physiol 129: 823-837

Giraud E, Ng S, Carrie C, Duncan O, Low J, Pong Lee C, Van Aken O, Millar AH, Murcha MW, Whelan J (2010) TCP transcription factors link the regulation of genes encoding mitochondrial proteins with the circadian clock in Arabidopsis thaliana. Plant Physiol

Graham IA (2008) Seed storage oil mobilization. Annu Rev Plant Biol 59: 115-142
Hammani K, Gobert A, Hleibieh K, Choulier L, Small I, Giege P (2011) An Arabidopsis dual-localized pentatricopeptide repeat protein interacts with nuclear proteins involved in gene expression regulation. The Plant cell 23: 730-740
Heazlewood JL, Verboom RE, Tonti-Filippini J, Small I, Millar AH (2007) SUBA: the Arabidopsis Subcellular Database. Nucleic acids research 35: D213-218
Hilson P, Allemeersch J, Altmann T, Aubourg S, Avon A, Beynon J, Bhalerao RP, Bitton F, Caboche M, Cannoot B, Chardakov V, Cognet-Holliger C, Colot V, Crowe M, Darimont C, Durinck S, Eickhoff H, de Longevialle AF, Farmer EE, Grant M, Kuiper MT, Lehrach H, Leon C, Leyva A, Lundeberg J, Lurin C, Moreau Y, Nietfeld W, Paz-Ares J, Reymond P, Rouze P, Sandberg G, Segura MD, Serizet C, Tabrett A, Taconnat L, Thareau V, Van Hummelen P, Vercruysse S, Vuylsteke M, Weingartner M, Weisbeek PJ, Wirta V, Wittink FR, Zabeau M, Small I (2004) Versatile gene-specific sequence tags for Arabidopsis functional genomics: transcript profiling and reverse genetics applications. Genome Res 14: 2176-2189
Hirayama T, Shinozaki K (2007) Perception and transduction of abscisic acid signals: keys to the function of the versatile plant hormone ABA. Trends in plant science 12: 343-351
Holdsworth MJ, Bentsink L, Soppe WJ (2008) Molecular networks regulating Arabidopsis seed maturation, after-ripening, dormancy and germination. The New phytologist 179: 33-54
Holdsworth MJ, Finch-Savage WE, Grappin P, Job D (2008) Post-genomics dissection of seed dormancy and germination. Trends in plant science 13: 7-13
Howell KA, Cheng K, Murcha MW, Jenkin LE, Millar AH, Whelan J (2007) Oxygen initiation of respiration and mitochondrial biogenesis in rice. J Biol Chem 282: 15619-15631
Howell KA, Millar AH, Whelan J (2006) Ordered assembly of mitochondria during rice germination begins with pro-mitochondrial structures rich in components of the protein import apparatus. Plant Mol Biol 60: 201-223
Howell KA, Narsai R, Carroll A, Ivanova A, Lohse M, Usadel B, Millar AH, Whelan J (2009) Mapping Metabolic and Transcript Temporal Switches during Germination in Rice Highlights Specific Transcription Factors and the Role of RNA Instability in the Germination Process. Plant Physiol 149: 961-980
Kappen C (2000) Analysis of a complete homeobox gene repertoire: implications for the evolution of diversity. Proceedings of the National Academy of Sciences of the United States of America 97: 4481-4486
Kecpczyński J, Kecpczyńska E (1997) Ethylene in seed dormancy and germination. Physiologia Plantarum 101: 720-726
Keren I, Bezawork-Geleta A, Kolton M, Maayan I, Belausov E, Levy M, Mett A, Gidoni D, Shaya F, Ostertzer-Biran O (2009) AtnMat2, a nuclear-encoded maturase required for splicing of group-II introns in Arabidopsis mitochondria. RNA 15: 2299-2311
Kim YO, Pan S, Jung CH, Kang H (2007) A zinc finger-containing glycine-rich RNA-binding protein, atRZ-1a, has a negative impact on seed germination and seedling growth of Arabidopsis thaliana under salt or drought stress conditions. Plant & cell physiology 48: 1170-1181
Kosugi S, Suzuika I, Ohashi Y (1995) Two of three promoter elements identified in a rice gene for proliferating cell nuclear antigen are essential for meristematic tissue-specific expression. The Plant journal : for cell and molecular biology 7: 877-886
Kotak S, Larkindale J, Lee U, von Koskull-Doring P, Vierling E, Scharf KD (2007) Complexity of the heat stress response in plants. Current opinion in plant biology 10: 310-316

Kotak S, Vierling E, Baumlein H, von Koskull-Doring P (2007) A novel transcriptional cascade regulating expression of heat stress proteins during seed development of Arabidopsis. The Plant cell 19: 182-195

Kuromori T, Wada T, Kamiya A, Yuguichi M, Yokouchi T, Imura Y, Takabe H, Sakurai T, Akiyama K, Hirayama T, Okada K, Shinozaki K (2006) A trial of phenome analysis using 4000 Ds-insertional mutants in gene-coding regions of Arabidopsis. Plant J 47: 640-651

Kwak KJ, Park SJ, Han JH, Kim MK, Oh SH, Han YS, Kang H (2011) Structural determinants crucial to the RNA chaperone activity of glycine-rich RNA-binding proteins 4 and 7 in Arabidopsis thaliana during the cold adaptation process. Journal of experimental botany

Li Z, Thomas TL (1998) PEI1, an embryo-specific zinc finger protein gene required for heart-stage embryo formation in Arabidopsis. The Plant cell 10: 383-398

Linkies A, Muller K, Morris K, Tureckova V, Wenk M, Cadman CS, Corbineau F, Strnad M, Lynn JR, Finch-Savage WE, Leubner-Metzger G (2009) Ethylene interacts with abscisic acid to regulate endosperm rupture during germination: a comparative approach using Lepidium sativum and Arabidopsis thaliana. The Plant cell 21: 3803-3822

Lister R, Carrie C, Duncan O, Ho LH, Howell KA, Murcha MW, Whelan J (2007) Functional definition of outer membrane proteins involved in preprotein import into mitochondria. Plant Cell 19: 3739-3759

Logan DC, Leaver CJ (2000) Mitochondria-targeted GFP highlights the heterogeneity of mitochondrial shape, size and movement within living plant cells. J Exp Bot 51: 865-871

Lu Y, Li C, Wang H, Chen H, Berg H, Xia Y (2011) AtPPR2, an Arabidopsis pentatricopeptide repeat protein, binds to plastid 23S rRNA and plays an important role in the first mitotic division during gametogenesis and in cell proliferation during embryogenesis. The Plant journal : for cell and molecular biology 67: 13-25

Meinke D, Muralla R, Sweeney C, Dickerman A (2008) Identifying essential genes in Arabidopsis thaliana. Trends in plant science 13: 483-491

Morohashi Y, Bewley JD, Yeung EC (1981) Biogenesis of Mitochondria in Imbibed Peanut Cotyledons : II. DEVELOPMENT OF LIGHT AND HEAVY MITOCHONDRIA. Plant physiology 68: 318-323

Nakabayashi K, Okamoto M, Koshita T, Kamiya Y, Nambara E (2005) Genome-wide profiling of stored mRNA in Arabidopsis thaliana seed germination: epigenetic and genetic regulation of transcription in seed. Plant J 41: 697-709

Nakagawa N, Sakurai N (2006) A mutation in At-nMat1a, which encodes a nuclear gene having high similarity to group II intron maturase, causes impaired splicing of mitochondrial NAD4 transcript and altered carbon metabolism in Arabidopsis thaliana. Plant Cell Physiol 47: 772-783

Narsai R, Castleden I, Whelan J (2010) Common and distinct organ and stress responsive transcriptomic patterns in Oryza sativa and Arabidopsis thaliana. BMC plant biology 10: 262
Narsai R, Howell KA, Carroll A, Ivanova A, Millar AH, Whelan J (2009) Defining core metabolic and transcriptomic responses to oxygen availability in rice embryos and young seedlings. Plant physiology 151: 306-322

Narsai R, Howell KA, Millar AH, O'Toole N, Small I, Whelan J (2007) Genome-wide analysis of mRNA decay rates and their determinants in Arabidopsis thaliana. The Plant cell 19: 3418-3436

Newman TC, Ohme-Takagi M, Taylor CB, Green PJ (1993) DST sequences, highly conserved among plant SAUR genes, target reporter transcripts for rapid decay in tobacco. Plant Cell 5: 701-714

Nishimura N, Kitahata N, Seki M, Narusaka Y, Narusaka M, Kuromori T, Asami T, Shinozaki K, Hirayama T (2005) Analysis of ABA hypersensitive germination revealed the pivotal functions of PARN in stress response in Arabidopsis. The Plant journal : for cell and molecular biology 44: 972-984

O'Toole N, Hattori M, Andres C, Iida K, Lurin C, Schmitz-Linnweber C, Sugita M, Small I (2008) On the expansion of the pentatricopeptide repeat gene family in plants. Mol Biol Evol 25: 1120-1128

Osakabe Y, Maruyama K, Seki M, Satou M, Shinozaki K, Yamaguchi-Shinozaki K (2005) Leucine-rich repeat receptor-like kinase1 is a key membrane-bound regulator of abscisic acid early signaling in Arabidopsis. The Plant cell 17: 1105-1119

Pagnussat GC, Yu HJ, Ngo QA, Rajani S, Mayalagu S, Johnson CS, Capron A, Xie LF, Ye D, Sundaresan V (2005) Genetic and molecular identification of genes required for female gametophyte development and function in Arabidopsis. Development 132: 603-614

Palaniswamy SK, James S, Sun H, Lamb RS, Davuluri RV, Grotewold E (2006) AGRIS and AtRegNet, a platform to link cis-regulatory elements and transcription factors into regulatory networks. Plant physiology 140: 818-829

Perrin R, Meyer EH, Zaepfle M, Kim YJ, Mache R, Grienenberger JM, Gualberto JM, Gagliardi D (2004) Two exoribonucleases act sequentially to process mature 3'-ends of atp9 mRNAs in Arabidopsis mitochondria. J Biol Chem 279: 25440-25446

Pesaresi P, Masiero S, Eubel H, Braun HP, Bhushan S, Glaser E, Salamini F, Leister D (2006) Nuclear photosynthetic gene expression is synergistically modulated by rates of protein synthesis in chloroplasts and mitochondria. The Plant cell 18: 970-991

Rajjou L, Gallardo K, Debeaujon I, Vandekerckhove J, Job C, Job D (2004) The effect of alpha-amanitin on the Arabidopsis seed proteome highlights the distinct roles of stored and neosynthesized mRNAs during germination. Plant Physiol 134: 1598-1613

Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and inparalogs from pairwise species comparisons. J Mol Biol 314: 1041-1052

Reumann S, Ma C, Lemke S, Babujee L (2004) AraPerox. A database of putative Arabidopsis proteins from plant peroxisomes. Plant physiology 136: 2587-2608

Riano-Pachon DM, Ruzicic S, Dreyer I, Mueller-Roeber B (2007) PlnTFDB: an integrative plant transcription factor database. BMC Bioinformatics 8: 42

Roberti M, Polosa PL, Bruni F, Manzari C, Dececchie S, Gadaleta MN, Cantatore P (2009) The MTERF family proteins: mitochondrial transcription regulators and beyond. Biochimica et biophysica acta 1787: 303-311

Russell L, Larner V, Kurup S, Bougourd S, Holdsworth M (2000) The Arabidopsis COMATOSE locus regulates germination potential. Development 127: 3759-3767
Saiga S, Furumizu C, Yokoyama R, Kurata T, Sato S, Kato T, Tabata S, Suzuki M, Komeda Y (2008) The Arabidopsis OBERON1 and OBERON2 genes encode plant homeodomain finger proteins and are required for apical meristem maintenance. Development 135: 1751-1759

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU (2005) A gene expression map of Arabidopsis thaliana development. Nature genetics 37: 501-506

Schmidt F, Marnef A, Cheung MK, Wilson I, Hancock J, Staiger D, Ladomery M (2010) A proteomic analysis of oligo(dT)-bound mRNP containing oxidative stress-induced Arabidopsis thaliana RNA-binding proteins ATGRP7 and ATGRP8. Molecular biology reports 37: 839-845

Schmitz-Linneweber C, Small I (2008) Pentatricopeptide repeat proteins: a socket set for organelle gene expression. Trends in plant science 13: 663-670

Schoffl F, Prandl R, Reindl A (1998) Regulation of the heat-shock response. Plant physiology 117: 1135-1141

Sreenivasulu N, Usadel B, Winter A, Radchuk V, Scholz U, Stein N, Weschke W, Strickert M, Close TJ, Stitt M, Graner A, Wobus U (2008) Barley grain maturation and germination: metabolic pathway and regulatory network commonalities and differences highlighted by new MapMan/PageMan profiling tools. Plant Physiol 146: 1738-1758

Tiwari SB, Hagen G, Guilfoyle TJ (2004) Aux/IAA proteins contain a potent transcriptional repression domain. Plant Cell 16: 533-543

Tyler L, Thomas SG, Hu J, Dill A, Alonso JM, Ecker JR, Sun TP (2004) Della proteins and gibberellin-regulated seed germination and floral development in Arabidopsis. Plant physiology 135: 1008-1019

Ulamasov T, Murfett J, Hagen G, Guilfoyle TJ (1997) Aux/IAA proteins repress expression of reporter genes containing natural and highly active synthetic auxin response elements. Plant Cell 9: 1963-1971

Usadel B, Nagel A, Steinauher D, Gibon Y, Blasing OE, Redestig H, Sreenivasulu N, Krall L, Hannah MA, Poree F, Fernie AR, Stitt M (2006) PageMan: an interactive ontology tool to generate, display, and annotate overview graphs for profiling experiments. BMC bioinformatics 7: 535

Wang BC, Wang HX, Feng JX, Meng DZ, Qu LJ, Zhu YX (2006) Post-translational modifications, but not transcriptional regulation, of major chloroplast RNA-binding proteins are related to Arabidopsis seedling development. Proteomics 6: 2555-2563

Weitbrecht K, Muller K, Leubner-Metzger G (2011) First off the mark: early seed germination. Journal of experimental botany

Yamauchi Y, Ogawa M, Kuwahara A, Hanada A, Kamiya Y, Yamaguchi S (2004) Activation of gibberellin biosynthesis and response pathways by low temperature during imbibition of Arabidopsis thaliana seeds. Plant Cell 16: 367-378

Yilmaz A, Mejia-Guerra MK, Kurz K, Liang X, Welch L, Grotewold E (2011) AGRIS: the Arabidopsis Gene Regulatory Information Server, an update. Nucleic acids research 39: D1118-1122

Yu QB, Jiang Y, Chong K, Yang ZN (2009) AtECB2, a pentatricopeptide repeat protein, is required for chloroplast transcript accD RNA editing and early chloroplast biogenesis in Arabidopsis thaliana. The Plant journal : for cell and molecular biology 59: 1011-1023
Figure legends

Figure 1. Overview of transcriptomic changes during stratification and germination. A) The number of genes significantly (p<0.05, PPDE>0.96) differentially expressed during dark stratification at 4 ºC (in black boxes) and following this, into continuous light. Stratification = S, continuous light following stratification = SL. The total number of genes up-regulated (red) or down-regulated (blue) in each comparison is shown above each of the bars. B) For each comparison, the significant fold-changes were analysed using the PageMan tool (Usadel et al., 2006). Statistical analysis of over-represented functional categories was carried out using the ORA Fisher method with Benjamini-Hochberg false discovery rate correction. For the larger functional categories, average z-scores are visualised. Functional categories that did not show significant changes were collapsed for display (detailed functional categories shown in Supplementary table 2). Statistical significances are represented by a false colour heat map (up, orange; down, green) where a z-score of 1.96 represents a p-value of 0.05.

Figure 2. Seed morphology, hierarchical clustering of expression profiles and GO representation of genes during Arabidopsis germination. A) Light microscopy was used to examine the changes in morphology over the germination time course. B) All genes called present at a minimum of one time point were normalised to the highest level of expression over the time course of the study and hierarchically clustered using average linkage based on Euclidian distance. Four clusters were identified: Cluster 1 (pink) – transcripts which increased in abundance over the time period examined, Cluster 2 (blue) – transcripts highly expressed in dry seed to 12 h S and decrease in abundance over time, Cluster 3 – transcripts with low or absent abundance in dry seeds, peaked between 48 h S and 6 h SL and then declined in abundance and Cluster 4 (black) – transcripts that displayed relatively stable levels of abundance throughout the time course. C) For each Cluster, all genes were analysed for representation in GO categories. Over/under-represented functional categories were identified by z-score analysis, where statistical significances are represented by a false colour heat map (up, orange; down, green) where a z-score of 1.96 represents a p-value of 0.05. D) Protein abundance profiles (from Fu et al., 2005) were compared to transcript abundance profiles during germination. The percentage of transcripts in each cluster that showed similar (up-regulated/down-regulated) protein abundance profiles during germination is indicated.
Figure 3. Identification of germination specific gene expression. A) Publically available microarrays carried out on wild-type tissues in the AtGenExpress developmental set (Schmid et al. 2005, E-AFMX-9) were normalised together with the arrays in this study. Hierarchical clustering of the relative expression levels for 775 unique genes identified as showing the highest expression during germination, with expression levels <50% of these levels in all other tissues. B) Examination of over/under-represented GO categories for these transiently expressed genes revealed significant over-representation of genes encoding proteins targeted to the mitochondria, nucleus and ribosomes, corresponding with the observed over-representation of DNA and RNA binding functions. The over-representation of these functions, specifically during germination implies that these genes are likely to encode the regulatory proteins required specifically for germination progression and later development. Over/under-represented functional categories were identified by z-score analysis. Statistical significances are represented by a false colour heat map (up, orange; down, green) where a z-score of 1.96 represents a p-value of 0.05. C) The distribution of genes into functional sub-categories for the genes categorised as i) mitochondrial and ii) nuclear localised. D) i) The percentage of PPR encoding genes in each cluster is displayed as a pie-chart below the percentage of genes in each cluster in the genome. ii) Expression levels during germination without (No stratification) and including stratification for the 75 PPR genes identified as possessing germination specific expression in this study. iii) Expression of the rice genes orthologous to the 75 Arabidopsis genes shown across germination under aerobic and anaerobic conditions in rice.

Figure 4. Functional analysis of proteins encoded by genes defined as germination specific. Of the 775 genes showing germination specific expression, 114 genes showed a published phenotype when mutated/silenced/knocked out. This number is significantly greater (p<0.001) than the expected percentage of genes showing phenotypes in random sets of 775 genes. A) The types of phenotypes observed for the 114 genes with known phenotypes in the germination specific set. The number of genes displaying each phenotype is indicated (details can be seen in Supplementary Table 6). The significant over-representation of seed lethal phenotypes is indicated with a red asterisk. Of the 114 genes, the number of those encoding RNA binding/processing functions is also shown in the centre of the pie chart. B) From a list of 775 randomly selected genes, 79 genes (on average, across 3 independent, randomly selected sets) were found to have phenotypes. The number of genes and distribution of phenotypes observed for these are shown. Of the 79 genes, the number of those encoding RNA binding/processing functions is also shown in the centre of the pie chart. Note that the pie chart
(B) is smaller than that of the 114 genes (A), as it is drawn to scale for the number of phenotypes observed.

**Figure 5** GFP analysis confirms localisation of proteins encoded by genes expressed during germination. A subset of 65 genes predicted to encode organelle targeted proteins were analysed to determine sub-cellular localisation. Only genes most highly expressed during germination i.e. showing low expression in dry seed and post-germination (48 h SL) were selected for analysis. (A) Examples of fluorescent images for proteins targeted to the mitochondria, plastids and peroxisomes. The images for every protein identified as targeting to an organelle can be found in Supplementary Figure 5. Fusion proteins containing GFP and the first 100 amino acids of the protein of interest (or the last 100 amino acids in the case of the peroxisomal predicted proteins) were constructed and transiently transformed into cell culture using biolistic transformation. Organelle targeting was verified using AOX-RFP as a mitochondrial control, SSU-RFP as a plastid control and RFP-SRL as a control for peroxisomal targeting. (B) A heat map showing expression levels during germination for the 65 genes analysed by GFP targeting. The Arabidopsis Gene Identifier for each gene, brief gene symbol/description and localisation determined by GFP is shown next to the heatmap. Predicted localisations are indicated in brackets (M=Mitochondria, Pl=Plastids, Per=Peroxisomes) next to the experimentally determined localisation. Note that genes which were defined as most highly expressed during germination in Figure 3A, are indicated by ^. Known phenotypes when gene is knocked out/silenced: Defective in - Flower (F), Stress (St), Embryo (ED), Growth (G), Seedling (SL). Reference for phenotypes: 1(Boavida et al., 2009); 2(Kwak et al., 2011); 3(Ding et al., 2006); 4(Keren et al., 2009); 5(Hilson et al., 2004); 6(Hammani et al., 2011); 7(Meinke et al., 2008); 8(Kuromori et al., 2006); 9(Chen et al., 2010); 10(Saiga et al., 2008); 11(Osakabe et al., 2005). AGIs in grey indicate those annotated as involved in RNA binding/processing.

**Figure 6. Analysis of putative regulatory factors affecting transcript abundance.** A) Parallel comparison showing the average profile of genes in Clusters 1-4 (i-iv) for transcripts expressed during germination without stratification (“No stratification”) and including stratification (“This study”). B) The mRNA half-lives of all genes expressed during germination is shown as a pie-chart indicating the percentage of genes that had mRNA half-lives of differing lengths i.e. 0 h to over 24 h (Genome). For each cluster (1-4; i-iv), the distribution of genes into the different mRNA half-life increments are indicated, with red/blue font indicating an over/under-representation of genes with mRNA half-lives for that increment. C) Expression profiles of all the
genes encoding transcription factors in Arabidopsis (Riano-Pachon et al., 2007; 1299 expressed during germination) was used to determine whether there was an overall over/under representation of TFs in each cluster. i) The percentage of TFs in each cluster is compared with the percentage of genes in that cluster in the genome. ii) The representation of specific TF families within each cluster was compared with the percentage present in the genome to show if specific families were over/under-represented, which is indicated by the red or blue font respectively (at p<0.05). TF families that were also over-represented and showed the same expression pattern in rice germination are indicated in black boxes.
Table captions

Table 1. Known motifs that were found to be significantly over-represented in genes expressed over germination. The specific hexamer, name of the known motif, reference (ref.) showing the source of the motif; 1=AGRIS, 2=Schoffl et al., 1998, 3=Kosugi et al., 1995 and occurrence in all genes expressed during germination, in each cluster and the germination specific (GS) subset are shown (bold indicates significant over-representation). Note the motifs with an * indicate that these were also identified as over-represented in dry seed by Nakabayashi et al., 2005.
| Hexamer | Motif name | Ref. | Genome (15,789) | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | GS |
|---------|------------|------|----------------|-----------|-----------|-----------|-----------|----|
| CACGTG* | ABRE, CBF2, G-box motif | 1 | 2927 (19%) | 886 (19%) | 978 (23%) | 375 (16%) | 688 (15%) | 98 (13%) |
| CGTTCG  | ABRE-like binding site motif | 1 | 1799 (11%) | 508 (11%) | 636 (15%) | 239 (10%) | 416 (9%) | 79 (10%) |
| ACCTTG  | ABRE-like binding site motif, Z-box | 1 | 2888 (18%) | 850 (18%) | 967 (23%) | 351 (15%) | 720 (16%) | 102 (13%) |
| ACACGT* | ACE promoter motif | 1 | 2916 (18%) | 854 (18%) | 973 (23%) | 365 (16%) | 724 (16%) | 99 (13%) |
| GCCGAC  | CBF1 BS in cor15a | 1 | 799 (5%) | 222 (5%) | 282 (7%) | 100 (4%) | 195 (4%) | 29 (4%) |
| CACCTG* | CBF2 binding site motif, GBF | 1 | 2240 (14%) | 701 (15%) | 729 (17%) | 276 (12%) | 534 (12%) | 96 (12%) |
| TGCCGA  | DRE-like promoter motif | 1 | 781 (5%) | 221 (5%) | 266 (6%) | 103 (4%) | 191 (4%) | 35 (5%) |
| TTCCCG  | E2F binding site motif | 1 | 1342 (9%) | 347 (7%) | 363 (9%) | 235 (10%) | 397 (9%) | 77 (10%) |
| CAAGGG  | EIL1 BS in ERF1 | 1 | 1246 (8%) | 338 (7%) | 345 (9%) | 225 (10%) | 342 (8%) | 56 (7%) |
| GCCGAC  | ERE promoter motif, GCC-box | 1 | 748 (5%) | 181 (4%) | 201 (5%) | 137 (6%) | 229 (5%) | 44 (6%) |
| TAGGGG  | LFY BS in AP3 | 1 | 992 (6%) | 274 (6%) | 243 (6%) | 184 (8%) | 291 (6%) | 66 (9%) |
| CGGCCC  | SORLIP2 | 1 | 715 (5%) | 194 (4%) | 194 (5%) | 127 (6%) | 200 (4%) | 57 (6%) |
| CCACGT* | CBF1 BS in cor15a | 1 | 2240 (14%) | 701 (15%) | 729 (17%) | 276 (12%) | 534 (12%) | 96 (12%) |
| TGCCGA  | DRE-like promoter motif | 1 | 781 (5%) | 221 (5%) | 266 (6%) | 103 (4%) | 191 (4%) | 35 (5%) |

Downloaded from www.plantphysiol.org on August 20, 2017. Published by www.plantphysiol.org Copyright © 2011 American Society of Plant Biologists. All rights reserved.
Supplementary Figures

Supplementary Figure 1. Gene expression profiles during germination without stratification (data from Nakabayashi et al., 2005) were also hierarchically clustered in parallel. Clusters showing similar expression profiles in both studies are indicated by the coloured lines.

Supplementary Figure 2. Gene expression across Arabidopsis germination and development. Publically available microarrays carried out on wild-type tissues in the AtGenExpress developmental set (Schmid et al. 2005, E AFMX-9) were normalised together with the arrays in this study. Hierarchical clustering of the log2 expression levels for the 15,789 unique genes expressed during germination.

Supplementary Figure 3. Expression of rice genes orthologous to 775 Arabidopsis GS genes across rice germination and development. Publically available microarrays carried out on rice wild-type tissues across germination and development (Supplementary Table 4) were normalised together (materials and methods). A) Relative expression values were hierarchically clustered and visualised to determine whether germination specific expression of these genes is conserved across rice and Arabidopsis. Genes in the yellow box indicates conserved high expression from dry seed to 30 h post imbibition. B) The sub-set of rice genes orthologous to the 75 germination specific PPR genes in Arabidopsis were hierarchically clustered across germination and development in rice, as in (A).

Supplementary Figure 4. Expression profiles of all 422 genes with known seed phenotypes (described in the SeedGenes database, Meinke et al., 2008). A) The percentage of seed-genes in each cluster is displayed. B) Expression levels during germination without (No stratification) and including stratification for the 35 seed-genes identified as showing germination specific expression in this study. C) Expression of the rice genes orthologous to the 35 Arabidopsis genes shown across germination under aerobic and anaerobic conditions in rice. A red/blue asterisk* indicates that genes were over/under-represented in a particular cluster, respectively.

Supplementary Figure 5. GFP fluorescence images for the 65 proteins; targeted to the mitochondria, plastids and/or peroxisomes. Fusion proteins containing GFP and the first 100 amino acids of the protein of interest (or the last 100 amino acids in the case of the peroxisomal
predicted proteins) were constructed and transiently transformed into cell culture using biolistic transformation.

**Supplementary Figure 6.** Pageman output for differentially expressed genes during Arabidopsis germination A) without stratification, original microarray data from Nakabayashi et al., 2005 and B) including stratification, as in this study. The Pageman Category and z-scores are shown, following ORA analysis using Fishers test after Benjamini-Hochberg FDR correction.

**Supplementary Figure 7.** Gene expression profiles during germination with stratification (this study) and without stratification (data from Nakabayashi et al., 2005) are visualised in parallel for genes encoding proteins involved in ethylene related functions (based on Pageman bins).

**Supplementary Tables**

**Supplementary Table 1.** Differential expression analysis in dry seeds, stratified and germinating seedlings. For each of the 15,789 genes that were expressed during the time course, the Affymetrix probeset identifier, Arabidopsis Gene Identifier (AGI), cluster number, fold change, p-value (Bayes.p) and FDR p-value (PPDE<p) are shown for each comparison. For each respective AGI, the mitochondrial, plastid and peroxisome localisation as predicted by/experimentally annotated in SUBA (Heazlewood et al., 2007) is indicated where 1=experimentally shown localisation, 2=predicted only, 3=predicted and experimentally shown. Given the minimal annotation for peroxisomal predicted/localised genes, for these genes, 4=as annotated in AraPerox (Reumann et al., 2004) as well.

**Supplementary Table 2.** Pageman output for differentially expressed genes during Arabidopsis germination. The Pageman Category, sub-category and z-scores for the ORA analysis is shown. These were calculated using the Fisher test after Benjamini-Hochberg FDR correction.

**Supplementary Table 3.** Details of the developmental tissue set (Schmid et al., 2005; E-AFMX-9) ATGE accessions, number of biological replications, the sample details, genotype, tissue, age, photoperiod, growth substrate, Number on Figure 3 (Array no.) are shown below.

**Supplementary Table 4.** Details of the developmental tissue microarrays used for rice. The array number (Array no.) as shown in Supplementary Figure , the publications addressing
these microarrays, the GEO or MIAME Gene express accessions, number of biological replications, experimental details, the individual sample details, cultivar, tissue and age are shown below.

**Supplementary Table 5.** Comparison of protein abundance data from Fu et al., 2005 with the parallel transcript abundance profiles in this germination study. The protein profile description and stage of protein expression refer directly to the findings from Fu et al., 2005. Arabidopsis gene identifiers (AGI) with ^ indicate that this protein was detected in Gallardo et al., 2001 as well, * indicates that this protein was also detected in Chibani et al., 2006. The column named Transcript cluster refers to the gene clusters as identified in Figure 2B of this study; Cluster 1 - increasing transcript abundance over time, Cluster 2 - high transcript abundance in dry seed and decreasing over time, Cluster 3 - transient expression during germination, Cluster 4 - slow/unchanging expression over germination.

**Supplementary Table 6. A)** The Affymetrix probe identifier, AGI (chromosome locus), TAIR annotation, Phenotype, Source database (DB) and reference is shown for each gene that was found to have a phenotype. **B)** Details showing the number of genes having known phenotypes from 3 sets of 775 random genes (RS 1-3), the average of these (as shown in Figure 4) and the numbers seen for the 775 GS set (as in Figure 4). **C)** Full references for genes from A.

**Supplementary Table 7.** A selection of genes (>75% from Cluster 3) with predicted organellar localisations were analysed by GFP tagging to confirm protein localisation. The Arabidopsis Gene Identifier, annotation, predicted location, GFP shown location, whether or not this gene has a known seed phenotype (SeedGenes phenotype - Meinke et al., 2008) and Cluster from which this gene was selected.
Figure 1. Overview of transcriptomic changes during stratification and germination. A) The number of genes significantly (p<0.05, PPDE>0.96) differentially expressed during dark stratification at 4 °C (in black boxes) and following this, into continuous light. Stratification = S, continuous light following stratification = SL. The total number of genes up-regulated (red) or down-regulated (blue) in each comparison is shown above each of the bars. B) For each comparison, the significant fold-changes were analysed using the PageMan tool (Usadel et al., 2006). Statistical analysis of over-represented functional categories was carried out using the ORA Fisher method with Benjamini-Hochberg false discovery rate correction. For the larger functional categories, average z-scores are visualised. By default, data shown that fold-change > 2, significant changes were collapsed for display (detailed functional categories shown in Supplementary table 2). Statistical significances are represented by a false colour heat map (up, orange; down, green) where a z-score of 1.96 represents a p-value of 0.05.
Figure 2. Seed morphology, hierarchical clustering of expression profiles and GO representation of genes during Arabidopsis germination. A) Light microscopy was used to examine the changes in morphology over the germination time course. B) All genes called present at a minimum of one time point were normalised to the highest level of expression over the time course of the study and hierarchically clustered using average linkage based on Euclidian distance. Four clusters were identified: Cluster 1 (pink) – transcripts which increased in abundance over the time period examined, Cluster 2 (blue) – transcripts highly expressed in dry seed to 12 h S and decrease in abundance over time, Cluster 3 – transcripts with low or absent abundance in dry seeds, peaked between 48 h S and 6 h SL and then declined in abundance and Cluster 4 (black) – transcripts that displayed relatively stable levels of abundance throughout the time course. C) For each Cluster, all genes were analysed for representation in GO categories. Over/under-represented functional categories were identified by z-score analysis, where statistical significances are represented by a false colour heat map (up, orange; down, green). A z-score of 1.96 represents a p-value of 0.05. D) Protein abundance profiles (from Fu et al., 2008) were compared to transcriptomic expression profiles during germination. The percentage of transcripts in each cluster that showed similar (up-regulated/down-regulated) protein abundance profiles during germination is indicated.
Figure 3. Identification of germination specific gene expression. A) Publicly available microarrays carried out on wild-type tissues in the AtGenExpress developmental set (Schmid et al. 2005, E-AFMY-9) were normalised together with the arrays in this study. Hierarchical clustering of the relative expression levels for 775 unique genes identified as showing the highest expression during germination, with expression levels <50% of these levels in all other tissues. B) Examination of over/under-represented GO categories for these transiently expressed genes revealed significant over-representation of genes encoding proteins targeted to the mitochondria, nucleus and ribosomes, corresponding with the observed over-representation of DNA and RNA binding functions. The over-representation of these functions, specifically during germination, is consistent with the phenotypic characteristics of genes required specifically for germination progression and later development. Over/under-represented functional categories were identified by gene-expression analysis. Statistical significances are represented by a false colour heat map (up, orange; down, green) where a Z-score of 1.96 represents a p-value of 0.05. C) The distribution of genes into functional sub-categories for the genes categorised as i) mitochondrial and ii) nuclear localised. D) i) The percentage of PPR encoding genes in each cluster is displayed as a pie-chart below the percentage of genes in each cluster in the genome. ii) Expression levels during germination without (No stratification) and including stratification for the 75 PPR genes identified as possessing germination specific expression in this study. iii) Expression of the rice genes orthologous to the 75 Arabidopsis genes shown across germination under aerobic and anaerobic conditions in rice.
Figure 4. Functional analysis of proteins encoded by genes defined as germination specific. Of the 775 genes showing germination specific expression, 114 genes showed a published phenotype when mutated/silenced/knocked out. This number is significantly greater (p<0.001) than the expected percentage of genes showing phenotypes in random sets of 775 genes. A) The types of phenotypes observed for the 114 genes with known phenotypes in the germination specific set. The number of genes displaying each phenotype is indicated (details can be seen in Supplementary Table 6). The significant over-representation of seed lethal phenotypes is indicated with a red asterisk. Of the 114 genes, the number of those encoding RNA binding/processing functions is also shown in the centre of the pie chart. B) From a list of 775 randomly selected genes, 79 genes (on average, across 3 independent, randomly selected sets) were found to have phenotypes. The number of genes and distribution of phenotypes observed for these are shown. Of the 79 genes, the number of those encoding RNA binding/processing functions is also shown in the centre of the pie chart. Note that the pie chart (B) is smaller than that of the 114 genes (A), as it is drawn to scale for the number of phenotypes observed.
Figure 5 GFP analysis confirms localisation of proteins encoded by genes expressed during germination. A subset of 65 genes predicted to encode organelle targeted proteins were analysed to determine sub-cellular localisation. Only genes most highly expressed during germination i.e. showing low expression in dry seed and post-germination (48 h SL) were selected for analysis. (A) Examples of fluorescent images for proteins targeted to the mitochondria, plastids and peroxisomes. The images for every protein identified as targeting to an organelle can be found in Supplementary Figure 5. Fusion proteins containing GFP and the first 100 amino acids of the protein of interest (or the last 100 amino acids in the case of the peroxisomal predicted proteins) were constructed and transiently transformed into cell culture using biolistic transformation. Organelle targeting was verified using AOX-RFP as a mitochondrial control, SSU-RFP as a plastid control and RFP-SRL as a control for peroxisomal targeting. (B) A heat map showing expression levels during germination for the 65 genes analysed by GFP targeting. The Arabidopsis Gene Identifier for each gene, brief gene symbol/description and localisation determined by GFP is shown next to the heatmap. Predicted localisations are indicated in brackets (M=mitochondria, Pl=plastids, Per=peroxisomes) next to the experimentally determined localisation. Note that genes which were defined as most highly expressed during germination in Figure 3A, are indicated by ^_. Known phenotypes when gene is knocked out/silenced: Defective in - Flower (F), Stress (St), Embryo (Ed), Growth (G), Seedling (SL). References for observed phenotypes: 1Bovida et al., 2009; 2KwaK et al., 2011; 3Ding et al., 2006; 4Keren et al., 2009; 5Hilson et al., 2004; 6Hammani et al., 2011; 7Meinke et al., 2008; 8Kurosaki et al., 2006; 9Chen et al., 2010; 10Saiga et al., 2008; 11Osakabe et al., 2005. AGIs in grey indicate those annotated as involved in RNA binding/processing.
Figure 6. Analysis of putative regulatory factors affecting transcript abundance. A) Parallel comparison showing the average profile of genes in Clusters 1-4 (i-iv) for transcripts expressed during germination without stratification (“No stratification”) and including stratification (“This study”). B) The mRNA half-lives of all genes expressed during germination is shown as a pie-chart indicating the percentage of genes that had mRNA half-lives of differing lengths i.e. 0 h to over 24 h (Genome). For each cluster (1-4; i-iv), the distribution of genes into the different mRNA half-life increments are indicated, with red/blue font indicating an over/under-representation of genes with mRNA half-lives for that increment. C) Expression profiles of all the genes encoding transcription factors in Arabidopsis (Riano-Pachon et al., 2007; 1299 expressed during germination) was used to determine whether there was an overall over/under representation of TFs in each cluster. i) The percentage of TFs in each cluster is compared with the percentage of genes in each cluster that they control. ii) The representation of specific TF families within each cluster was compared with the percentage present in the genome to determine whether specific families were over/under-represented, which is indicated by the red or blue font respectively (at p<0.05). TF families that were also over-represented and showed the same expression pattern in rice germination are indicated in black boxes.