Supplemental Data file accompanying the Dekkers et al manuscript entitled "Transcriptional Dynamics of Two Seed Compartments with Opposing Roles in Arabidopsis Seed Germination"

SUPPLEMENTAL DATA

Comparisons with other seed transcriptome datasets (with Supplemental Fig. S3).

1. Endosperm/Embryo Dataset. Taking the 18 .cel files published as part of the article (Penfield et al., 2006), [http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE5751], we renormalize the chips using the RMA normalization procedure and the custom CDF as detailed in the material & methods, to ensure compatibility with our data. The resulting probe set distribution suggests that the noise region of the data is <5 (log_2 RMA, Supplemental Fig. S3A). We calculate which probe sets are differentially expressed at a 5-fold level, thresholding the data at 4 and then perform a t-test to check each expressed probe set is significantly different at a p-value of 0.05. This resulted in 445 (434) genes which are 5 fold up-regulated in the post-radicle emergence Endosperm (Embryo), when using our methods of analysis (see Supplemental materials & methods). Compared to a list of genes which are 5 fold different between our MCE 38 HAS ER and RAD 38 HAS ER samples (the equivalent to their radicle protrusion at 24 HAS post-stratification), 277/445 of genes were in both Endosperm up lists, and 145/432 genes are in both Embryo up lists (Supplemental Fig. S3B). This represents a significant overlap certainly if one takes into account that the Penfield data compared embryo vs whole endosperm of stratified seeds while in our case non-stratified were used and compared and RAD vs halved endosperm, the MCE. This also provides an explanation why the overlap between EMB Up list is smaller, as this is likely due to the absence of COT specific genes in our RAD sample. There are only a few genes in the EMB Up/MCE 38 ER Up and the END Up / RAD 38 ER Up overlap (7 and 2 genes, respectively), mostly genes which are in the Cotyledons but not the Radicle (or vice versa). Compared to our compartment-specific lists, (which include the entire time course, not just the post-germination time point), we find there is less overlap with 91/452 and 76/432 genes that are 5-fold up-regulated in the Penfield dataset being tissue specific to our stringent definition (Supplemental Fig. S3B).

2. Microdissected Seed Development Dataset.

Taking the 87 .cel files published as part of the article on microdissected seeds (Le et al., 2010), [http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE15165], we renormalize the chips using the RMA normalization procedure and the custom CDF as in the Supplemental materials and methods. The resulting probe set distribution suggests that the noise region of the data is <3 (log_2 RMA, Supplemental Fig. S3C). To generate the tissue-specific lists we find genes which are 2-fold different between the two samples, having thresholded the data at 3, then perform a t-test on each gene found significant by this method (at a p-value of 0.05). As each tissue only has two replicates, a significant number of genes are called non-significant by this t-test and discarded. Comparing with the

Mature Green (MG) stage of the microdissected data from (Le et al., 2010), we find that some of the genes which are 2-fold higher expressed between our MCE or PE samples at 3, 16, 31 HAS are also 2-fold higher in the appropriate microdissected samples. Many of the genes which are specific in our time course are not expressed (no mean over 4) in the MG endosperm data (Le et al., 2010), and vice versa. Of the genes which are expressed in the seed development data, approximately 10% of the genes which are up-regulated in the time course are specific to the same part of the developing seed endosperm (Supplemental Fig. S3D). The data of (Le et al., 2010) separates the endosperm into Micropylar, Chalazal and Peripheral samples, and thus we need to make additional comparisons to compare with our data.

Similarly, when the END- and EMB- lists (without genes that are only expressed over 6 at 38 HAS) are investigated, we find 28 of the genes are also specific in the comparison between the microdissected Embryo and all three Endosperm samples (14 specific to the Endosperm and 14 specific to the Embryo, see Supplemental Fig. S3E), with only one gene (AT5G42200) which is higher in the Endosperm in the Mature Green sample but specific to the Embryo in our time course. These 28 genes are therefore specific to the Embryo/Endosperm in both our time course and the seed development data. If we compare our Endosperm/Embryo specific genelists with the comparisons between the Embryo and the three Endosperm samples, we find 51.4% of our Endosperm specific genes that are expressed in the developing seeds are higher in the Endosperm than Embryo in the developing seeds (Supplemental Fig. S3E). Lower overlap is seen for the Embryo specific list, but only a few genes are specific to the opposite tissue than in our data.

Confirmation of tissue specific gene expression by RT-qPCR (with Supplemental Fig. S4).
To confirm tissue specific expression found in the microarray data we performed gene expression analysis using RT-qPCR. Therefore seeds were isolated at the 31 HAS time point at which all seeds showed TR (Supplemental Fig. S4A) and dissected in the tissues described in Supplemental materials & methods (see also Fig. 1D) except that the MCE tissue was further dissected in the micropylar endosperm (ME) and the chalazal endosperm (CE). In total 20 genes were tested, the majority of which were either specific to the MCE or higher expressed in the MCE compared to the PE. Relative expression levels from microarray data and RT-qPCR data were compared (Supplemental Fig. S4C) and both analysis showed similar expression patterns and thereby confirmed the gene expression patterns found in the microarray dataset. Interestingly, the majority of the genes that are either specifically or highly expressed in the MCE tissue at 31 HAS are much more prominently expressed in the ME compared to the CE tissue (Supplemental Fig. S4C).

Correlation networks (with Supplemental Fig. S5 and S6).
To further investigate the topological features of these networks we have used TopoGSA (Glaab et al.,

2010), available at http://www.topogsa.net/. We computed four topological features and their distributions: node degree (number of connections to other nodes), length of the shortest paths (how far from all other nodes in the network a given node is), local clustering coefficient (how interconnected a group of nodes is to each other) and node betweenness centrality (how many network shortest paths go through a given node) (Supplemental Fig. S5). Clusters of special interest are identified by comparing the average distribution of a given topological feature (e.g. node degree) to the distribution found for that feature in the entire network. While almost 90% of the clusters have higher mean node degree (Supplemental Fig. S5A) than their respective networks and while cluster 1 in both networks is the most connected (with mean node degree 4 times greater than the network average) the clusters node degree distributions are noticeable different for RadNet and EndoNet with the latter's average lower than the former. The mean length of shortest paths is higher for RadNet than for EndoNet (Supplemental Fig. S5B). Over 35% of the RadNet clusters have an average shortest path length greater than their network's average while this figure is only 20% for EndoNet's clusters. Taking together the average node degree (Supplemental Fig. S5A) in both networks (and their clusters) with the average clustering coefficients for clusters in RadNet and EndoNet (Supplemental Fig. S5C) suggests that, although clusters are well-defined, they are not internally dense (average clustering coefficient is never higher than 0.7) and cluster members have many connections outside the clusters they belong to. Approximately half of the clusters in both networks have higher betweenness than their networks averages (see Supplemental Fig. S5D). Notably, within EndoNet, clusters 7, 14, 15, 17 and 19 have the top five largest betweenness centrality score (thus are the most important hubs). For RadNet the clusters with highest betweenness centrality are 15, 21, 22, 25, 30. GO classes overrepresented in these important hubs are depicted in Supplemental Fig. S5E,F.

The largest 30 clusters of the EndoNet network were investigated using overrepresentation analysis (ORA (Keller et al., 2008)), revealing cluster-specific overrepresentation of specific biological processes (Supplemental Fig. S6). For example, clusters 7 and 14 from EndoNet contain almost exclusively ribosome and translation related genes (Supplemental Fig. S6). Investigation of promoter elements in these clusters identified a strong enrichment of a telomere motif (TELO-box), a promoter element found in the Arabidopsis eEF1A (elongation factor) gene promoter and known to be present in numerous genes related to translation (Tremousaygue et al., 1999). Almost all connections in EndoNet clusters 7 and 14 (98% and 88% respectively) are also present in the RadNet (Fig. 3B), showing that genes related to the ribosome and translation are strongly co-expressed in both compartments. Despite this strong co-expression within both networks, the expression profile is different between the two compartments, being induced in both but subsequently repressed in the endosperm.

SUPPLEMENTAL MATERIAL & METHODS

Seed material. For this experiment the *Arabidopsis thaliana* accession Columbia-0 (Col-0, N60000) was used. Arabidopsis plants were grown on rockwool in a climate cell at 22°C and 70% humidity in a 16h light/8h dark cycle for seed production. Plants were watered with a Hyponex nutrient solution (1g/L, www.hyponex.co.jp). For germination and water content measurements, seeds were sown on 0.7% water agarose (Eurogentec) and incubated in a germination cabinet at 22°C with continuous lighting.

Water content measurements. To obtain the initial water content of the "dry" seeds, 5-7 mg of seeds were weighed on an AD-4 Autobalance (Perkin-Elmer). These seeds were dried in an oven at 104˚C for 17h (ISTA, 2009) and weighed again. To measure the water content of imbibed seeds a sample of weighed dry seeds were sown on 0.7% water agarose. After the indicated time points seeds were removed from the agarose plate and dried on filter paper to remove the access of water on the outside of the seeds. After that all seeds were collected and weighed on a balance and from this weight value (taking into account the initial dry weight) the water content was calculated on a dry weight basis.

Seed dissections and RNA isolation. After the indicated hours of imbibition seeds were harvested and dissected using forceps and a scalpel knife. To obtain the micropylar end of the endosperm the Arabidopsis seeds were dissected transversely (slightly out of the middle towards the micropylar end). This endosperm sample includes both the micropylar endosperm (endosperm layer over the radicle tip) as well as the chalazal endosperm (over the cotyledon tips). Therefore we call these samples the micropylar and chalazal endosperm (MCE). The remainder of the endosperm was sampled as peripheral endosperm (PE). Since the endosperm and seed coat are difficult to separate the endosperm was isolated including the seed coat tissue. Since the seed coat is a dead tissue we assumed that this does not interfere with endosperm transcriptome analysis. To obtain the embryo parts the seeds were carefully opened and the embryo was gently removed from the endosperm/seed coat tissue. To obtain the RAD sample the axis was cut just underneath the cotyledons meaning that the RAD sample includes the root tissue and the majority of the hypocotyl. Therefore this sample included the region that has been shown to elongate during germination (Sliwinska et al., 2009). The remainder of the embryo was collected as the COT sample which, next to the cotyledons, also consisted of remainder of the axis, i.e. the top of the hypocotyl and the shoot apical meristem. For the embryo parts approx. 100 seeds and for the endosperm sections 200 seeds were dissected per individual sample. Material was flash frozen in liquid nitrogen and ground in a dismembrator (Mikro-dismembrator U, B. Braun Biotech International) using stainless steel beads. For the isolation of RNA a commercial kit of Stratagene (Agilent Technologies, Absolutely RNA Nanoprep kit, 50 preps, cat# 400753) was used according to the manual. The only modification was the addition of polyvinylpolypyrrolidone (PVPP, 60mg/ml) to the extraction buffer for RNA extraction of endosperm samples, to inactivate phenolic

compounds present in the seed coat.

RNA concentration of the samples was measured using a Nanodrop ND1000 spectrophotometer. To assess quality and integrity of the RNA the samples were analyzed using the Shimadzu MultiNA and Agilent Bioanalyzer. In total 100ng of RNA was used to synthesize Biotin-labelled cRNA (using the Affymetrix 3" IVT-Express Labelling Kit) and the concentration and size of the cRNA was assessed. Denaturized cRNA was hybridized on the Affymetrix GeneChips Arabidopsis ATH1 Genome Array.

Normalization of microarray data. The raw .cel files were background corrected and normalized using the Robust Microarray Averaging (RMA) procedure (Irizarry et al., 2003), with a custom chip definition file (.cdf) from the CustomCDF project (Ath1121501_At_TAIRG.cdf v14.0.0, released 22nd March 2011 (Dai et al., 2005)), using the Bioconductor 'affy' package in the programming language R. This CDF maps the individual probes on the Affymetrix chip, using recent sequencing information contained in The Arabidopsis Information Repository (TAIR), with their corresponding genes. This eliminates the many-many relationship which exists between the Affymetrix probe sets and gene targets as is traditionally used. In particular this bijective mapping ensures that gene AGI codes may be used as the primary identifier in the correlation networks with no question of how to deal with multiple probe sets, with sometimes markedly differing behaviours, corresponding to the same gene. The resulting probe sets have varying numbers of probes with a minimum of three, although the majority of probe sets have the eleven probes from an original Affymetrix probe set. After removing the control probes, 21313 genes remain.

Fold Changes. Throughout this paper, when the fold changes are calculated, the data means are first clipped at level of 4 ($\log_2$) in expression level – replacing anything less than four with four. This may slightly underestimate the number of differentially expressed genes (or their level of fold change), but helps prevent the noise region, between 2 and ~4.5-5 in this case, from heavily influencing the results of the fold changes.

Differential Gene Expression. A gene is considered differentially expressed between two conditions if the difference between the condition means is sufficiently large (with the clipping as detailed above), and the values are statistically significant at a p-value of 0.05.

Comparison with the seed development data set (Le et al., 2010). The 20 Affymetrix GeneChip microarrays from the dataset (GEO Accession GSE680) (Le et al., 2010) were normalized using RMA with the CustomCDF in the same way as detailed above. The histogram of the normalized data suggests a noise level of 5, and so the means were clipped at 4. The means of the two replicates of the WT Cotyledon Stage and the WT Post-Mature Green stage (PMG) samples were analyzed. Genes with

a mean at least 5-fold higher in one condition were tested using a t-test for significance at a p-value of 0.05, with no False Discovery Rate applied. This process resulted in 907 genes which were higher expressed in the Cotyledon stage, and 602 which were higher in the later PMG stage. Without applying the t-test an additional 301 and 139 genes respectively would be considered expressed.

Comparison with the touch data set (Lee et al., 2005). In the original experiment by (Lee et al., 2005) for the response of leaves to touch or darkness, any genes which were not expressed in all conditions were ignored. We therefore obtained the original MAS normalized data (.cel files were not available) from the original authors, thresholded the data at 20 (not $\log_2$) and generated a list of genes which were both at least 2 fold differentially regulated and had a p-value of less than 0.05 in a t-test.

Correlation Networks. For both tissue types, endosperm (MCE and PE combined) and radicle (RAD), we filter the genes by keeping those probe sets which have at least one sample with mean expression (averaged over the four replicates) greater than or equal to 6. This means we only consider genes which have a significant amount of expression in at least one time point, both reducing the number of genes under consideration and removing those genes whose expression is noisy. This results in 11,525 and 11,645 expressed genes in the endosperm and RAD samples respectively. The two types of endosperm samples were combined to give more information into the Endosperm network, as they are very similar, whereas we decided not to combine the cotyledon samples with the radicle samples due to their significant expression as well as functional differences.

To identify interactions between the expressed genes the Pearson correlation coefficient between all pairs of genes is calculated. A cutoff, y, may then be applied to the resulting correlation matrix to produce a set of edges which are above this cutoff. In order to choose the value of this cutoff we calculate, for a range of y, the cumulative frequency of the edge degree of each node in the resulting graph. This may then be plotted, resulting in an approximately straight line in a log-log plot for many values of y (Supplemental Fig. S10). This suggests that the underlying network is obeying a power law distribution over several orders of magnitude, and over a wide range of number of edges.

The node degree distribution for a given cutoff y is approximately scale-free (Clauset et al., 2009), giving a straight line in a log-log plot of the node degree distribution. A simple log-log regression may be fitted to the log-transformed node degree distribution (excepting the degree 1 nodes), and the resulting adjusted r-squared value used as a measure of the linearity of the fit, for each value of y. For the MCE and PE samples (in the combined endosperm network, EndoNet) we choose a cutoff of $y = 0.932$, resulting in 577,846 edges. This choice balances the conflicting demands of the number of edges and the linearity of the power law fit, leading to an adjusted r-squared value of 0.986 (Supplemental Fig. S10). Using the RAD samples on their own to create a network (RadNet) results in

higher correlations due to fewer chips being included, so we choose a cutoff such that approximately the same number of edges as in the MCE network are retained, that is $y = 0.946$ and 586,746 edges, in order to make the resulting networks somewhat comparable. The adjusted r-squared value is 0.996. We note that these choices of cutoff are essentially arbitrary, and very similar networks are generated by increasing or decreasing the cutoff, with more/fewer edges in the order of 20,000-30,000 for each 0.01 the cutoff is adjusted by.

Once the cutoff has been determined, each correlation above that number is considered as an edge and a table of edges between nodes is exported into Cytoscape v2.8.1 (Shannon et al., 2003; Smoot et al., 2011), along with the correlation value between each edge. The yGraph Organic or the Edge-Weighted Force Directed Biolayout methods for arranging the nodes were then used to display the resulting networks.

From these correlation networks, the Cytoscape plug-in ClusterMaker (Morris et al., 2011) is used to partition the overall network into distinct clusters. In particular, the Transitivity Clustering method is used (Wittkop et al., 2010) (with parameters Max Subcluster Size =400, Max Time = 10, using the correlation values as an edge weight) to generate small, well-connected clusters in the network. These resulting clusters contain almost all of the possible edges between the nodes involved, ensuring that all the genes considered correlate well with each other. For example, cluster 1 in EndoNet contains 18862 edges between 195 genes, 99.7% of the possible edges, and therefore these genes therefore have very similar expression profiles.

Plotting the genes present in each cluster allows us to see that this method produces very similar looking clusters as expected, ensuring the genes involved do have very similar behaviour, avoiding the problem of an averaging clustering process like k-means. To determine how similar the correlations are between the two tissues, the edges in an EndoNet cluster are then investigated for correlation in the RAD samples (at varying levels of c), and *vice versa*. This gives a percentage expressing the similarity between these genes in the opposing tissue. In the 111 EndoNet clusters which contain at least 10 genes, 70 (63%) of them contain at least 70% of the corresponding edges at c=0.75 in the RAD samples, showing that many of the genes have correlate in the opposing tissue, although at a lower threshold. The clusters in which only a minority of edges exist in the opposing tissue are interesting, as they may either be tissue specific processes or contain genes which are only present in one tissue. Conversely, of the 95 RadNet clusters containing at least 10 genes, only 47 (49%) have at least 70% corresponding edges at c=0.75 in the endosperm samples.

Overrepresentation analysis. The gene lists of the 30 largest clusters of the EndoNet network (Supplemental Fig. S6) and the compartment specific gene sets (Supplemental Fig. S2) were analyzed

using Genetrial (Keller et al., 2008) (http://genetrail.bioinf.uni-sb.de) for GO categories that are overrepresented. For ORA of the gene classes overrepresented in time and tissue were analyzed using Pageman (Usadel et al., 2006; Sreenivasulu et al., 2008) (http://mapman.mpimp-golm.mpg.de/pageman/). A mapping file described by (Joosen et al., 2011) was used. The only modification was addition of a bin containing genes related to aging which was obtained from TAIR (www.arabidopsis.org). The results of Pageman analysis were summarized and redrawn in Fig. 5, Fig. 7 and Supplemental Fig. S7.

Phylotranscriptomic analysis. The determination of the evolutionary age of the *A. thaliana* protein coding genes was performed as described in (Quint et al., 2012). The resulting phylostratigraphic map is identical to (Quint et al., 2012) with one exception: as phylostratum (PS) 10 contained only 18 genes, PS 9 and 10 were fused. Hence, instead of 13 PS, the resulting phylostratigraphic map contains only 12 PS. Relative expression levels were computed as described previously (Domazet-Loso and Tautz, 2010). In brief, the mean expression level e_{js} of phylostratum j and developmental stage s was computed for each j and s as the arithmetic mean of expression levels e_{is} of all genes i belonging to phylostratum j. The mean expression levels e_{js} were linearly transformed to the interval [0,1] according to

$$f_{js} = \frac{e_{js} - e_{jmin}}{e_{jmax} - e_{jmin}}$$

where $e_{j\min}/e_{j\max}$ is the minimum/maximum mean expression level of phylostratum j over the seven developmental stages s. This linear transformation corresponds to a shift by $e_{j\min}$ and a subsequent shrinkage by $e_{j\max} - e_{j\min}$. As a result, the relative expression level f_{js} of developmental stage s with minimum e_{js} is 0, the relative expression level f_{js} of the developmental stage s with maximum e_{js} is 1, and the relative expression levels f_{js} of all other stages s range between 0 and 1, accordingly (Quint et al., 2012). Mean relative expression levels of genes in PS1-PS2, PS3-PS5 and PS6-PS12 were computed in each sampled developmental stage. Error bars represent the standard error of the relative expression levels in PS1-PS2, PS3-PS5 and PS6-PS12 in each developmental stage. Statistical significance of the differences between mean relative expression levels of different phylostrata classes was tested by one-way ANOVA.

RT-qPCR. For RT-qPCR, RNA was isolated from radicle, cotyledon and endosperm tissue used as indicated above. Seeds were dissected using forceps and a scalpel knife. For the radicle and cotyledon samples, material of approximately 300 seeds and for the endosperm samples material of approximately 1300 seeds were used. Genomic DNA was removed using a DNase treatment (RNase-free DNase set, Qiagen). Absence of DNA was checked by comparing cDNA samples with RNA

samples which were not reverse transcribed (minus RT control) and the difference was at least 5 Cq values as suggested (Nolan et al., 2006). RNA integrity of all samples was assessed by analysis on a 1% agarose gel. For all Arabidopsis samples clear ribosomal rRNA bands were visible and the OD 260/280 ratios (measured using a Nanodrop ND-1000, Nanodrop Technologies Inc.) were close to 2.0 for all samples used in this experiment.

cDNA synthesis, RT-qPCR conditions and primer design. RNA was reverse transcribed using the iScript^{TM} cDNA synthesis kit (Bio-Rad), with 500ng of total RNA being reverse transcribed according to the kit protocol. cDNA samples were diluted in a total volume of 360µl using sterile milliQ water. Per qPCR reaction 5µl sample, 12.5µl iQ SYBR Green Supermix (Bio-Rad), 0.5µl of primer (from a 10µM work solution) was added and supplemented with water to a final volume 25µl. The RT-qPCR reactions were run on a MyiQ (Bio-Rad). The qPCR program run consisted of a first step at 95°C for 3 min. and afterwards 40 cycles alternating between 15 sec. at 95°C and 1 min. at 60°C.

Primers for the target genes were designed preferably in the 3' part of the transcript. When possible the primer or primer pair was designed in such a way that it spanned an intron/exon border. The T_m of the primers was between 59 and 62°C. The primer sequences are described in Supplemental Table S2. Routinely a melting curve analysis was performed after the qPCR run (between 55°C and 95°C with 0.5°C increments for 10 sec. each) and for all primers a single peak was observed.

RT-qPCR data analysis. For analyzing our RT-qPCR data we used qbasePLUS (Hellemans et al., 2007) which is commercially available software (Biogazelle, Ghent, Belgium, www.biogazelle.com). For normalizing the data we mined our microarray data for stably expressed genes using a set that was recently tested for stable expression in seeds (Graeber et al., 2011; Dekkers et al., 2012). Six genes (AT1G13320, AT1G17210, AT2G28390, AT3G18780, AT4G34270 and AT5G25760) appeared to be stably expressed and their expression in our samples was confirmed using the geNORM program (Vandesompele et al., 2002), which is integrated into the qbasePLUS software. In the calculation we corrected for primer efficiency which was calculated from the amplification curve using LinReg PCR (Ramakers et al., 2003; Ruijter et al., 2009).

SUPPLEMENTAL FIGURES

Fig. S1. ATH1 Genechip quality assessment and reproducibility. A,B, All 116 ATH1 Genechip arrays showed similar patterns of raw probe intensity. Slide images were manually inspected, with no noticeable spatial artefacts. C, RNA degradation plot shows comparable slopes for all arrays. D, After RMA normalization (Irizarry et al., 2003) the data distributions become comparable, although lower median values are found for the dry and shortly imbibed seeds, which are samples that were isolated

from metabolically less active material. E, The histogram of the normalized data shows separated peaks for noise and signal, and the plot indicates a value of five (on a $\log_2$ scale) as being potentially expressed. F, The correlation between individual replicates are all above 0.980, with the majority (143 out of 174 comparisons) being over 0.990 (Supplemental Table S1). Six individual samples needed to be re-done, with RNA being isolated, labelled and hybridized at a later time. The correlation for these samples was slightly lower, but still above 0.980.

Fig. S2. General expression numbers and the identification and analysis of endosperm and embryo specific gene sets. A, Number of gene expressed (i.e. over 5 on $\log_2$ scale) in different tissues over the whole germination time course. The majority of the genes are shared by all seed compartments. B, Number of genes expressed increased during germination in all compartments. C, Small compartment specific gene sets were identified for the endosperm, embryo, MCE, PE, RAD and COT. D, Simplified reproduction of the ORA of the endosperm specific gene set. E, Simplified reproduction of the ORA of the embryo specific gene set. F, The endosperm and embryo specific gene sets are overrepresented for TFs. The table shows the TF classes and indicates the numbers of each family present on the chip, the number of expressed in the germination time course, and the number of TF genes expressed specifically in endosperm and embryo. p-value is calculated Chi-square test using a Yates correction. TF = transcription factor.

Fig. S3. Comparisons with two other seed microarray datasets. A, Histogram of the probe set values of the Penfield dataset (Penfield et al., 2006). B, Venn diagram showing the overlap between endosperm specific genes in our set at germination (MCE>RAD, 38 HAS ER) or over the whole time course compared the Penfield set (ENDO>EMB, using a 5 fold cutoff). C, Histogram of the probe set values of the Le dataset (Le et al., 2010). D, Table indicates overlap of expression in the endosperm between microdissected data at the post mature green stage and our set at 3, 16 and 31 HAS. E, Overlap of the endosperm and embryo specific sets from the germination time course compared to the microdissected seed development set (embryo and all three endosperm samples).

Fig. S4. RT-qPCR confirms tissue specific expression found in the microarray dataset. A, Indicates the different time points and stages that were sampled along the germination time course. B, The expression pattern of five example genes is depicted on pictograms that represent all 29 samples. Red indicates that the gene is expressed. C, The relative expression level in the different tissues at 31 HAS was calculated based on the microarray data. The seed compartment with the highest expression was set to 1 and indicated by the green colour and low expression was indicated by an orange to red colouring. Similarly the relative expression levels of the qPCR were depicted, with the micropylar (ME) and chalazal endosperm (CE) collected as separate samples. * = genes are part of the MCE specific gene list. Genes indicated in bold are also shown in B. HAS = hours after sowing

Fig. S5. Topological features of the EndoNet and RadNet. Four topological features were computed for both networks A, node degree; B, mean length of shortest paths; C, mean average clustering coefficients; and D, mean betweenness centrality score. The red line in both plots marks the mean value of the feature for entire network. We identified overrepresented GO classes for the five clusters with the highest mean betweenness centrality score (the most important hubs) in E, EndoNet and F, RadNet.

Fig. S6. Overrepresentation analysis of the 30 largest clusters from the EndoNet co-expression network. Clusters were grouped based on their expression pattern ('DOWN', 'UP and DOWN' or 'UP'). The graphs are divided in two parts and show the expression pattern of all genes in the cluster in both the endosperm (left side of each graph) as well as the expression pattern of the same set of genes in the RAD (right), see the schematic graph left of the legend for details. Clusters were analyzed by ORA using Genetrail. In total 25 out of 30 clusters showed overrepresented gene categories which are summarized underneath the graphs.

Fig. S7. ORA using Pageman of genes that are either higher expressed in the MCE or the RAD. Pageman analysis was comparing both tissues at each time point along the time course. Selected classes of the Pageman output were redrawn showing the most obvious differences between both tissues. Red colour indicates gene classes that are overrepresented while the blue colour indicates the underrepresented ones.

Fig. S8. Seed tissues differentiate during germination. A, The number of endosperm and embryo specific genes expressed increase along the germination time course. B, Graphs show the expression along the germination time course of exemplar genes related to stomatal development (Bergmann and Sack, 2007; Liu et al., 2010) and root development (Blilou et al., 2005; Overvoorde et al., 2010; Petricka et al., 2012) including examples of the core auxin biosynthetic pathway (Mashiguchi et al., 2011), auxin transport (Blakeslee et al., 2005) and auxin perception (Mockaitis and Estelle, 2008). The genes related to stomatal development were detected in the COT at 3, 16 and 31 HAS. The other genes were detected in the RAD throughout the whole time course (from 1 to 38 HAS).

Fig. S9. Expression of evolutionary old and young genes during Arabidopsis seed germination. A, The genes encoded on Arabidopsis genome are subdivided in 12 evolutionary age classes (phylostrata) depicted in a phylostratigrapic map. B,C,D, Mean relative expression in the MCE of PS1 and 2, PS3-5 and PS6-12 respectively. E,F,G, Mean relative expression in the RAD of PS1 and 2, PS3-5 and PS6-12 respectively.

Fig. S10. The node degree distribution for the correlation networks, showing power-law behaviour. A, A log-log cumulative frequency plot of the node degree distribution for the combined endosperm network, EndoNet and B, the radicle network, RadNet.

LITERATURE CITED

Bergmann DC, Sack FD (2007) Stomatal development. Annu Rev Plant Biol **58:** 163-181

Blakeslee JJ, Peer WA, Murphy AS (2005) Auxin transport. Curr Opin Plant Biol **8:** 494-500

Blilou I, Xu J, Wildwater M, Willemsen V, Paponov I, Friml J, Heidstra R, Aida M, Palme K, Scheres B (2005) The PIN auxin efflux facilitator network controls growth and patterning in Arabidopsis roots. Nature **433:** 39-44

Clauset A, Shalizi CR, Newman MEJ (2009) Power-Law Distributions in Empirical Data. Siam Review **51:** 661-703

Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res **33:** e175

Dekkers BJ, Willems L, Bassel GW, van Bolderen-Veldkamp RP, Ligterink W, Hilhorst HW, Bentsink L (2012) Identification of reference genes for RT-qPCR expression analysis in Arabidopsis and tomato seeds. Plant Cell Physiol **53:** 28-37

Domazet-Loso T, Tautz D (2010) A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature **468:** 815-818

Glaab E, Baudot A, Krasnogor N, Valencia A (2010) TopoGSA: network topological gene set analysis. Bioinformatics **26:** 1271-1272

Graeber K, Linkies A, Wood AT, Leubner-Metzger G (2011) A guideline to family-wide comparative state-of-the-art quantitative RT-PCR analysis exemplified with a Brassicaceae cross-species seed germination case study. Plant Cell **23:** 2045-2063

Hellemans J, Mortier G, De Paepe A, Speleman F, Vandesompele J (2007) qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. Genome Biol **8:** R19

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4:** 249-264

ISTA (2009) International rules for seed testing. International Seed Testing Association, Basserdorf

Joosen RVL, Ligterink W, Dekkers BJW, Hilhorst HWM (2011) Visualization of molecular processes associated with seed dormancy and germination using MapMan. Seed Science Research **21:** 143-152

Keller A, Backes C, Al-Awadhi M, Gerasch A, Kuntzer J, Kohlbacher O, Kaufmann M, Lenhof HP (2008) GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments. BMC Bioinformatics **9:** 552

Le BH, Cheng C, Bui AQ, Wagmaister JA, Henry KF, Pelletier J, Kwong L, Belmonte M, Kirkbride R, Horvath S, Drews GN, Fischer RL, Okamuro JK, Harada JJ, Goldberg RB (2010) Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. Proc Natl Acad Sci U S A **107:** 8063-8070

Lee D, Polisensky DH, Braam J (2005) Genome-wide identification of touch- and darkness-regulated Arabidopsis genes: a focus on calmodulin-like and XTH genes. New Phytol **165:** 429-444

Liu YK, Liu YB, Zhang MY, Li DQ (2010) Stomatal development and movement: the roles of MAPK signaling. Plant Signal Behav **5:** 1176-1180

Mashiguchi K, Tanaka K, Sakai T, Sugawara S, Kawaide H, Natsume M, Hanada A, Yaeno T, Shirasu K, Yao H, McSteen P, Zhao Y, Hayashi K, Kamiya Y, Kasahara H (2011) The main auxin biosynthesis pathway in Arabidopsis. Proc Natl Acad Sci U S A **108:** 18512-18517

Mockaitis K, Estelle M (2008) Auxin receptors and plant development: a new signaling paradigm. Annu Rev Cell Dev Biol **24:** 55-80

Morris JH, Apeltsin L, Newman AM, Baumbach J, Wittkop T, Su G, Bader GD, Ferrin TE (2011) clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics **12:** 436

Nolan T, Hands RE, Bustin SA (2006) Quantification of mRNA using real-time RT-PCR. Nat Protoc **1:** 1559-1582

Overvoorde P, Fukaki H, Beeckman T (2010) Auxin control of root development. Cold Spring Harb Perspect Biol **2:** a001537

Penfield S, Li Y, Gilday AD, Graham S, Graham IA (2006) Arabidopsis ABA INSENSITIVE4 regulates lipid mobilization in the embryo and reveals repression of seed germination by the endosperm. Plant Cell **18:** 1887-1899

Petricka JJ, Winter CM, Benfey PN (2012) Control of Arabidopsis root development. Annu Rev Plant Biol **63:** 563-590

Quint M, Drost HG, Gabel A, Ullrich KK, Bonn M, Grosse I (2012) A transcriptomic hourglass in plant embryogenesis. Nature **490:** 98-101

Ramakers C, Ruijter JM, Deprez RH, Moorman AF (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci Lett **339:** 62-66

Ruijter JM, Ramakers C, Hoogaars WM, Karlen Y, Bakker O, van den Hoff MJ, Moorman AF (2009) Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. Nucleic Acids Res **37:** e45

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

Sliwinska E, Bassel GW, Bewley JD (2009) Germination of Arabidopsis thaliana seeds is not completed as a result of elongation of the radicle but of the adjacent transition zone and lower hypocotyl. J Exp Bot **60:** 3587-3594

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics **27:** 431-432

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

Tremousaygue D, Manevski A, Bardet C, Lescure N, Lescure B (1999) Plant interstitial telomere motifs participate in the control of gene expression in root meristems. Plant J **20:** 553-561

Usadel B, Nagel A, Steinhauser 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

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol **3:** RESEARCH0034

Wittkop T, Emig D, Lange S, Rahmann S, Albrecht M, Morris JH, Bocker S, Stoye J, Baumbach J (2010) Partitioning biological data with transitivity clustering. Nat Methods **7:** 419-420

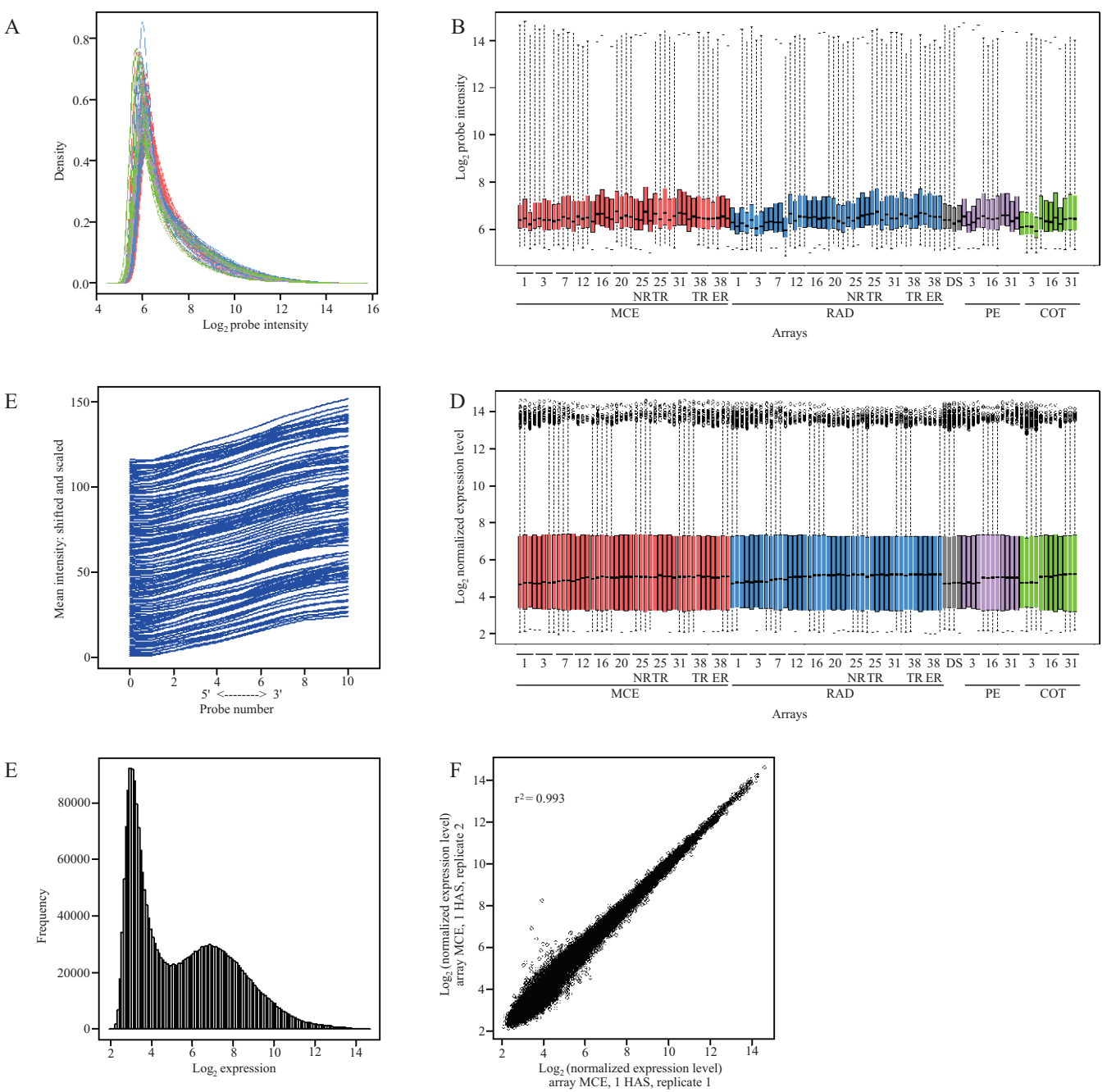
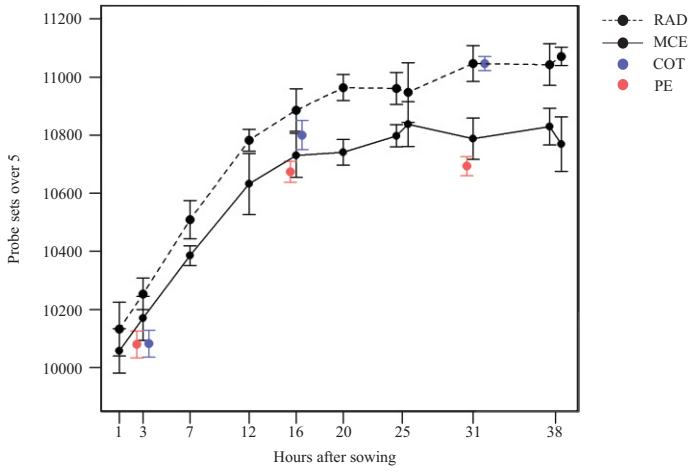


Fig. S1. ATH1 Genechip quality assessment and reproducibility. A,B, All 116 ATH1 Genechip arrays showed similar patterns of raw probe intensity. Slide images were manually inspected, with no noticeable spatial artefacts. C, RNA degradation plot shows comparable slopes for all arrays. D, After RMA normalization (Irizarry et al., 2003) the data distributions become comparable, although lower median values are found for the dry and shortly imbibed seeds, which are samples that were isolated from metabolically less active material. E, The histogram of the normalized data shows separated peaks for noise and signal, and the plot indicates a value of five (on a log2 scale) as being potentially expressed. F, The correlation between individual replicates are all above 0.980, with the majority (143 out of 174 comparisons) being over 0.990 (Supplemental Table S1). Six individual samples needed to be re-done, with RNA being isolated, labelled and hybridized at a later time. The correlation for these samples was slightly lower, but still above 0.980.

A

Number of genes expressed (>5 log) in:											
At least one sample	Dry seeds	Embryo	RAD	COT	Endosperm	MCE	PE	At least once in Endo+Emb	All tissues (excluding dry seeds)	All tissues (including dry seeds)	All 29 samples
14,317	10,060	13,617	13,279	12,346	13,147	13,085	12,282	12,457	11,298	9,661	6,867

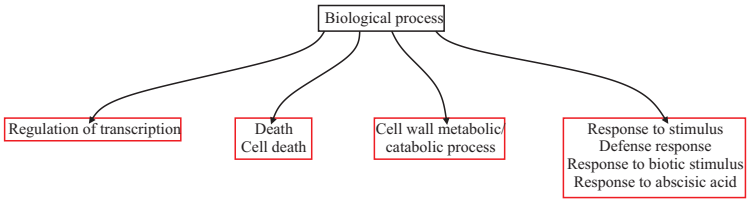
B



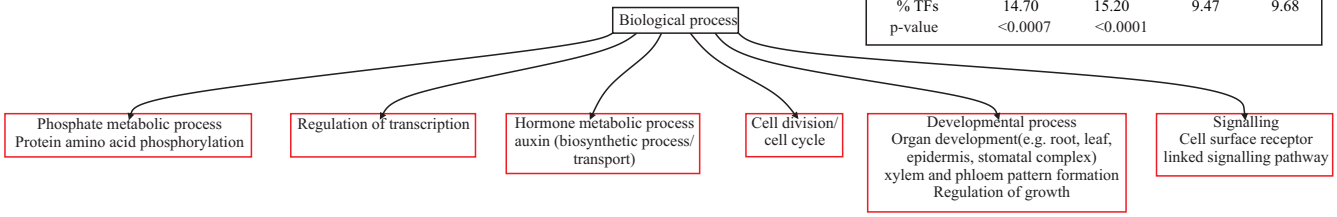
C

Sample	Expressed (>6 log):	Number of genes	Percentage
ALL	In either endosperm or embryo	12,856	100%
	In both endosperm and embryo	10,801	84.01%
ENDOSPERM	In endosperm	11,606	90.27%
	Endosperm specific	415	3.22%
	In MCE	11,525	89.65%
	MCE specific	21	0.16%
	In PE	10,660	82.92%
	PE specific	2	0.01%
EMBRYO	In embryo	12,051	93.73%
	Embryo specific	546	4.24%
	In RAD	11,645	90.58%
	RAD specific	106	0.82%
	In COT	10,664	82.95%
	COT specific	47	0.37%

D



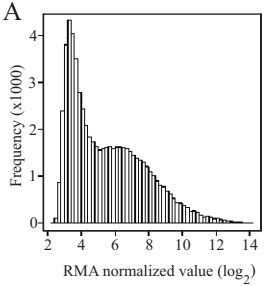
E

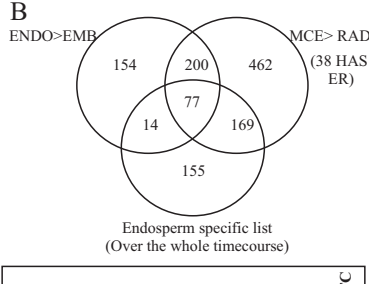


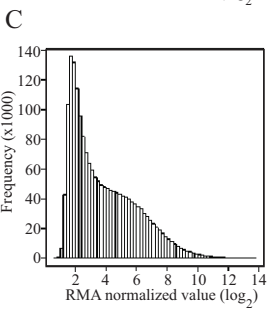
F

TF class	Number of TF expressed: Specifically in endosperm	Specifically in embryo	In total	# of TF present on chip
ABI3VP1	0	1	11	44
Alfin-like	0	0	7	7
AP2-EREBP	3	6	83	128
ARF	1	1	15	18
ARID	0	0	6	9
ARR-B	0	0	5	10
AUX/IAA	0	2	17	28
BBR/BPC	0	0	5	6
BES1	0	0	6	6
bHLH	1	10	61	113
BSD	0	0	6	7
bZIP	1	4	43	67
C2C2-CO-like	2	0	10	15
C2C2-Dof	2	3	11	32
C2C2-GATA	0	2	17	20
C2C2-YABBY	0	1	3	5
C2H2	3	6	83	145
C3H	9	3	128	172
CAMTA	0	0	8	8
CCAAT	0	1	28	39
CPP	0	0	3	6
CSD	0	0	4	4
DBP	0	0	2	4
DDT	0	0	3	3
E2F-DP	0	1	6	7
EIL	0	0	5	6
FAR1	0	0	12	15
FHA	0	0	8	13
G2-like	0	5	20	34
GeBP	0	0	5	10
GNAT	0	0	22	27
GRAS	0	1	23	31
GRF	0	0	6	7
HB	1	10	42	84
HMG	0	3	11	11
HSF	0	1	15	22
Jumonji	0	0	14	15
LIM	0	0	5	7
LOB	3	3	12	26
MADS	0	0	9	73
mTERF	0	1	19	25
MYB	6	5	40	141
MYB-related	2	1	42	55
NAC	15	0	43	93
OFP	0	0	0	9
Orphans	3	3	32	63
PHD	0	1	32	43
PLATZ	0	0	5	7
Pseudo	0	0	3	4
RWP-RK	0	0	7	11
SBP	0	0	5	14
SET	0	1	22	28
Sigma70-like	0	1	5	6
Smalls	0	0	26	35
SNF2	1	2	25	37
SRS	0	1	2	6
SWI/SNF BAF60b	0	0	10	14
SWI/SNF-SWI3	0	0	0	5
TAZ	0	0	5	7
TCP	0	1	13	20
Tify	0	0	12	13
TRAF	0	0	14	18
Trihelix	1	1	23	38
TUB	0	0	8	11
WRKY	7	1	37	63
zf-HD	0	0	9	14
Total TFs	61	83	1219	2064
Total	415	546	12865	21313
% TFs	14.70	15.20	9.47	9.68
p-value	<0.0007	<0.0001		

Fig. S2. General expression numbers and the identification and analysis of endosperm and embryo specific gene sets. A, Number of gene expressed (i.e. over 5 on log2 scale) in different tissues over the whole germination time course. The majority of the genes are shared by all seed compartments. B, Number of genes expressed increased during germination in all compartments. C, Small compartment specific gene sets were identified for the endosperm, embryo, MCE, PE, RAD and COT. D, Simplified reproduction of the ORA of the endosperm specific gene set. E, Simplified reproduction of the ORA of the embryo specific gene set. F, The endosperm and embryo specific gene sets are overrepresented for TFs. The table shows the TF classes and indicates the numbers of each family present on the chip, the number of expressed in the germination time course, and the number of TF genes expressed specifically in endosperm and embryo. p-value is calculated Chi-square test using a Yates correction. TF = transcription factor.







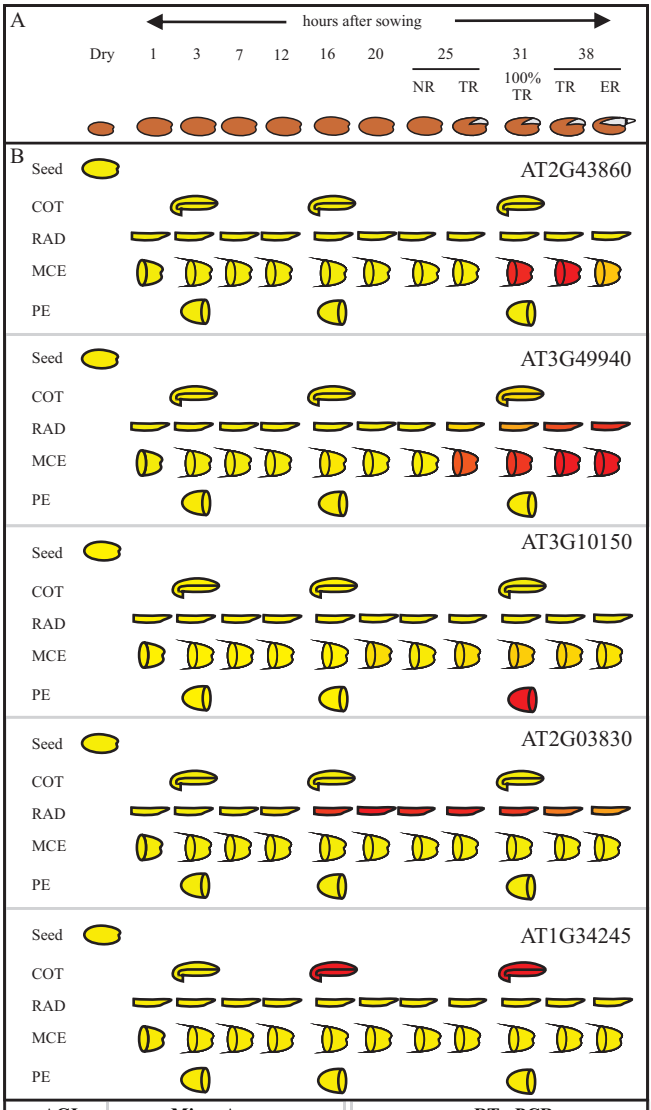
D

	ME > PE	CE > PE	Total ME&CE Up	PE > ME	PE > CE	Total PE Up	Expressed in SD	Expressed in GERM TC
MCE3 > PE3	2	2	4	0	0	0	19	48
MCE16 > PE16	4	1	5	0	4	4	57	153
MCE31 > PE31	11	7	18	0	13	13	180	468
PE3 > MCE3	0	0	0	1	0	1	1	9
PE16 > MCE16	0	0	0	0	2	2	31	53
PE31 > MCE31	0	0	0	2	5	6	53	116
Total	78	193	266	46	228	270	9823	

E

	ME > EMB	PE > EMB	CE > EMB	Total END Up	EMB > ME	EMB > PE	EMB > CE	Total EMB Up	All END > EMB	EMB > All END	Expressed in SD	Total
END specific (except 38HAS)	27	31	32	52	5	1	3	6	14	0	101	264
EMB specific (except 38HAS)	4	3	3	8	23	22	31	38	1	14	121	309
Total	460	441	485	908	385	326	496	805	125	98	10423	21313

Fig. S3. Comparisons with two other seed microarray datasets. A, Histogram of the probe set values of the Penfield dataset (Penfield et al., 2006). B, Venn diagram showing the overlap between endosperm specific genes in our set at germination (MCE>RAD, 38 HAS ER) or over the whole time course compared the Penfield set (ENDO>EMB, using a 5 fold cutoff). C, Histogram of the probe set values of the Le et al. dataset (Le et al., 2010). D, Table indicates overlap of expression in the endosperm between microdissected data at the post mature green stage and our set at 3, 16 and 31 HAS. E, Overlap of the endosperm and embryo specific sets from the germination time course compared to the microdissected seed development set (embryo and all three endosperm samples).

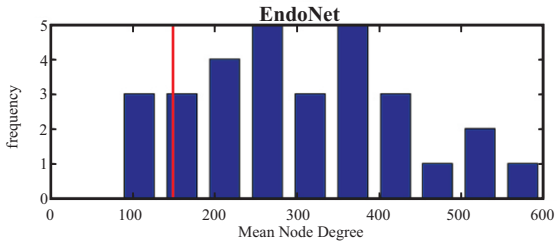


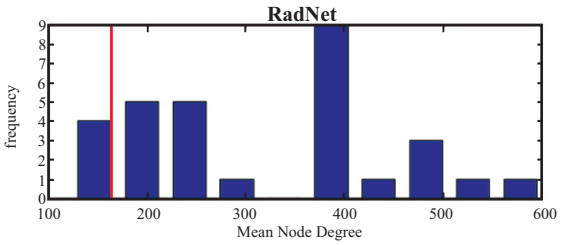
C

AGI	Micro Array				RT-qPCR				
	MCE	PE	RAD	COT	ME	CE	PE	RAD	COT
AT1G70690*	1	0.16	0.13	0.13	1	0.08	0.09	0.00	0.00
AT2G30670*	1	0.21	0.21	0.21	1	0.02	0.01	0.00	0.00
AT2G38320*	1	0.20	0.20	0.20	1	0.10	0.04	0.04	0.04
AT2G43860*	1	0.02	0.01	0.01	1	0.05	0.00	0.00	0.00
AT2G45420*	1	0.05	0.05	0.05	1	0.07	0.02	0.00	0.00
AT3G02240*	1	0.06	0.06	0.06	1	0.03	0.01	0.00	0.00
AT3G11870*	1	0.15	0.18	0.15	1	0.10	0.02	0.15	0.00
AT3G22060*	1	0.10	0.10	0.10	1	0.05	0.03	0.03	0.00
AT4G22470*	1	0.18	0.09	0.09	1	0.07	0.12	0.02	0.00
AT4G35060*	1	0.22	0.22	0.22	1	0.13	0.07	0.07	0.00
AT5G08480*	1	0.17	0.16	0.16	1	0.20	0.18	0.05	0.08
AT4G30140*	1	0.06	0.06	0.06	1	0.05	0.00	0.00	0.00
AT1G29280	1	0.14	0.54	0.06	1	0.09	0.06	0.41	0.00
AT1G79580	1	0.13	0.41	0.13	1	0.06	0.00	0.21	0.00
AT2G18980	1	0.02	0.01	0.01	1	0.05	0.01	0.00	0.00
AT3G49940	1	0.13	0.42	0.19	1	0.10	0.09	0.51	0.29
AT1G28290	0.29	0.02	1.00	0.05	0.41	0.02	0.01	1	0.04
AT3G10150	0.26	1.00	0.17	0.17	0.13	0.39	1	0.00	0.00
AT2G03830	0.16	0.16	1.00	0.16	0.01	0.03	0.03	1	0.01
AT1G34245	0.18	0.18	0.18	1.00	0.04	0.11	0.13	0.05	1

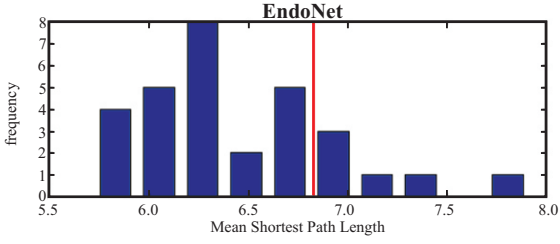
Fig. S4. RT-qPCR confirms tissue specific expression found in the microarray dataset. A, Indicates the different time points and stages that were sampled along the germination time course. B, The expression pattern of five example genes is depicted on pictograms that represent all 29 samples. Red indicates that the gene is expressed. C, The relative expression level in the different tissues at 31 HAS was calculated based on the microarray data. The seed compartment with the highest expression was set to 1 and indicated by the green colour and low expression was indicated by an orange to red colouring. Similarly the relative expression levels of the qPCR were depicted, with the micropylar (ME) and chalazal endosperm (CE) collected as separate samples. * = genes are part of the MCE specific gene list. Genes indicated in bold are also shown in B. HAS = hours after sowing

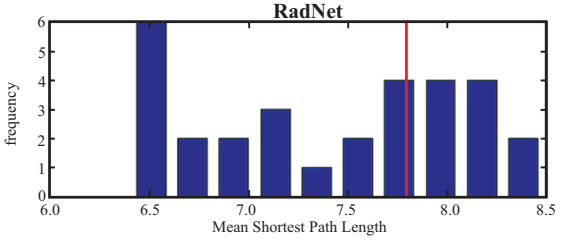
A



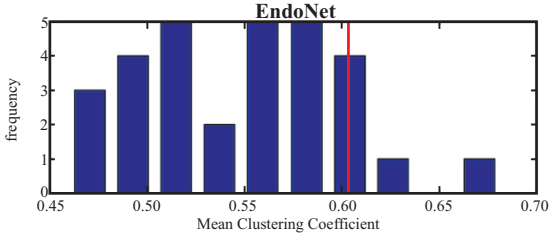


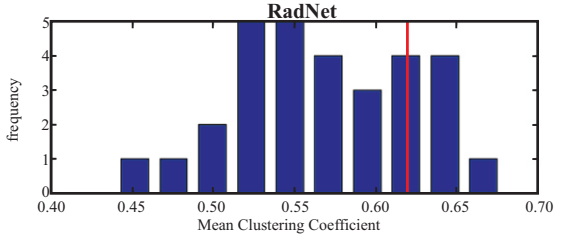
B



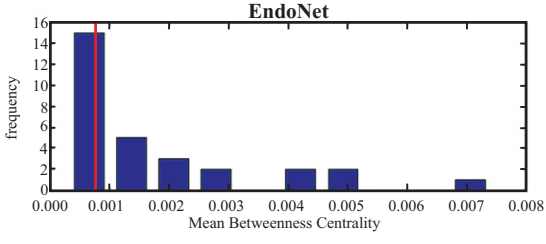


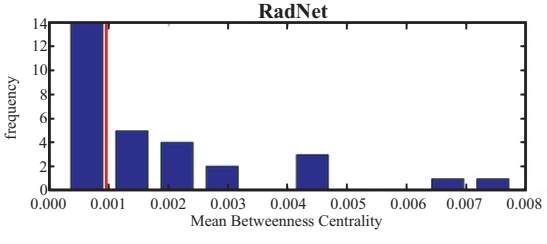
C





D





E

Cluster	Size	Examples of overrepresented GO classes
EndoNet 7	82	ribosome, translation, nucleus, cell wall
EndoNet 14	51	ribosome, translation, nucleus, cell wall, mitochondrion
EndoNet 15	48	nucleus, DNA-directed RNA polymerase complex, nitrogen compound metabolic process
EndoNet 17	44	ribosome, ncRNA processing, RNA metabolic process, nucleus
EndoNet 19	39	ribosome, translation, unfolded protein binding, response to metal ion, ATP binding

F

Cluster	Size	Examples of overrepresented GO classes
RadNet 15	48	ribosome, translation, nucleus, mitochondrion, ncRNA metabolic process, nitrogen compound metabolic process, nucleic acid binding
RadNet 21	32	-
RadNet 22	31	DNA-directed RNA polymerase complex, nucleus, ATPase activity coupled
RadNet 25	29	RNA binding, methylation, cellular nitrogen compound metabolic process
RadNet 30	23	exonuclease activity, heterotrimeric G-protein complex

Fig. S5. Topological features of the EndoNet and RadNet. Four topological features were computed for both networks A, node degree; B, mean length of shortest paths; C, mean average clustering coefficients; and D, mean betweenness centrality score. The red line in both plots marks the mean value of the feature for entire network. We identified overrepresented GO classes for the five clusters with the highest mean betweenness centrality score (the most important hubs) in E, EndoNet and F, RadNet.

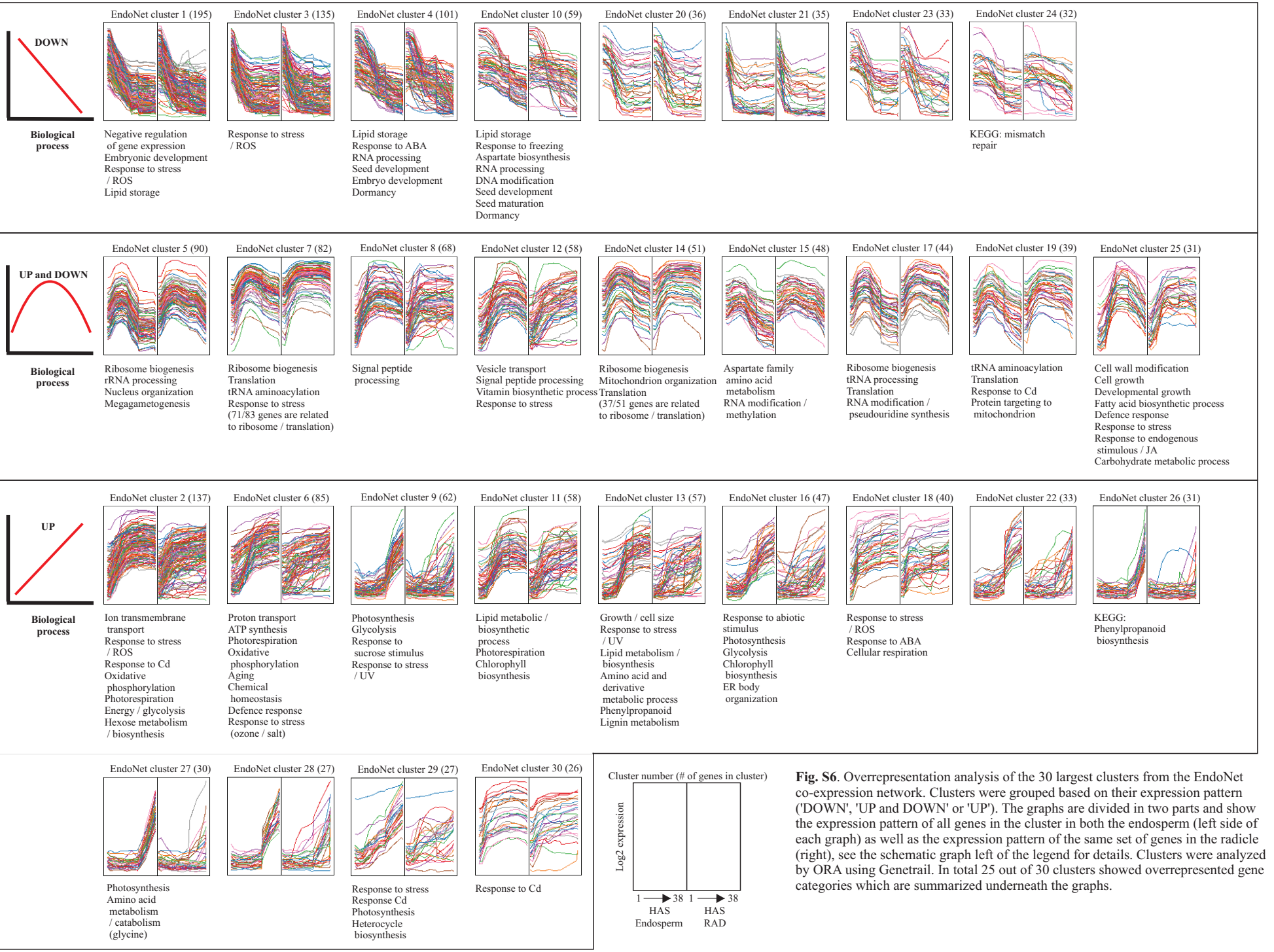


Fig. S6. Overrepresentation analysis of the 30 largest clusters from the EndoNet co-expression network. Clusters were grouped based on their expression pattern ('DOWN', 'UP and DOWN' or 'UP'). The graphs are divided in two parts and show the expression pattern of all genes in the cluster in both the endosperm (left side of each graph) as well as the expression pattern of the same set of genes in the radicle (right), see the schematic graph left of the legend for details. Clusters were analyzed by ORA using Genetrail. In total 25 out of 30 clusters showed overrepresented gene categories which are summarized underneath the graphs.

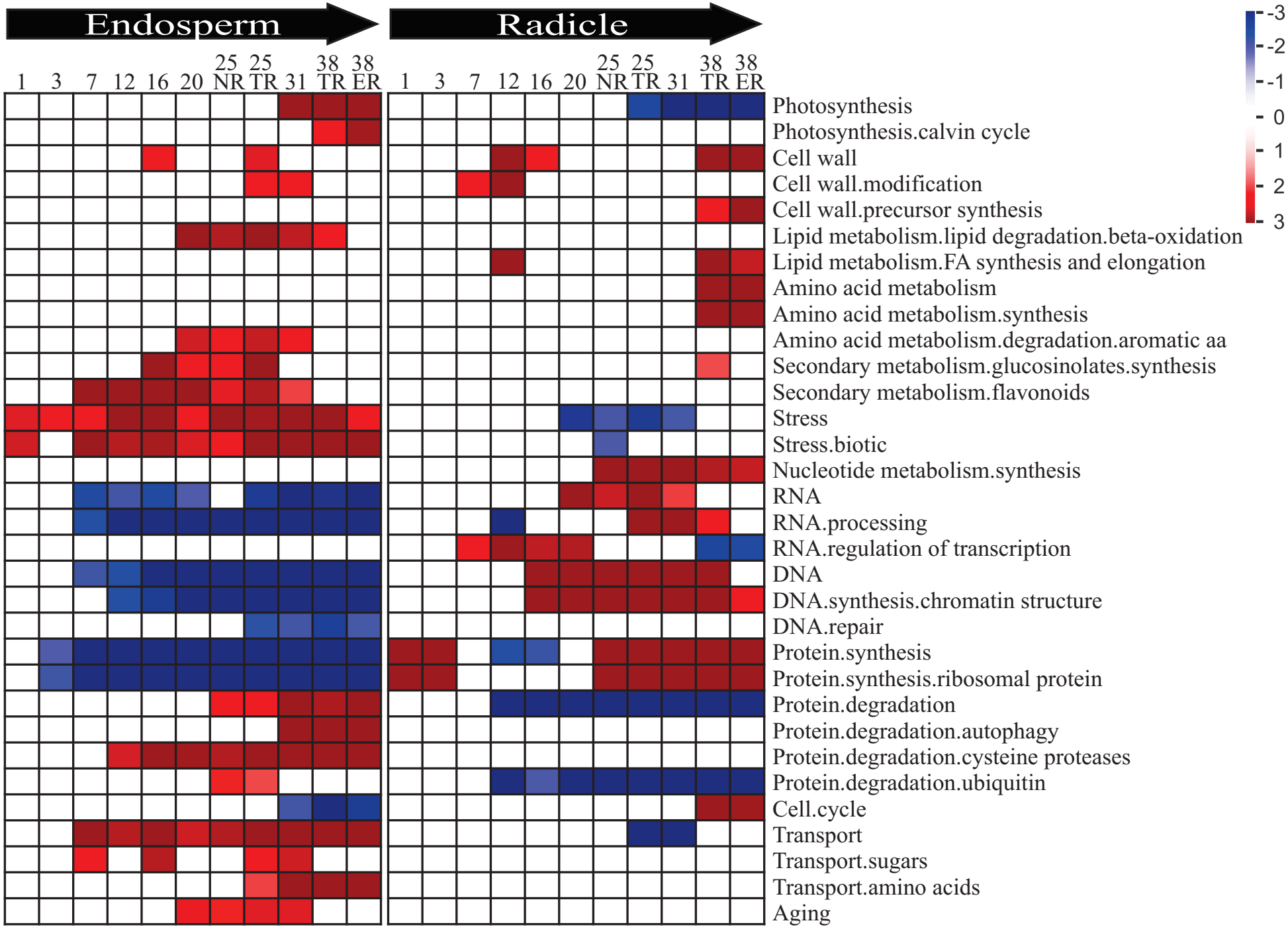


Fig. S7. ORA using Pageman of genes that are either higher expressed in the MCE or the RAD. Pageman analysis was comparing both tissues at each time point along the time course. Selected classes of the Pageman output were redrawn showing the most obvious differences between both tissues. Red colour indicates gene classes that are overrepresented while the blue colour indicates the underrepresented ones.

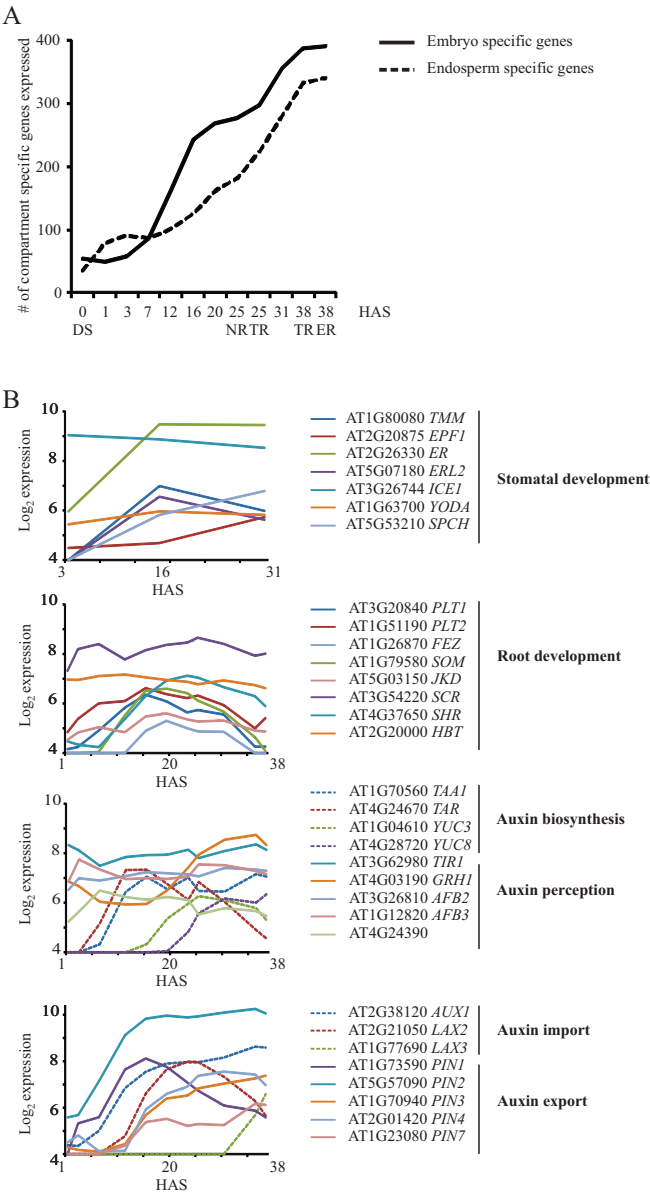


Fig. S8. Seed tissues differentiate during germination. A, The number of endosperm and embryo specific genes expressed increase along the germination time course. B, Graphs show the expression along the germination time course of exemplar genes related to stomatal development (Bergmann and Sack, 2007; Liu et al., 2010) and root development (Blilou et al., 2005; Overvoorde et al., 2010; Petricka et al., 2012) including examples of the core auxin biosynthetic pathway (Mashiguchi et al., 2011), auxin transport (Blakeslee et al., 2005) and auxin perception (Mockaitis and Estelle, 2008). The genes related to stomatal development were detected in the COT at 3, 16 and 31 HAS. The other genes were detected in the RAD throughout the whole time course (from 1 to 38 HAS).

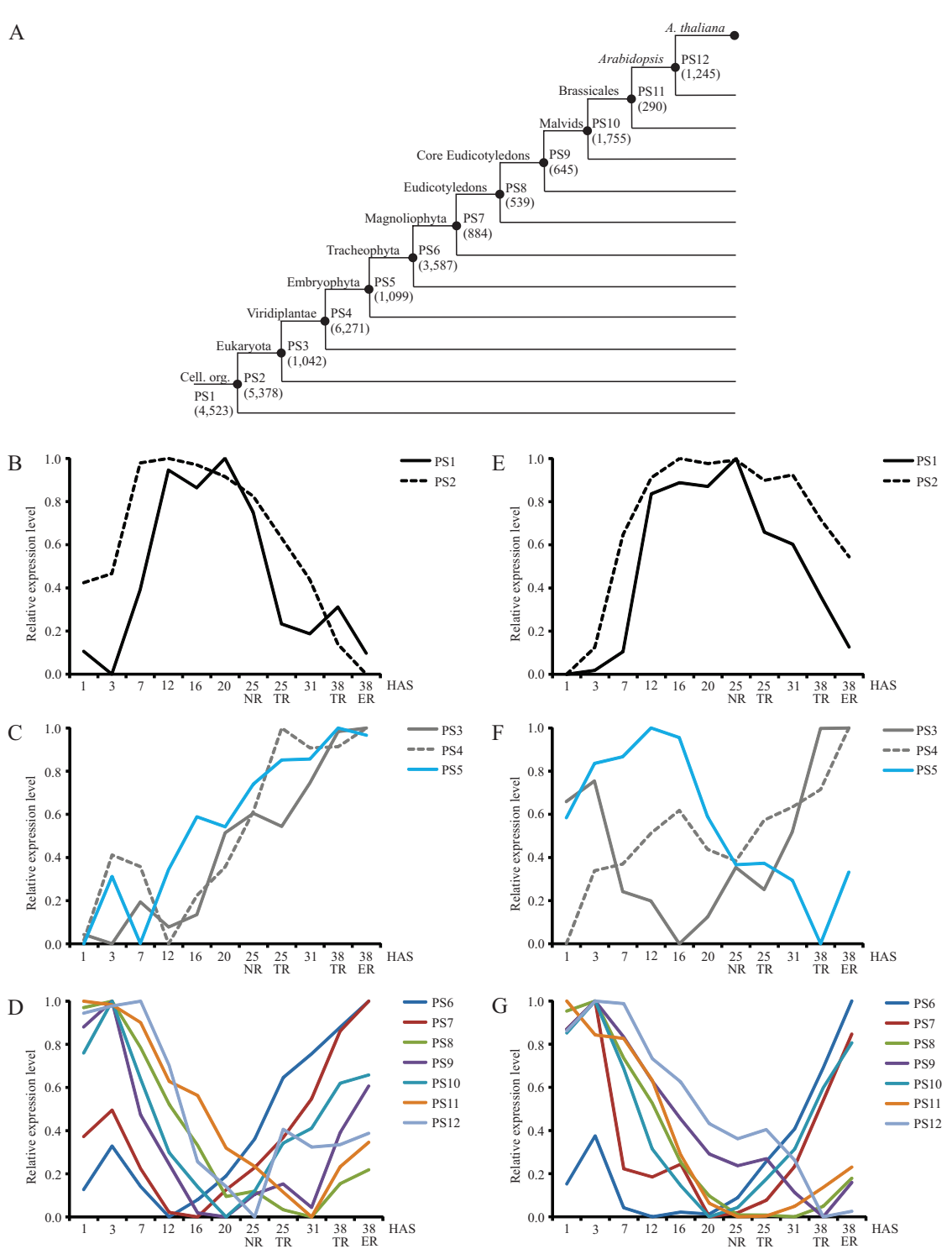


Fig. S9. Expression of evolutionary old and young genes during Arabidopsis seed germination. A, The genes encoded on Arabidopsis genome are subdivided in 12 evolutionary age classes (phylostrata) depicted in a phylostratigrapic map. B,C,D, Mean relative expression in the MCE of PS1 and 2, PS3-5 and PS6-12 respectively. E,F,G, Mean relative expression in the RAD of PS1 and 2, PS3-5 and PS6-12 respectively.

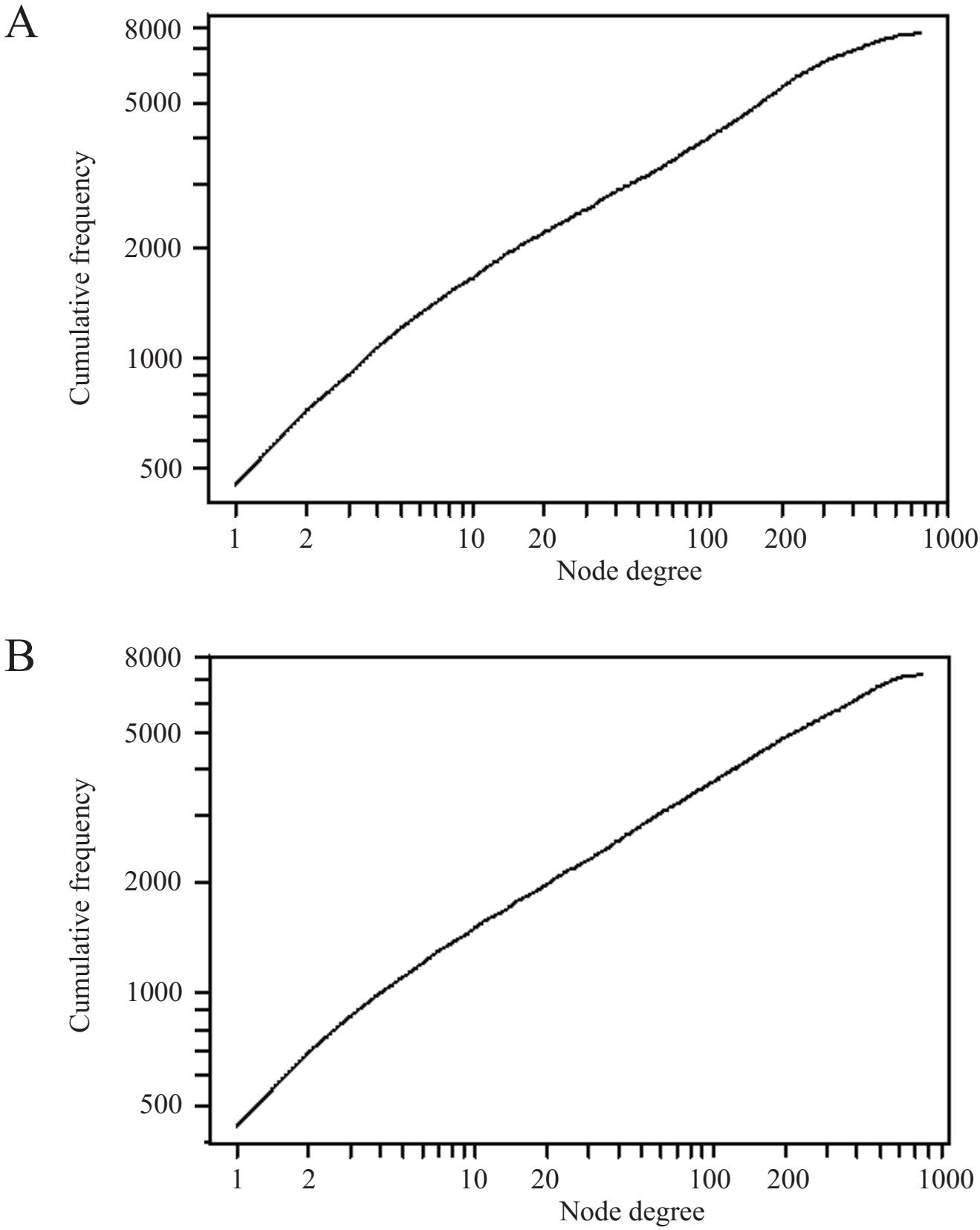


Fig. S10. The node degree distribution for the correlation networks, showing power-law behaviour. A, A log-log cumulative frequency plot of the node degree distribution for the combined endosperm network, EndoNet and B, the radicle network, RadNet.

SUPPLEMENTAL TABLES

Table S1. Correlations between the sample replicates. Values in italics indicate correlations involving samples which were redone.

	Correlation (r^2) between replicates:					
	(1,2)	**(1,3)**	**(1,4)**	**(2,3)**	**(2,4)**	**(3,4)**
DRY SEEDS	0.993	0.992	0.992	0.991	0.992	0.994
MCE1	0.993	0.988	0.992	0.990	0.994	0.990
MCE3	0.992	0.992	0.991	0.995	0.994	0.995
MCE7	*0.981*	0.992	0.992	*0.990*	*0.990*	0.996
MCE12	*0.987*	0.992	0.992	*0.991*	*0.991*	0.995
MCE16	0.996	0.991	0.990	0.993	0.992	0.997
MCE20	0.995	0.994	0.995	0.995	0.995	0.997
MCE25NR	*0.985*	*0.986*	*0.988*	0.994	0.994	0.994
MCE25TR	0.993	0.995	0.993	0.995	0.995	0.995
MCE31	0.992	0.994	*0.989*	0.993	*0.982*	*0.992*
MCE38TR	0.993	0.992	0.992	0.994	0.994	0.993
MCE38ER	0.992	0.992	0.988	0.995	0.992	0.993
RAD1	0.989	0.990	0.985	0.991	0.986	0.987
RAD3	0.987	0.991	0.991	0.990	0.988	0.992
RAD7	0.992	*0.980*	0.991	*0.985*	0.994	*0.985*
RAD12	0.991	0.991	0.990	0.994	0.994	0.996
RAD16	0.994	0.993	0.993	0.994	0.996	0.993
RAD20	0.996	0.993	0.993	0.994	0.994	0.993
RAD25NR	0.992	0.992	0.992	0.995	0.994	0.996
RAD25TR	0.993	0.994	0.992	0.996	0.994	0.995
RAD31	0.993	0.993	0.995	0.994	0.995	0.996
RAD38TR	0.992	0.992	0.991	0.995	0.994	0.995
RAD38ER	0.993	0.994	0.992	0.995	0.993	0.994
PE3	0.990	0.992	0.993	0.990	0.990	0.994
PE16	0.996	0.994	0.991	0.996	0.993	0.996
PE31	0.991	0.994	0.993	0.991	0.989	0.995
COT3	*0.987*	0.990	0.990	*0.987*	*0.989*	0.989
COT16	0.995	0.996	0.995	0.994	0.996	0.994
COT31	0.990	0.992	0.990	0.995	0.996	0.997

Table S2. Primer information of the genes tested by RT-qPCR. For = forward, rev = reverse, I/E = intron/exon border.

AGI		Primer sequence	Primer length (bp)	T_m	Fragment length (bp)	Over I/E
AT1G70690	for	GTAGAAAATTACAAGATGAGAAATGGTG	28	61	91	yes
	rev	AACTAACAGTGTGGCTCGAC	20	61		
AT2G30670	for	TGGCTCAACCTTTTTTCAAAGAC	23	62	95	yes
	rev	TGATGAAACCTCATTTGGCTCTC	23	62		
AT2G38320	for	ATGAAAGGGCTGAAGAATGGG	21	62	98	yes
	rev	ATCTGACCCTCTTCCAGTATATCC	24	62		
AT2G43860	for	GCCCTAAACAGGAATCAGGAG	21	62	93	yes
	rev	AATCTAACAAAACCGCCACTTC	22	61		
AT2G45420	for	CAAGCATTAGCACATGAACTC	21	59	123	no
	rev	TAACAATGCGGAGATGTATCG	21	59		
AT3G02240	for	TGAAACCGATCAATCCAACAAAG	23	61	97	no
	rev	TCCCTTAGCTCAACATATACTTCTC	25	61		
AT3G11870	for	CCACAGAGAGATACTTGATGATCC	24	62	106	yes
	rev	TTCCATCATCAAATTCGGGAAAC	23	61		
AT4G22060	for	GGTGGGAGCTGTAACTTTAGG	21	62	111	no
	rev	TAATCACAAATTAAGACCGTGTTTGG	26	62		
AT4G22470	for	TTAGCTGCCCATGATGACC	19	61	81	no
	rev	AGGCTGTCTTTGTATATTTAAGACTAG	27	60		
AT4G35060	for	CTAGCTCCACTGAGGTTCG	19	61	75	no
	rev	AACATCACATAACAACACAAGC	22	59		
AT5G08480	for	CATCGAAGAAGAGGAGAAAGC	21	60	77	no
	rev	TATACCCCGGTTTAGACCTTG	21	60		
AT4G30140	for	TGCATACTACGGAAGTGATAAATAC	25	60	79	no
	rev	GATATGTTGAACGGACTAGCT	21	60		
AT1G29280	for	GACGAGTTCGCATGGTTTAC	18	61	120	no
	rev	AAAGAACACCGCCACGTC	20	62		
AT1G79580	for	CTTGGGAAACGACGCTAATC	20	60	61	no
	rev	GATTCAACAAATCCGATTGATCTAC	25	60		
AT2G18980	for	TCACTAATTAACACGTGAACAATCC	25	61	66	no
	rev	CGGGTTGGAGTTAAAACCG	19	60		
AT3G49940	for	GGCGTCGTTTTTGTTATGAGG	21	62	115	no
	rev	TAAATCACGTAATTTGCTACATTTCAC	27	60		
AT1G28290	for	AACCCATCGAAGGTGCTAC	19	61	79	yes
	rev	TTCTTGTCTGTCGTCGTCTC	20	61		
AT3G10150	for	GATCCTCCGTTAAGGCAGTG	20	62	81	yes
	rev	AAAGCCAGAGTTTATCTTTGTACG	24	61		
AT2G03830	for	TCACTTCTTCAGGAAAATCTAAGG	24	60	90	yes
	rev	CCATCGATAGATATTTCGCTAGAG	24	60		
AT1G34245	for	GCACCACAAGAAGGAAATAAGC	22	62	78	yes
	rev	CGTATGAACAATCCGGTAAGC	21	61		