SCIENTIFIC REPORTS

OPEN

Plant organ evolution revealed by phylotranscriptomics in *Arabidopsis thaliana*

Received: 21 March 2017
Accepted: 4 July 2017
Published online: 08 August 2017

Li Lei[1], Joshua G. Steffen[2], Edward J. Osborne[3] & Christopher Toomajian[1]

The evolution of phenotypes occurs through changes both in protein sequence and gene expression levels. Though much of plant morphological evolution can be explained by changes in gene expression, examining its evolution has challenges. To gain a new perspective on organ evolution in plants, we applied a phylotranscriptomics approach. We combined a phylostratigraphic approach with gene expression based on the strand-specific RNA-seq data from seedling, floral bud, and root of 19 *Arabidopsis thaliana* accessions to examine the age and sequence divergence of transcriptomes from these organs and how they adapted over time. Our results indicate that, among the sense and antisense transcriptomes of these organs, the sense transcriptomes of seedlings are the evolutionarily oldest across all accessions and are the most conserved in amino acid sequence for most accessions. In contrast, among the sense transcriptomes from these same organs, those from floral bud are evolutionarily youngest and least conserved in sequence for most accessions. Different organs have adaptive peaks at different stages in their evolutionary history; however, all three show a common adaptive signal from the Magnoliophyta to Brassicale stage. Our research highlights how phylotranscriptomic analyses can be used to trace organ evolution in the deep history of plant species.

The evolution of species' phenotypes occurs at two levels, not only through changes in protein sequence but also in gene expression levels (*i.e.*, changes in the transcriptome)[1]. Though early molecular evolutionary studies focused on protein evolution, data from subsequent research supported the important role gene expression changes played in phenotypic evolution in both animals[2] and plants[3]. However, examining the evolution of gene expression in different organs presents novel challenges compared to examining the evolution of protein coding sequence. Yang and Wang, through comparing the transcriptomes in different tissues of maize, rice and *Arabidopsis*, found a correlation across genes in the rates of sequence evolution and divergence in gene expression patterns, and also evidence that the expression differences of the orthologous genes among different species varied between different organs[4]. Nevertheless, the age and sequence divergence level of those transcriptomes in different organs and how those organs adapted in different stages along their evolutionary history are unknown.

Transcriptome profiling by RNA-sequencing facilitates comparisons of gene expression across organs in the context of coding sequence evolution. Comparisons of gene expression in different organs combined with their coding sequence evolution will illuminate how these processes jointly shape the phenotypic diversity of the animal and plant kingdoms.

Due to DNA's double-stranded complementarity, the mRNA transcripts that are used as templates for protein translation look like what is referred to as the sense strand of the DNA with respect to the protein-coding gene, and the whole process called sense expression. However, certain genes display measurable antisense expression, in which mRNAs are transcribed in the opposite (antisense) direction and thus contain sequence complementary to the sense mRNA. Natural antisense transcripts (NAT) produced by antisense expression can regulate the transcript abundance of their complements by triggering the biogenesis of natural antisense short interfering RNAs (nat-siRNAs) that subsequently guide transcript cleavage[5–7]. Additionally, antisense expression can have a function in antisense transcript-induced RNA splicing, alternative splicing, and polyadenylation, which can regulate gene expression abundance and protein-coding complexity[8].

[1]Kansas State University, Department of Plant Pathology, Manhattan, KS, 66506, USA. [2]Colby-Sawyer College, Natural Sciences Department, New London, NH, 03257, USA. [3]University of Utah, Department of Biology, Salt Lake City, UT, 84111, USA. Correspondence and requests for materials should be addressed to L.L. (email: llei@umn.edu) or C.T. (email: toomajia@ksu.edu)

A novel approach for comparing the gene expression in different organs in conjunction with sequence evolution is to estimate the contribution of genes from different "phylostrata"[9,10] to gene expression. The evolutionary origin of any gene can be traced by sequence similarity searches in genomes representing the whole tree of life, an approach known as "phylostratigraphy". In this approach, every gene within a genome has a phylogenetic rank, and is associated with a "phylostratum"[10] based on its inferred phylogenetic emergence. By combining phylogenetic hierarchy and gene expression to examine the developmental hourglass model, which predicts the pattern of morphological divergence for different developmental stages, in zebrafish, Domazet-Loso *et al.* first introduced the transcriptome age index (TAI), which integrates the age of a gene with its expression level at a given developmental stage and sums this over all genes expressed at the respective stage[10]. Similarly, Quint *et al.* introduced the transcriptome divergence index (TDI), which integrates the sequence divergence of a gene with its expression level at a given developmental stage and sums this over all genes expressed at the respective stage, to explore the embryonic developmental hourglass in *Arabidopsis*[11]. Recently, using TAI and TDI, Cheng *et al.* investigated the developmental hourglass in fungi[12]. Drost *et al.* summarized the cross-kingdom comparison of the developmental hourglass with TAI and TDI[13]. Although these studies applied the indices TAI and TDI, their application was limited to the examination of the developmental hourglasses in different organisms. More recently, several studies used these two indices to investigate the age of genes and their contributions to transcriptomes in different developmental stages, especially in reproductive tissues, for example, in pollen[14], seed germination and flower development[15], and plant reproductive development[16]. However, these reports focused more on comparisons of the reproductive tissues or stages from single accessions from different species, and did not explore the adaptive peaks (over-represented phylostrata) at different stages in their evolutionary history.

It is well known that root, flower, stem and leaves (the latter two are major components of seedlings) are important organs for the majority of plants. Thus, knowing more about the evolutionary properties of their transcriptomes, and how those organs adapted in different stages of evolutionary history could be of great significance in studying a plant organ's origin or development, and even for improving a plant's adaptation to the environment. Nevertheless, how and when these organs evolved is not well understood, although there are estimates of those organs' origins. Some of the earliest fossil evidence of roots comes from 419–408 million years ago (mya) in clubmosses (*e.g.*, *Drepanophycus spinaeformis*) and their close relatives in the extinct zosterophylls (*e.g.*, *Bathurstia denticulate*)[17]. Flower-like structures first appeared in the fossil record ~130 mya in the Cretaceous[18]. Leaves first evolved during the Late Devonian to Early Carboniferous, diversifying rapidly until the designs settled down in the mid Carboniferous[19]. Although one can infer the origin of each organ of plants from the fossil record, details of the adaptive history for each organ are still lacking, specifically, in which phylostrata are the organ-specific expressed genes enriched (indicating an adaptive stage for plant organs). Several pioneering studies can serve as models for examining the adaptive profiles of plant organs because they have used the approach of mapping genes specifically expressed from domains in the vertebrate head sensory system and brain to their phylostrata[20,21] in order to reveal the adaptive history of each domain.

Here we examine the age and sequence divergence of the transcriptomes from different organs in 19 *Arabidopsis thaliana* accessions to reveal the adaptive profile of each organ along its evolutionary history. We applied the phylostratigraphic approach and combined it with gene expression based on the strand-specific RNA-seq data from seedling, floral bud, and root. Our results indicate that, among the sense and antisense transcriptomes of these organs, the sense transcriptomes of seedlings are the evolutionarily oldest across all accessions and are the most conserved in amino acid sequence for most accessions. In contrast, among the sense transcriptomes from these same organs those from floral bud are evolutionarily youngest and least conserved in sequence for the majority of accessions.

Results

Analyses of gene expression in *Arabidopsis* seedling, root and floral bud. We collected the strand-specific transcriptomes from three organs: root, seedling and stage 12 floral buds[22] (flower for short hereafter) from 19 *Arabidopsis* accessions, which are the founders of the MAGIC lines[23,24]. Then we normalized the expression as reads per million (RPM). Due to the noise inherent in estimating gene expression by RNA-seq and the potential for polymorphism across the 19 accessions, we classified a gene as expressed in a certain organ (separately for both sense and antisense expression) if the expression abundance in at least 5 accessions was above 0.8 RPM in that organ. Figure 1A (sense expression) & 1B (antisense expression) indicate the expression breadth of all expressed genes. Given our three organs, we defined seven expression groups: specific to root (R), flower (F), and seedling (S); shared between root-seedling (RS), flower-root (FR), and seedling-flower (RS), and shared among root-seedling-flower (RSF). More than half of sense expressed (85.63%) and antisense expressed (55.46%) transcripts are expressed in more than a single organ (Fig. 1A and B). Organ-specific expressed genes and genes expressed in only two organs are much more frequent for antisense compared with sense expression, while the genes shared by three organs are underrepresented in antisense expression (Fig. 1A and B; Table S1).

Additionally, we clustered each organ from each accession based on gene expression (Fig. 1C and D). Sense expression between accessions within organs is highly correlated while antisense expression is more variable across accessions. Also, sense expression variation among accessions is very minor compared to that among organs.

Phylostrata. Any gene can be mapped to the time point—the phylostratum—when its oldest domain emerged in evolution[10]. Following the approach of Drost, *et al.*[25], the 25,260 protein-coding genes of *A. thaliana* (TAIR 10) were assigned into 12 phylostrata (Fig. 2A). The phylostrata of those genes cover a time span from the origin of cellular organisms (ca. 2520 Ma) to the terminal lineage (0 Ma), that is, *A. thaliana*. More than half of genes (64.87%) originated from the three most ancient phylostrata (Cellular organisms, Eukaryota, and Viridiplantae), which indicates that relatively few genes originated during the subsequent evolutionary processes

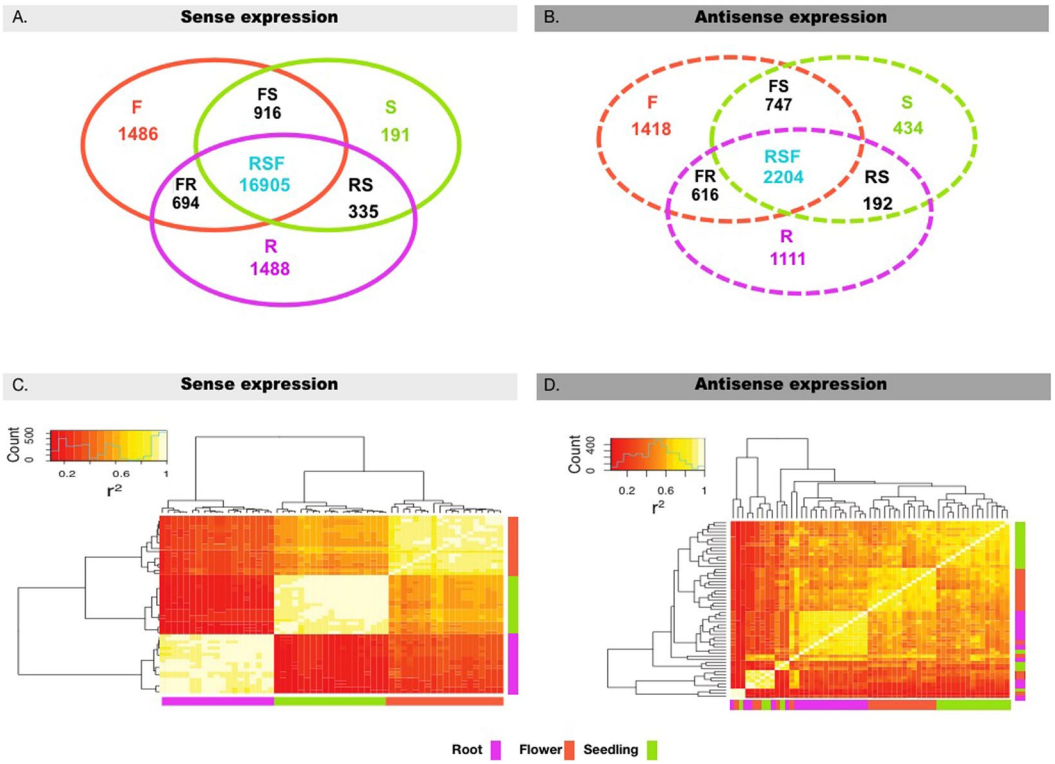

Figure 1. Gene expression in root, flower and seedling across 19 *A. thaliana* accessions. (**A**) Venn diagram of the counts of genes with sense expression in root, flower and seedling across 19 accessions. (**B**) Venn diagram of the counts of genes with antisense expression in root, flower and seedling across 19 accessions. (**C**) Heatmap of sense expression in three organs across 19 accessions. (**D**) Heatmap of antisense expression in three organs across 19 accessions. F: Flower-specific expressed genes; S: Seedling-specific expressed genes; R: Root-specific expressed genes; FS: genes with expression shared by Flower and Seedling only; FR: genes with expression shared by Flower and Root only; RS: genes with expression shared by Root and Seedling only; RSF: genes with expression shared by Root, Seedling and Flower.

along the land plant specific portion of the lineage leading to *A. thaliana*. Nevertheless, these later appearing genes may have played an important role in the development and divergence of seed plants. For the rest of the phylostrata, the largest set of genes appeared in Embryophyta (10.43%), then *A. thaliana* (6.94%), and then Magnoliophyta (6.22%). That could indicate that those three phylostrata could be key phylostrata in their ultimate contribution to *A. thaliana* evolution.

Evolutionary age of *Arabidopsis* seedling, root and flower transcriptomes. In order to estimate the age of each transcriptome, we computed a mean TAI and standard error based on phylostratigraphy for each organ of each accession with 1000 bootstrap replicates. The TAI quantifies the mean evolutionary age of the transcriptome in a certain organ; the lower the TAI, the evolutionarily older the transcriptome[9,11]. The standard errors of TAI in both sense and antisense across 19 accessions are small, less than 0.3% of the mean for sense and less than 0.8% of the mean for antisense (Fig. 3A and B and Table S2). For sense expression across 19 accessions, the evolutionarily older genes tended to be expressed in seedling (consistently lowest average TAI across all accessions), while the younger genes tended to be expressed in flower (highest average TAI in all but three accessions, Wil-2, Zu-0 and Bur-0, in which root has the highest average TAI rather than flower) (Fig. 3A and Tables S2 and 3). This is consistent with previous studies, which reported that young genes tend to contribute more to the transcriptomes of the reproductive tissues compared with the old genes[14,16]. Cui, *et al.* also found that root has a TAI nearly as high as the reproductive tissues[14], which is also observed in our results (Fig. 3A). This result is also clearly evident in the comparison of relative expression between old genes (PS1–PS3) and young genes (PS4–PS13) (Fig. 3E and Figure S3). It shows that old genes have much higher relative sense expression in seedling than young genes; in contrast, in root and flower, young genes have relatively higher sense expression than old genes. In particular, the magnitude of the difference between young genes and old genes tends to be bigger in flower than in root (Fig. 3E and Figure S3). This matches previous reports that young genes in reproductive stages tend to have higher expression than old genes[14,16]. Additionally, the pattern of TAI is fairly consistent across accessions, with the minor exception of three accessions in which root has the highest average TAI rather than flower. It should be noted that due to the lack of biological replication for each accession-organ combination, we cannot quantify their biological variability, nor make statistically supported statements about how the TAI patterns across organs differs among the accessions.

A.

		Phylostrata	Clade-specific genes#	Percentage (%)	Age(Ma)
High	Arabidopsis thaliana	PS12	1752	6.94	0
	Arabidopsis	PS11	188	0.74	13
	Brassicales	PS10	970	3.84	71
	Rosids	PS9	352	1.39	104
	Core eudicotyledons	PS8	276	1.09	116
	Eudicotyledons	PS7	282	1.12	125
	Magnoliophyta	PS6	1571	6.22	170
	Tracheophyta	PS5	846	3.35	496
	Embryophyta	PS4	2635	10.43	547
	Viridiplantae	PS3	782	3.10	968
	Eukaryota	PS2	6503	25.74	1628
	Cellular organisms	PS1	9103	36.04	2520
Low	Total		25260		

B.

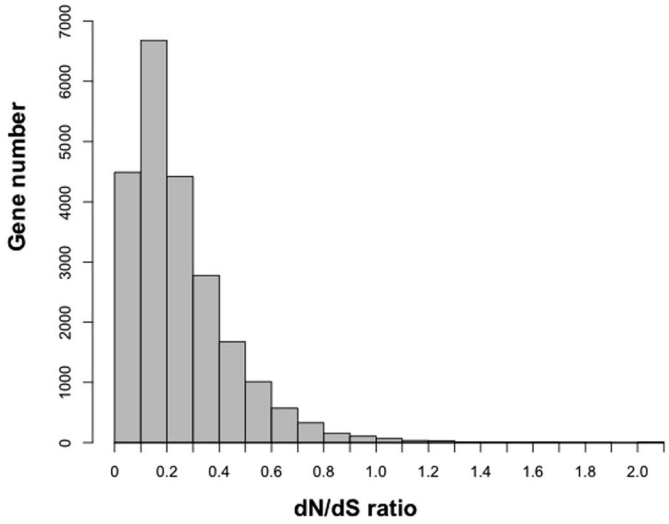

Figure 2. Phylostrata and *dN*/*dS* ratio for protein coding genes in *A. thaliana*. (**A**) Phylostrata and the number of genes mapped to each phylostratum. (**B**) The distribution of Nonsynonymous substitution rate (*dN*)/ Synonymous substitution rate (*dS*) for orthologous gene pairs of *A. thaliana* and *A. lyrata*.

In contrast, for antisense expression, TAI is much higher for all three organs, *i.e*, antisense expression tends to occur in younger genes, and there are no major consistent differences in TAI across organs (Fig. 3A,B and F; Figure S3). Instead, two accessions (Col-0 and Wil-2) have a much higher value for TAI (Fig. 3B and Table S2), though due to lack of biological replication we cannot assess the significance of the higher Col-0 and Wil-2 values. Nevertheless, a large number of young genes that show antisense expression in a restricted set of accessions, including Col-0 and/or Wil-2, could produce their higher TAI values. A bias in the analysis of accession-specific genes, which also must correspond to the *A. thaliana* phylostratum (PS12) if they resulted from gene duplication or creation, predicts this result for Col-0. Specifically, though accession-specific genes are expected in all of the accessions in the study, using Col-0 as the reference to define the set of genes for the phylostrata analysis (as we have done) can create a bias in TAI calculations across the set of accessions, because genes specific to other accessions but absent in Col-0 are not included in the analyses. On the other hand, genes specific to Col-0 are included in the TAI calculation but must not be expressed in other accessions. As a consequence, Col-0 and the accessions most closely related to it are expected to have slightly higher average expression from the class of the youngest genes relative to other accessions. This would tend to increase TAI for Col-0, which is what we see, at least for the TAI calculation for antisense expression.

Sequence divergence of *Arabidopsis* seedling, root and flower transcriptomes. The *dN/dS* ratio is used to assess the selective pressure on protein-coding regions and reflects natural selection. The majority of the orthologous gene pairs between *A. thaliana* and other species in Brassicaceae have a *dN/dS* ratio less than 1 (Fig. 2B and Figure S1), indicating purifying selection predominates in the majority of gene pairs. However, we found a significant enrichment for Gene Ontology terms related to lipid localization, transport and binding, and

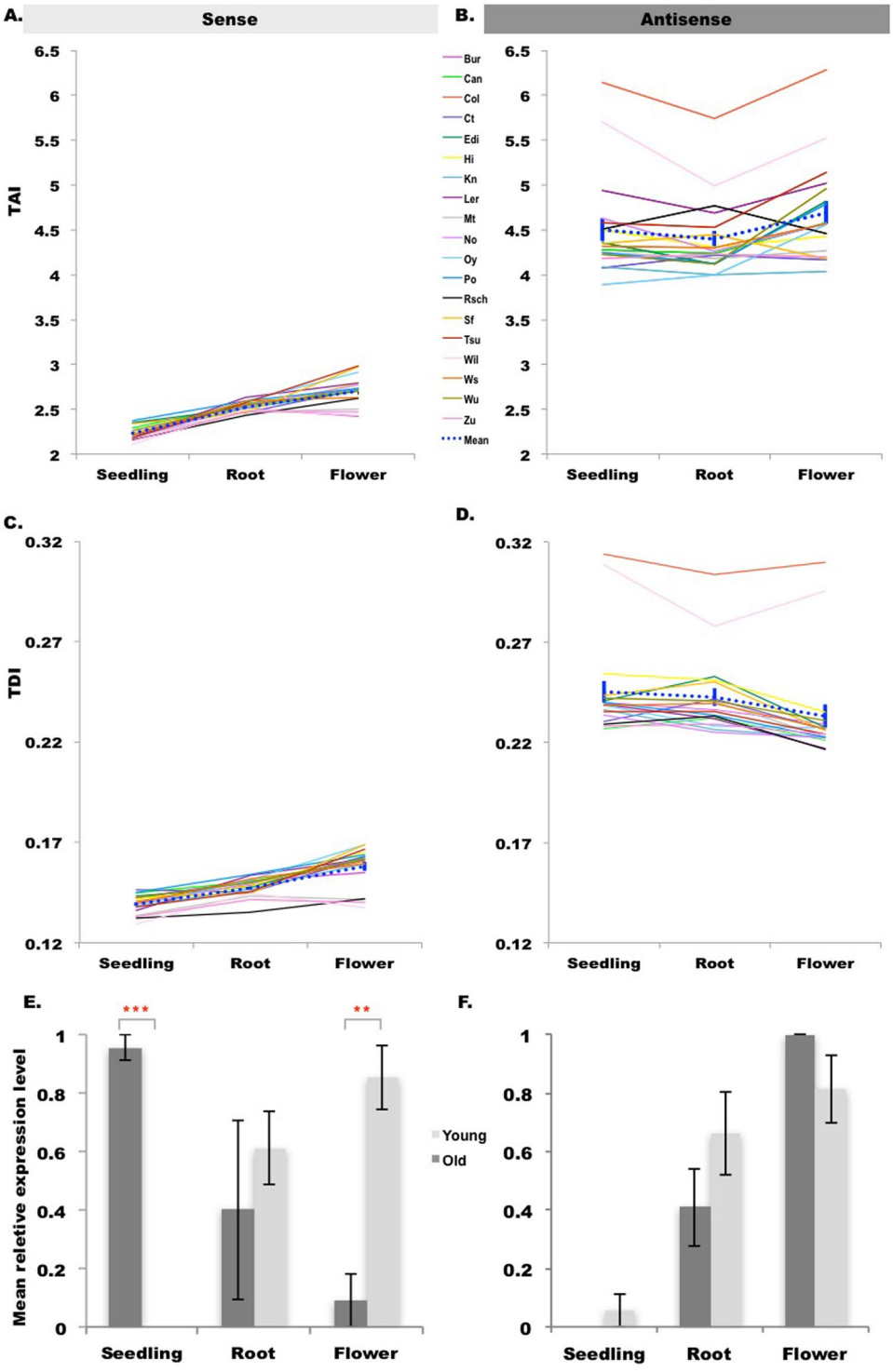

Figure 3. Transcriptome indices of seedling, root and flower across 19 *A. thaliana* accessions. (**A**) The sense transcriptome age index (TAI) profile of three organs across 19 accessions. (**B**) The antisense transcriptome age index (TAI) profile of three organs across 19 accessions. (**C**) The sense transcriptome divergence index (TDI) profile of three organs across 19 accessions. (**D**) The antisense transcriptome divergence index (TDI) profile of three organs across 19 accessions. The standard errors of TAI and TDI, estimated by bootstrap analysis (1000 replicates), are extremely small so that they cannot be seen in (**A**–**D**), but are presented in Tables S2, 4, 6, 8 and 10. The comparisons of TAI and TDI between any two organs are significant by the Mann Whitney test, and the p-values are presented in Tables S3, 5, 7, 9 and 11. (**E**) Mean relative sense expression levels of evolutionarily old (PS1–PS3) and young (PS4–PS12) genes in different organs in a representative accession: Ler-0. (**F**) Mean relative antisense expression levels of evolutionarily old (PS1–PS3) and young (PS4–PS13) genes in different organs in Ler-0. The comparisons of relative antisense expression levels between old and young genes in different organs were performed by t test: ***means p-value < 0.001; **means p-value < 0.01; *means p-value < 0.05; no *means not significant.

the endomembrane system for genes with a ratio that exceeds one, but only when comparing to homologs from *A. lyrata* and not from two more distant relatives (Table S23). We also investigated the distribution of genes with a ratio greater than one across the 12 phylostrata, and we found that all phylostrata between Magnoliophyta and *Arabidopsis* inclusive show an enrichment for these genes, but the enrichment is only significant by a hypergeometric test after multiple testing correction in Eudicotyledons, Rosids, Brassicales, and *Arabidopsis* (Table S24). Only cellular organisms had a significant under-representation of genes with $dN/dS > 1$.

To estimate the transcriptome divergence, we computed a transcriptome divergence index (TDI) based on *dN/dS* ratios for each organ. TDI quantifies the mean selection force acting on the coding sequence of a transcriptome; the further below 1 the TDI, the stronger the purifying selection[26]. We found that TDI exhibited a similar pattern to the TAI for both sense and antisense expression. Compared to sense expression, the TDI of antisense expression is much higher across all three organs. For sense expression, seedling tended to have the lowest average TDI, giving the strongest signal of purifying selection and greatest amino acid conservation, and flower tended to have the highest average TDI, suggesting its transcriptome was experiencing weaker purifying selection and therefore greater amino acid divergence (Fig. 3C; Figure S2A,C and E, Table S4, 5, 6, 7, 8, 9, 10 and 11). This pattern is also consistent with previous studies, which reported that plant reproductive tissues tend to have more divergent transcriptomes and less selective pressure[4,14,16]. But the order of the transcriptomes divergence in the seedling and root is opposite that observed by Yang and Wang, though that could be because of different methods used to estimate the expression divergence and the use of only a single accession of *A. thaliana*[4]. The pattern of TDI for sense expression is relatively consistent across accessions, with only five accessions deviating from this general pattern when using TDI computed with *dN/dS* ratios estimated from comparing *A. thaliana* to any of four related species. The lowest average TDI value occurs in root rather than seedling for Ct-1 (across all four TDI calculations) and Sf-2 (in two of the four TDI calculations), while the highest average TDI value occurs in root rather than floral bud for Wil-2 (across all four TDI calculations), Zu-0 (in two of the four TDI calculations) and Mt-0 (in one of the four TDI calculations).

Similar to TAI, the standard errors of TDI in both sense and antisense expression across 19 accessions are also small, less than 0.7% of the mean in every case (Fig. 3C and D and Tables S4, 6, 8 and 10). In contrast to the consistent pattern of TDI for sense expression across accessions, for antisense expression there is no consistent pattern of TDI with respect to organs and relatively higher variance in TDI across accessions (Fig. 3D and Figure S2B,D and F). Instead, two accessions (Col-0 and Wil-2) have a much higher value for TDI when using *dN/dS* ratios computed from comparisons of *A. thaliana* with *A. lyrata* (Fig. 3D, Table S4), but do not show similar elevated TDI values when using *dN/dS* ratios computed from a comparison of *A. thaliana* with *T. halophila* (Figure S2B and Table S6), *C. rubella* (Figure S2D and Table S8) or *B. rapa* (Figure S2F and Table S10). Though lack of biological replication prevents our assessing the significance of the high TDI for the two accessions, this pattern could result from a reason similar to the bias in TAI due to accession-specific genes explained above. In this case, the relevant genes must be present in the closest related species (or else divergence could not be calculated), but nevertheless many genes can still be polymorphic for presence/absence among *A. thaliana* accessions. The bias occurs because genes are only included in the analysis when they are found in Col-0, but can be absent from other accessions. Assuming that presence/absence genes tend to have less selective constraint, and hence a higher *dN/dS* ratio, Col-0 is expected to have an inflated ratio. Once we make comparisons with relatives more distant than *A. lyrata*, the inflation in the Col-0 TDI greatly decreases or disappears. This would be expected if genes with presence/absence polymorphisms are largely limited to those that arose after the *A. thaliana* lineage split from all but its closest relative, so that the bias specific to Col-0 and the accessions most closely related to it disappears once the analysis is restricted to genes for which divergence from the more distant relatives can be estimated.

Adaptive patterns in *Arabidopsis* seedling, root and floral bud. To further shed light on the history of adaptation (estimated as the enrichment of the organ-specific expressed genes in different phylostrata) in seedling, root and flower, we mapped seedling-specific (S), root-specific (R) and flower-specific (F) expressed genes (Fig. 1A and B) (for both sense and antisense expression) to their corresponding phylostrata, and plotted the adaptive profiles of these genes (Fig. 4 and Figure S4). We also mapped the genes expressed in all three organs (RSF) (Fig. 1A and B) to phylostrata, and plotted the adaptive profile for comparison (Fig. 4 and Figure S4). Generally, different organs have different adaptive peaks (over-represented phylostrata) (Fig. 4). The adaptive peaks for floral bud are at the Embryophyta, Magnoliophyta, Eudicotyledons, core Eudicotyledons, Rosids and Brassicales phylostrata, although only in Eudicotyledons and Brassicales is the over-representation significant. The first adaptive peak appears at Embryophyta, though the level of over-representation is low, suggesting that flower formation could partially be traced back to genes originating in Embryophyta ancestors. The second adaptive peak appears at the Magnoliophyta phylostratum. Flower-specific genes are also overrepresented in Eudicotyledons, core Eudicotyledons, Rosids and Brassicales, suggesting that novel flower-specific genes continued to originate at different stages of plant evolution after Magnoliophyta to sustain the complex character and activity of flowers. But after Brassicales there is no enrichment in *Arabidopsis* and *A. thaliana*, possibly indicating that new genes that have evolved since Brassicales are not often expressed in flowers.

The adaptive peaks for root are at Embryophyta, Tracheophyta, Magnoliophyta, Eudicotyledons, core Eudicotyledons, Rosid, Brassicales and *A. thaliana* (Fig. 4). The biggest adaptive peak for root is in Tracheophyta. This indicates that Tracheophyta is the most important phase during the evolutionary history of the organ, consistent with the fact that starting from Tracheophyta, plants have had independent roots, and that the earliest fossils of roots indicate they originated from early species within Tracheophyta[27].

The adaptive peaks for seedling are at Embryophyta, Magnoliophyta, Eudicotyledons, core Eudicotyledons, Rosids, Brassicales and *A. thaliana* (Fig. 4). Additionally, we also observed that before Embryophyta, none of the previous phylostrata are overrepresented in the three sets of organ-specific sense-expressed genes, but the RSF

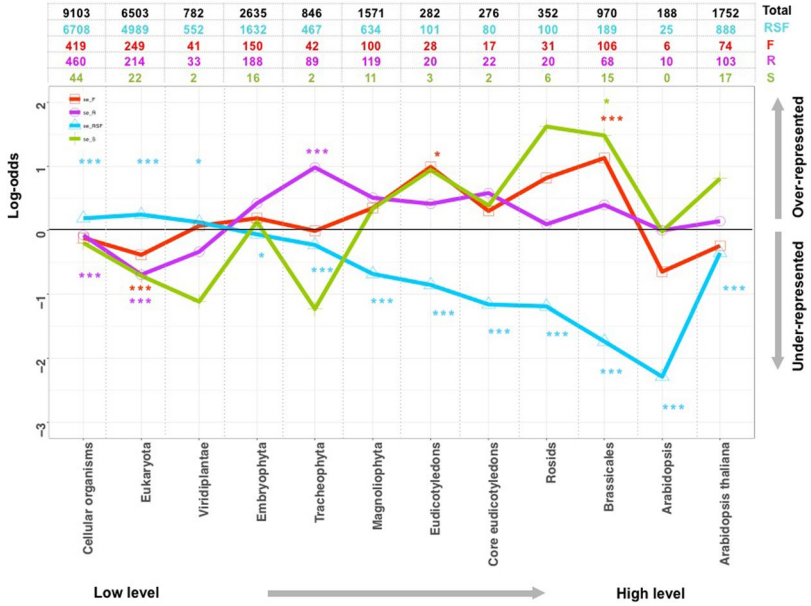

Figure 4. Enrichment analysis of organ-specific sense-expressed genes in different phylostrata. RSF: genes with expression shared by Root, Seedling and Flower; S: Seedling-specific expressed genes; R: Root-specific expressed genes; F: Flower-specific expressed genes; Total: all the protein coding genes with an assigned phylostratum. Gray line: A log-odds of zero, which corresponds to the actual number of organ-specific genes in each phylostratum equaling the predicted number. Because there are zero S genes from phylostratum *Arabidopsis*, the log-odds value is undefined here. In plotting the figure, we conservatively adjusted the observed count to one and re-calculated the log-odds value. (*p-value < 0.05, **p-value< 0.01 and ***p-value < 0.001).

shared genes show over-representation for each of the three early phylostrata (Fig. 4), which could be because root, seedling and floral bud were not differentiated before this phylostratum, also indicating that few new genes with broad expression across all 3 organs have originated since root, seedling, and floral bud have differentiated.

Investigating the adaptive pattern of genes with antisense expression in different organs (Figure S4), we found that for all the groups, in all of the phylostrata before Magnoliophyta, genes with organ-specific antisense expression were distributed fairly evenly without strong over-representation (with log-odd less than 0.5) except in Viridiplantae. In phylostrata 11 (*Arabidopsis*), all groups of genes were strongly under-represented, which indicates that antisense expression of genes that originated in this phylostratum is very rare, and hence could not play a major role in regulating gene expression for these genes. The overrepresentation value only exceeds 0.5 in core Eudicots and Viridiplantae; and both are in root. This could indicate that during the evolution of root, antisense expression could have been important to regulate the gene expression in the phylostrata of Viridiplantae and core Eudicots. Also in Embryophyta, even though the log-odds are less than 0.5, the root and the RSF expressed genes show significant enrichment, which could suggest that antisense expression played an important role in regulating gene expression in Embryophyta.

Gene ontology analysis for organ-specific genes and genes expressed in all three organs. Since we investigated the adaptive patterns for seedling, flower and root, we were curious about the functional categories to which organ-specific genes belong, focusing on sense expression specifically. We performed the Gene Ontology (GO) analysis for R, S and F genes considering all the genes expressed in at least one organ as background. In parallel, we performed a GO analysis for genes in the RSF set (Table S14). We found that flower-specific genes tend to be enriched in several terms, including: pollen tube development, plant-type cell wall and cell morphogenesis related, and reproductive processes related (Table S12). We specifically examined 12 genes with well documented roles in flowering according to Marín, *et al.*[28]: FLC, CO, FT, SOC1, LFY, GAI, FY, FRI, VRN1, VRN2, LHY and TOC1. We found that sense expression of only CO, FT, and LFY is flower specific; CO and FT can be tracked back to Eukaryota while LFY originated from Embryophyta. TOC1, from Cellular organisms, has no sense expression in the three tissues but has flower-specific antisense expression. The remaining eight flowering genes are expressed in all three organs. GAI and FI are from Cellular organisms; SOC1, FLC, VRN2, and LHY are from Eukaryota; FRI and VRN1 are from Embryophyta.

Root-specific genes tend to be enriched in several terms, including: oxidoreductase related, transport related terms related to response to different biotic and abiotic factors, and root development related terms (Table S13). Genes expressed in three organs tend to be enriched in the following terms: organelle related processes, cytoplasmic related processes, cellular metabolite processes, etc. (Table S14). We also examined three well known root related genes, CaS (CALCIUM SENSING RECEPTOR)[29,30], EIR (ETHYLENE INSENSITIVE ROOT 1)[31], and RHL1 (ROOT HAIRLESS 1)[32] to see to which phylostratum they belong and if their expression is root specific. CaS, from Viridiplantae, is sense expressed in all three organs. EIR1 is sense expressed in root and flower, and it

originated from Cellular organisms. RHL1 is only expressed in root and it originated in Embryophyta. However, we found no GO terms enriched in seedling-specific genes, which could be because of the small number of seedling-specific sense-expressed genes.

Gene ontology analysis for genes in each phylostratum. In order to explore if there is an association of certain cellular components or functions with certain phylostrata, we performed the Gene Ontology analysis for genes in each phylostratum considering all the genes expressed in at least one organ as background. We found that with the exception of four phylostrata (Cellular organisms, Eukaryota, Eudicotyledons, and Core eudicotyledons), GO terms are enriched in the remaining phylostrata (Table S15–22). For example, genes from Viridiplantae and Embryophyta tend to be enriched in some of the same terms, including: regulation of transcription, biosynthetic process, nucleobase, nucleoside, nucleotide and nucleic acid metabolic process, and primary metabolic process, and response to stimulus (Tables S15 and 16). However, both phylostrata are also enriched in distinct GO terms. For instance, genes from Viridiplantae are enriched in intracellular organelle, photosynthetic membrane, and regulation of timing of transition from vegetative to reproductive phase (Table S15); while genes from Embryophyta are enriched in aging, innate immune response, organ development, like shoot morphogenesis and meristem initiation (Table S16). Genes from Tracheophyta are enriched in endomembrane system, anchored to membrane, plasmodesma, symplast, and junction between and within cell (Table S17). Genes from Magnoliophyta are enriched in pectinesterase inhibitor activity, and lipid localization, binding, and transportation (Table S18). Genes from Rosids are enriched in receptor binding, pectinesterase inhibitor activity and pectinesterase activity, apoplast, and endomembrane system (Table S19). Genes from Brassicales are enriched in endomembrane system, receptor binding, extracellular region, apoplast and structural constituent of cell wall (Table S20). Genes from Arabidopsis are enriched in beta-galactosidase complex, beta-galactosidase and galactosidase activity (Table S21). Genes from *Arabidopsis thaliana* are enriched in endomembrane system and plasma membrane (Table S22).

Discussion

Plants play an important role in life on land, and the emergence of their morphological differences correlates with speciation, the ecological consequences of variation in physiological or developmental traits, and the adaptive evolution of different species[33]. Developing an understanding of the evolutionary origins of different organ systems lies at the heart of evolutionary biology. A direct approach to examine the origins and evolution of plant organs uses fossils[34,35]. The fossil record alone cannot comprehensively reveal the evolutionary processes for plant species[35]. Here we took a novel approach, phylotranscriptomics. This indirect approach combines gene expression and sequence evolution to investigate the age and sequence divergence of the transcriptomes in different organs and the adaptive profile of those organs along their evolutionary history.

TAI reflects long-term evolutionary changes covering 4 billion years since the origin of life, and TDI reflects short-term evolutionary changes covering 5–16 million years since the divergence of *A. thaliana* and four other Brassicaceaes[49,51–53]. The results of TAI and TDI analyses for sense expression indicate that the transcriptome in flower is youngest and has the highest amino acid sequence divergence, the root transcriptome is much older with much lower sequence divergence, and finally the transcriptome in seedling is oldest and most conserved in sequence, experiencing more purifying selection. The result that the transcriptome in flower is youngest is consistent with previous reports that young genes tend to contribute more to the transcriptomes in the reproductive tissues or stages[14,16]. This may be consistent with the theory of plant organ evolution and the earliest fossil records of plant organs. According to Hagemann's theory, the most primitive land plants that gave rise to vascular plants were flat, thalloid, leaf-like, without axes, somewhat like a liverwort or fern prothallus. Axes such as stems and roots evolved later as new organs[36,37]. Roots evolved in the sporophytes of at least two distinct lineages of early vascular plants (Tracheophyta) during their initial major radiation on land in Early Devonian times (c. 410–395 million years ago)[27]. Real flowers are modified leaves possessed only by Magnoliophyta, a group that originated and diversified during the Early Cretaceous[38]. Flowers are relatively late to appear in the fossil record. However, there are some fossil records of flower-like organs in ferns or cycads and gnetales[22]. Seedlings include the stems and leaves, both of which are more primitive than roots and flowers[27,38]. This is consistent with our conclusion that the transcriptome in seedling is oldest and most conserved at the amino acid sequence level, compared with root and flower because root and flower evolved later than stem and leaves.

By comparing their adaptive profiles, we obtained a global picture of the evolution of three important organs. Root, seedling and flower-specific sense-expressed genes are over-represented at the phase of Embryophyta, indicating an early adaptive peak of those organs at Embryophyta, although there is no fossil evidence supporting the origin of root, leaves and stems, and flower during this stage[27,36,37,39]. Thus, the Embryophyta stage could have included preadaptive events for each organ, where genes that play a role in an organ initially emerged before the organ differentiated. Then from Magnoliophyta until Brassicales, organ-specific expressed genes show over-representation during these phylostrata, suggesting they are key stages for the evolution of these organs in the lineage leading to *A. thaliana*. A significant adaptive signal appearing in Magnoliophyta is expected, because fossil evidence suggests that innovation of real flowers originated within Magnoliophyta[38]. Flowering plants are diverse as a group, with 250,000 to 400,000 species of Magnoliophyta[40–42]. This compares to around 12,000 species of moss[43] or 11,000 species of pteridophytes[44]. The adaptive peak of root and seedling at the stage of Magnoliophyta could result from the need for other organs like roots, stems, and leaves to adapt within that great diversity of Magnoliophyta species. According to the Angiosperm Phylogeny Group III system (APGIII)[45] angiosperms (Magnoliophyta) can be divided into Eudicotyledons and Monocots. Between these, there are major differences in leaves, flowers, roots, and other organs[46,47], so it is meaningful that root, seedling and flower have an adaptive peak at this stage. Similarly, eudicotyledons, core eudicotyledons, Rosids and Brassicales include many different subgroups of plants. For example, core eudicotyledons include Rosids, Gunnerales, Dilleniaceae,

Berberidopsidales, Santalales, Caryophyllales and asterids. These groups have large differences in their morphology, either in root, leaves or flower, which could explain the adaptive peaks in these phylostrata.

In higher eukaryotic organisms, 4% to 26% of protein coding genes are predicted to generate NAT[48]. In *A. thaliana*, Yuan *et al*. identified 7,962 genes with antisense expression[49], somewhat higher than our results (6327). The difference could be due to the strict threshold (0.8 RPM, which is based on the distribution of the usually more-abundant sense expression) we set to define the genes with antisense expression. Our results showed that gene antisense expression tended to be organ-limited (expressed in either one or two of the three organs) compared to sense expression (Figs 1A and B and S1). NAT may play an important role in regulating the expression or translation of genes that are specifically expressed in a certain organ because 100% of the organ specific sense expressed genes have antisense expression (organ specific sense expressed genes: Root (1486), Seedling (191), Flower (1486)); in contrast, only 28.11% of the genes with sense expression shared by three organs (RSF: 16905) have antisense expression. This is supported by the result that genes with both sense and antisense expression have dramatically lower sense expression levels than genes with only sense expression (in seedling, root and flower, 29.8 vs. 58.0 RPM, p-value < 0.001; 42.8 vs. 53.4 RPM, p-value < 0.01; 34.9 vs. 52.5 RPM, p-value < 0.001, respectively). However, patterns of antisense expression are qualitatively different from those of sense expression in important ways. Sense expression from all of our samples clustered strongly by organ, with only minor differences between accessions, but the antisense expression clustering into organs was weak and inconsistent (Fig. 1C and D). Also, while TAI and TDI values clearly differed between organs in a relatively consistent pattern across accessions for sense expression, this was not true for antisense expression.

These major patterns of antisense expression call into question how much of the observed antisense expression is either functional or adaptive[50]. Another hypothesis compatible with our results assumes most of antisense expression is not adaptive (and indeed much may be maladaptive) but instead a noisy molecular byproduct of transcriptional machinery. Under this hypothesis, genes with the most critical function, highest expression, and/or widest expression might suppress this potentially maladaptive antisense byproduct, while other genes with lower or more limited expression and less critical functions may more frequently have antisense expression because negative selective pressure to eliminate their antisense expression byproduct is weaker. Though our data do not allow us to rule out either the adaptive or transcriptional noise hypotheses, a combination of both likely explains our observed antisense expression patterns.

The phylostratigraphic approach has been widely used to infer patterns of genome evolution, and has been applied to address questions related to the ontogenic patterns predicted by the developmental hourglass model, the adaptive history of different organs, and *de novo* gene origination[9–13,15,20,21,51]. Central to this approach is the inference of the earliest emergence of target sequences at a particular phylogenetic node usually by using the similarity search algorithm BLAST[52] on a set of genomes that represent the nodes. BLAST has known limitations when sequences are highly diverged, especially in detecting remote homologues of short and fast-evolving sequences[53,54]. Due to these limitations of the BLAST algorithm, published phylostratigraphic analyses have recently been criticized due to a predicted bias where certain short or fast-evolving genes will erroneously appear younger than they truly are as a result of BLAST false negatives. Through simulation, Moyers and Zhang argued that genomic phylostratigraphy underestimates gene age for 14% and 11% of the sequences they simulated for *Drosophila* and yeast, respectively[54,55]. They argue that these potential errors for 11 or 14% of genes create spurious patterns in the distributions of certain gene properties, such as length or rate of evolution, across age groups. However, Domazet-Lošo, *et al*.[56] re-assessed these simulations and identified problems that call into question the conclusions of Moyers and Zhang[54]. The problems include irreproducibility, statistical flaws, the use of unrealistic parameter values, and the reporting of only partial results from those simulations[56]. Domazet-Lošo, *et al*.[56] argued that, even with a possible overall BLAST false negative rate between 5–15%, the large majority ($\geq$74%) of sequences assigned to a recent evolutionary origin by phylostratigraphy is unaffected by the technical concerns about BLAST[56]. Further, when they removed from the analysis genes that Moyers and Zhang[54] found susceptible to the BLAST error in their simulations (192 out of 4157 genes with expression) the general profiles of biological patterns remained largely unaffected, indicating such potentially misplaced genes do not distort the major results[56]. They concluded that phylostratigraphic analyses of patterns of gene emergence and evolution are robust to the false negative rate of BLAST[56].

Taken together, the phylotranscriptomics approach, through combining gene expression and sequence evolution, works well to investigate the age and sequence divergence of the transcriptomes in different *Arabidopsis* organs and the adaptive profile of those organs along their evolutionary history. This study opens a new door in the investigation of plant organ evolution, and gives new indirect evidence on the adaptation of plant organs.

Materials and Methods

Sample collection and transcriptome sequencing. Our study used seedlings (11–12 days old after 4 true leaves had emerged), root (from 10-day old seedlings) and floral bud (stage-12) RNA samples without biological replication from each of the 19 *A. thaliana* accessions that are also MAGIC founders[23,24]: Bur-0, Can-0, Col-0, Ct-1, Edi-0, Hi-0, Kn-0, Ler-0, Mt-0, No-0, Oy-0, Po-0, Rsch-4, Sf-2, Tsu-0, Wil-2, Ws-0, Wu-0 and Zu-0. Seedling and root tissue was collected from plants grown at 20 °C under long day conditions (16 hours light 8 hours dark). In order to illicit nearly simultaneous flowering all accessions were vernalized for six weeks at 4 °C under short day conditions followed by a shift back to long day conditions described above. Detailed procedures for plant growth, tissue collection, preparation of RNA, polyA-selected strand-specific library construction, and Illumina sequencing are as previously described[19]. The RNA-seq read data for Col-0 and Can-0 were released previously as part of Gan *et al*.[24] under GEO accession numbers GSM764077-GSM764082. RNA-seq read data for the other 17 MAGIC founders have been released under GEO series GSE53197.

Quantification of gene expression. The RNA-seq reads for all 19 MAGIC founders as available in the GSM764077-GSM764082 and GSE53197 releases for each of the three organs (seedling, root, and stage-12 floral bud) were aligned with PALMapper[57] to the *Arabidopsis* genome for read quantification after Gan *et al.*[24]. The quantification of gene expression and normalization were as described previously[24], and normalized expression (sense and antisense) of each gene for each organ from each accession (reads per million, or RPM) are available in a supplemental dataset (dataset1; https://github.com/lilei1/TAI_TDI_A.thaliana). For each gene, if its sense (or antisense) expression in a certain organ in at least 5 accessions was $\geq$0.8 RPM, then this gene was defined as sense (or antisense, respectively) expressed in this organ. We set 0.8 RPM as the threshold based on the distribution of the sense and antisense expression per gene per accession. In each organ half of the genes per accession have above 0.8 RPM sense expression, while less than 1% of genes per accession have above 0.8 RPM antisense expression.

In order to detect the overall similarity of sense and antisense transcriptome profiles across accessions and organs, we calculated Pearson's correlation coefficient pairwise across all samples. Using the "heatmap.2" package in R, we represented the expression correlations in a heatmap, and performed clustering of samples using the Euclidean distance metric.

Phylostratigraphy and *dN/dS* ratio. Our phylostratigraphic analysis was performed similarly to those presented previously[9–11]. The data about each gene's phylostratum were used directly from Drost *et al.*[25]. Briefly, the first gene model for each *A. thaliana* gene from the TAIR10 release was assigned to one of 12 phylostrata, starting from the origin of cellular organisms and ending at *A. thaliana*. The age of each phylostratum, or the estimated time since the diversification of the corresponding clade from its most recent common ancestor, was extracted from Arendsee *et al.*[58].

The data on the *dN/dS* ratios for each gene were used directly from Quint, *et al.*[11]. Specifically, orthologous gene pairs of *A. thaliana* and *A. lyrata* or the closely related Brassicaceaes, *T. halophila*, *C. rubella* or *B. rapa*, were determined with the method of best hits using blastp[59–63]. And *dN/dS* ratios of only those orthologous gene pairs with $dN < 0.5$, $dS < 5$ and $dN/dS < 2$ were retained for further analysis[11].

Calculation of TAI and TDI. As described previously[10–12, 25], the TAI of organ *s* was calculated as the weighted mean of the evolutionary age (PS) ps_i of gene *i* weighted by the expression level e_{is} of gene *i* at organ *s*:

$$TAI_s = \frac{\sum_{i=1}^{n} ps_i e_{is}}{\sum_{i=1}^{n} e_{is}} \tag{1}$$

where *n* is the total number of genes analyzed. And the expression level was estimated as RPKM. Low PS values correspond to evolutionarily old genes, thus the lower the TAI value, the older the transcriptome is. Likewise, high PS values correspond to evolutionarily young genes, so the higher the TAI value, the evolutionarily younger the transcriptome is.

Analogously, the TDI of organ *s* was calculated as the weighted mean of the sequence divergence (*dN/dS*) of gene *i* weighted by the expression level e_{is} of gene *i* at organ *s*

$$TDI_s = \frac{\sum_{i=1}^{n} \left(\frac{dN_i}{dS_i}\right) e_{is}}{\sum_{i=1}^{n} e_{is}} \tag{2}$$

where *n* is the total number of genes analyzed. Low/high *dN/dS* ratios correspond to conserved/divergent genes, so low/high TDI values correspond to conserved/divergent transcriptomes. To investigate the dependence of the results on the reference species for the calculation of TDI, the analysis was repeated for each reference species (*T. halophila*, *C. rubella* and *B. rapa*). The results are presented in Figure S2 and Tables S6–11.

Estimating standard errors of TAI and TDI by bootstrap analysis. The TAI is written alternatively as the sum of products between the phylostratum ps_i of gene *i* and the partial concentration f_{is} of gene *i* at organ *s*

$$TAI_s = \sum_{i=1}^{n} ps_i f_{is} = ps_1 f_{1s} + ps_2 f_{2s} + \ldots + ps_n f_{ns} \tag{3}$$

Analogously, the TDI is written alternatively as:

$$TDI_s = \sum_{i=1}^{n} \left(\frac{dN_i}{dS_i}\right) f_{is} = \left(\frac{dN_1}{dS_1}\right) f_{1s} + \left(\frac{dN_2}{dS_2}\right) f_{2s} + \ldots + \left(\frac{dN_n}{dS_n}\right) f_{ns} \tag{4}$$

where the partial concentration f_{is} is calculated as the expression level e_{is} from equation (1) divided by the denominator of equation (1) and *n* is the total number of genes analyzed.

Bootstrapping was used to approximate the standard errors of TAI (or TDI)[11, 12]. In brief, for each accession, resampling 25,078 genes (25,078 out of 25,260 genes with an assigned phylostratum have expression) with replacement was done 1,000 times. Then, using the phylostratum, *dN/dS* ratio, and expression abundance (RPKM) in each organ associated with each gene, TAI and TDI can be computed according to equations (3) and (4) for each bootstrap sample of genes. A sample mean of TAI and TDI was obtained according to the results of the 1,000 bootstrap samples. Similarly, the standard errors of the distribution of the sample means were calculated based on the results of the 1,000 bootstrap samples, and were used to estimate the standard error of TAI (or TDI). The script to do bootstrap analysis is available at https://github.com/lilei1/TAI_TDI_A.thaliana.

Statistical significance of the TAI and TDI profiles. For each accession, we performed the Mann Whitney test on TAI and TDI for comparison of each pair of organs based on the results obtained from 1,000 bootstraps.

Relative expression of genes for a given phylostratum. As described previously[10–12,25], the relative expression RE_{ls} of the genes in PS l in organ s was calculated as

$$RE_{ls} = \frac{\overline{g}_{ls} - \overline{g}_{lmin}}{\overline{g}_{lmax} - \overline{g}_{lmin}} \quad (5)$$

Where $\overline{g}_{ls}$ denotes the average partial expression of the genes in PS l at organ s and $\overline{g}_{l\ \max}/\overline{g}_{l\ \min}$ is the maximum/minimum average partial expression across all organs. The value of RE_{ls} ranges from 0 to 1, where 0 denotes the organ where genes in PS l show minimum average partial expression and 1 denotes the organ where genes in PS l show maximum average partial expression. We further grouped the relative expression levels into two PS classes, where the first PS class consists of relative expression levels of genes belonging to the three oldest phylostrata PS1–PS3 and the second PS class consists of relative expression levels of genes belonging to the younger phylostrata PS4–PS12[11]. This grouping was chosen because it distinguishes phylostrata unique to land plants (PS4–PS12) that are evolutionarily relatively young, from older phylostrata (PS1–PS3) that arose before the origin of land plants.

Phylostrata enrichment calculations for organ-specific expressed genes. If organ-specific expressed genes are enriched in a certain phylostratum, this indicates that there are adaptive signals in this phylostratum[20]. In order to find the adaptive signals of three main organs of *A. thaliana*, we firstly defined the organ-specific expressed genes (indicated as R, S and F), separately for sense and antisense expression, and then mapped those genes back to the phylostratigraphy profile that spanned 12 phylostrata[11,20,21,25,64]. In order to have another set for comparison, we also mapped the genes expressed in three organs (RSF) to each phylostrata.

For each of the groups R, S, F and RSF, we performed an over-representation analysis by comparing the actual frequency of each group in a phylostratum with their expected frequency based on the product of the marginal frequencies of the four groups and the proportion of all genes with assigned phylostrata found within each of the 12 different phylostrata[20,21]. The logarithm of the ratio of actual and expected frequencies, or log-odds ratio, was obtained to indicate the enrichment of a group within a certain phylostratum. Then we performed the two-tailed hypergeometric test[65] to test the significance of the over- or under-representation, and controlled for multiple comparisons through a Bonferroni correction.

Gene Ontology for organ-specific expressed genes. Enrichment analysis of gene ontology (GO) terms was performed on organ-specific sense-expressed genes (R, S and F) and genes sense-expressed in all three organs (RSF) using Agri GO[66]. All GO analyses used as background all genes with sense expression in at least one organ. The significance was assessed with hypergeometric tests and corrected by False Discovery Rate. The significance threshold was set at $P = 0.05$, and GO annotations that did not appear in at least 5 entries were not shown, that is, the minimum number of mapping entries was set to 5.

References

1. King, M. C. & Wilson, A. C. Evolution at 2 Levels in Humans and Chimpanzees. *Science* **188**, 107–116 (1975).
2. Carroll, S. B. Endless forms: the evolution of gene regulation and morphological diversity. *Cell* **101**, 577–580 (2000).
3. Doebley, J. & Lukens, L. Transcriptional regulators and the evolution of plant form. *Plant Cell* **10**, 1075–1082 (1998).
4. Yang, R. L. & Wang, X. F. Organ Evolution in Angiosperms Driven by Correlated Divergences of Gene Sequences and Expression Patterns. *Plant Cell* **25**, 71–82 (2013).
5. Borsani, O., Zhu, J., Verslues, P. E., Sunkar, R. & Zhu, J. K. Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in *Arabidopsis*. *Cell* **123**, 1279–1291, doi:10.1016/j.cell.2005.11.035 (2005).
6. Katiyar-Agarwal, S. *et al.* A pathogen-inducible endogenous siRNA in plant immunity. *Proc Natl Acad Sci USA* **103**, 18002–18007, doi:10.1073/pnas.0608258103 (2006).
7. Ron, M., Alandete Saez, M., Eshed Williams, L., Fletcher, J. C. & McCormick, S. Proper regulation of a sperm-specific cis-nat-siRNA is essential for double fertilization in *Arabidopsis*. *Genes Dev* **24**, 1010–1021, doi:10.1101/gad.1882810 (2010).
8. Jen, C. H., Michalopoulos, I., Westhead, D. R. & Meyer, P. Natural antisense transcripts with coding capacity in *Arabidopsis* may have a regulatory role that is not linked to double-stranded RNA degradation. *Genome Biol* **6**, R51, doi:10.1186/gb-2005-6-6-r51 (2005).
9. Domazet-Loso, T., Brajkovic, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet* **23**, 533–539 (2007).
10. Domazet-Loso, T. & Tautz, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468**, 815–U107 (2010).
11. Quint, M. *et al.* A transcriptomic hourglass in plant embryogenesis. *Nature* **490(7418)**, 98 (2012).
12. Cheng, X. J., Hui, J. H. L., Lee, Y. Y., Law, P. T. W. & Kwan, H. S. A "Developmental Hourglass" in Fungi. *Mol Biol Evol* **32**, 1556–1566 (2015).
13. Drost, H. G., Janitza, P., Grosse, I. & Quint, M. Cross-kingdom comparison of the developmental hourglass. *Curr Opin Genet Dev* **45**, 69–75, doi:10.1016/j.gde.2017.03.003 (2017).
14. Cui, X. *et al.* Young Genes out of the Male: An Insight from Evolutionary Age Analysis of the Pollen Transcriptome. *Mol Plant* **8**, 935–945, doi:10.1016/j.molp.2014.12.008 (2015).
15. Drost, H. G. *et al.* Post-embryonic Hourglass Patterns Mark Ontogenetic Transitions in Plant Development. *Mol Biol Evol* **33**, 1158–1163, doi:10.1093/molbev/msw039 (2016).
16. Gossmann, T. I., Saleh, D., Schmid, M. W., Spence, M. A. & Schmid, K. J. Transcriptomes of Plant Gametophytes Have a Higher Proportion of Rapidly Evolving and Young Genes than Sporophytes. *Mol Biol Evol* **33**, 1669–1678, doi:10.1093/molbev/msw044 (2016).
17. Gensel, P. G., Kotyk, M. & Basinger, J. F. in *Plants invade the land: evolutionary and environmental perspectives* Vol. 83–102 (eds P. G. Gensel & D. Edwards) (Columbia University Press, 2001).

18. Lawton-Rauh, A. L., Alvarez-Buylla, E. R. & Purugganan, M. D. Molecular evolution of flower development. *Trends Ecol Evol* **15**, 144–149 (2000).
19. Boyce, C. K. & Knoll, A. H. Evolution of developmental potential and the multiple independent origins of leaves in Paleozoic vascular plants. *Paleobiology* **28**, 70–100 (2002).
20. Sestak, M. S., Bozicevic, V., Bakaric, R., Dunjko, V. & Domazet-Loso, T. Phylostratigraphic profiles reveal a deep evolutionary history of the vertebrate head sensory systems. *Front Zool* **10** (2013).
21. Sestak, M. S. & Domazet-Loso, T. Phylostratigraphic Profiles in Zebrafish Uncover Chordate Origins of the Vertebrate Brain. *Mol Biol Evol* **32**, 299–312 (2015).
22. Smyth, D. R., Bowman, J. L. & Meyerowitz, E. M. Early Flower Development in *Arabidopsis*. *Plant Cell* **2**, 755–767 (1990).
23. Kover, P. X. *et al*. A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in *Arabidopsis thaliana*. *PLoS Genet* **5** (2009).
24. Gan, X. C. *et al*. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **477**, 419–423 (2011).
25. Drost, H. G., Gabel, A., Grosse, I. & Quint, M. Evidence for active maintenance of phylotranscriptomic hourglass patterns in animal and plant embryogenesis. *Mol Biol Evol* **32**, 1221–1231, doi:10.1093/molbev/msv012 (2015).
26. Hurst, L. D. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet* **18**, 486–487 (2002).
27. Raven, J. A. & Edwards, D. Roots: evolutionary origins and biogeochemical significance. *J Exp Bot* **52**, 381–401 (2001).
28. Castro Marin, I. *et al*. Nitrate regulates floral induction in *Arabidopsis*, acting independently of light, gibberellin and autonomous pathways. *Planta* **233**, 539–552, doi:10.1007/s00425-010-1316-5 (2011).
29. Luschnig, C., Gaxiola, R. A., Grisafi, P. & Fink, G. R. EIR1, a root-specific protein involved in auxin transport, is required for gravitropism in *Arabidopsis thaliana*. *Genes Dev* **12**, 2175–2187 (1998).
30. Slovak, R. *et al*. A Scalable Open-Source Pipeline for Large-Scale Root Phenotyping of *Arabidopsis*. *Plant Cell* **26**, 2390–2403, doi:10.1105/tpc.114.124032 (2014).
31. Schneider, K. *et al*. The ROOT HAIRLESS 1 gene encodes a nuclear protein required for root hair initiation in *Arabidopsis*. *Genes Dev* **12**, 2013–2021 (1998).
32. Kwasniewski, M., Janiak, A., Mueller-Roeber, B. & Szarejko, I. Global analysis of the root hair morphogenesis transcriptome reveals new candidate genes involved in root hair formation in barley. *J Plant Physiol* **167**, 1076–1083, doi:10.1016/j.jplph.2010.02.009 (2010).
33. Edwards, D. & Kenrick, P. The early evolution of land plants, from fossils to genomics: a commentary on Lang (1937) 'On the plant-remains from the Downtonian of England and Wales'. *Philos T R Soc B* **370** (2015).
34. Willis, K. & McElwain, J. *The evolution of plants*. (Oxford University Press, 2013).
35. Smith, A. B. *Systematics and the fossil record: documenting evolutionary patterns*. (John Wiley & Sons, 2009).
36. Hagemann, W. A. F. Cormophytes - Alternative Hypothesis to Telome Theory. *Plant Syst Evol* **124**, 251–277 (1976).
37. Hagemann, W. Towards an organismic concept of land plants: the marginal blastozone and the development of the vegetation body of selected frondose gametophytes of liverworts and ferns. *Plant Syst Evol* **216**, 81–133 (1999).
38. Feild, T. S. *et al*. Fossil evidence for Cretaceous escalation in angiosperm leaf vein evolution. *P Natl Acad Sci USA* **108**, 8363–8366 (2011).
39. Bateman, R. M., Hilton, J. & Rudall, P. J. Morphological and molecular phylogenetic context of the angiosperms: contrasting the 'top-down' and 'bottom-up' approaches used to infer the likely characteristics of the first flowers. *J Exp Bot* **57**, 3471–3503 (2006).
40. Thorne, R. F. How many species of seed plants are there? *Taxon* **51**, 511–512 (2002).
41. Scotland, R. W., Olmstead, R. G. & Bennett, J. R. Phylogeny reconstruction: The role of morphology. *Syst Biol* **52**, 539–548 (2003).
42. Govaerts, R. How many species of seed plants are there? - a response. *Taxon* **52**, 583–584 (2003).
43. Goffinet, B. & William, R. B. *Systematics of the Bryophyta (Mosses): From molecules to a revised classification*". *Monographs in Systematic Botany*. Vol. 98: 205–239 (Molecular Systematics of Bryophytes (Missouri Botanical Garden Press) 2004).
44. Raven, P. H., E., R. F. & E., S. E. *Biology of Plants*. 7th edn, (W. H. Freeman and Company, 2005).
45. Bremer, B. *et al*. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot J Linn Soc* **161**, 105–121 (2009).
46. Simpson, M. G. *Plant systematics*. 2nd edn, 740 (Elsevier Inc., 2010).
47. Mader, S. S. In *Biology* 473–536 (Wm. C. Brown Publishers, 2007).
48. Solda, G. *et al*. Non-random retention of protein-coding overlapping genes in Metazoa. *BMC Genomics* **9**, 174, doi:10.1186/1471-2164-9-174 (2008).
49. Yuan, C. *et al*. Genome-wide view of natural antisense transcripts in *Arabidopsis thaliana*. *DNA Res* **22**, 233–243, doi:10.1093/dnares/dsv008 (2015).
50. Beiter, T., Reich, E., Williams, R. W. & Simon, P. Antisense transcription: a critical look in both directions. *Cell Mol Life Sci* **66**, 94–112, doi:10.1007/s00018-008-8381-y (2009).
51. Neme, R. & Tautz, D. Evolution: dynamics of de novo gene emergence. *Curr Biol* **24**, R238–240, doi:10.1016/j.cub.2014.02.016 (2014).
52. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410, doi:10.1016/S0022-2836(05)80360-2 (1990).
53. Elhaik, E., Sabath, N. & Graur, D. The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. *Mol Biol Evol* **23**, 1–3, doi:10.1093/molbev/msj006 (2006).
54. Moyers, B. A. & Zhang, J. Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol Biol Evol* **32**, 258–267, doi:10.1093/molbev/msu286 (2015).
55. Moyers, B. A. & Zhang, J. Evaluating Phylostratigraphic Evidence for Widespread De Novo Gene Birth in Genome Evolution. *Mol Biol Evol* **33**, 1245–1256, doi:10.1093/molbev/msw008 (2016).
56. Domazet-Loso, T. *et al*. No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. *Mol Biol Evol*. doi:10.1093/molbev/msw284 (2017).
57. Jean, G., Kahles, A., Sreedharan, V. T., De Bona, F. & Ratsch, G. RNA-Seq read alignments with PALMapper. *Curr Protoc Bioinformatics* **Chapter 11**, Unit 11 16, doi:10.1002/0471250953.bi1106s32 (2010).
58. Arendsee, Z. W., Li, L. & Wurtele, E. S. Coming of age: orphan genes in plants. *Trends Plant Sci* **19**, 698–708, doi:10.1016/j.tplants.2014.07.003 (2014).
59. Oh, D. H. *et al*. Genome structures and halophyte-specific gene expression of the extremophile *Thellungiella parvula* in comparison with *Thellungiella salsuginea* (*Thellungiella halophila*) and *Arabidopsis*. *Plant Physiol* **154**, 1040–1052, doi:10.1104/pp.110.163923 (2010).
60. Zuber, H. *et al*. The seed composition of *Arabidopsis* mutants for the group 3 sulfate transporters indicates a role in sulfate translocation within developing seeds. *Plant Physiol* **154**, 913–926, doi:10.1104/pp.110.162123 (2010).
61. Koch, M. A., Haubold, B. & Mitchell-Olds, T. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Mol Biol Evol* **17**, 1483–1498 (2000).
62. Arakaki, M. *et al*. Contemporaneous and recent radiations of the world's major succulent plant lineages. *Proc Natl Acad Sci USA* **108**, 8379–8384, doi:10.1073/pnas.1100628108 (2011).

63. Koch, M. A. & Kiefer, M. Genome evolution among cruciferous plants: a lecture from the comparison of the genetic maps of three diploid species–*Capsella rubella*, *Arabidopsis lyrata* subsp. *petraea*, and *A. thaliana*. *Am J Bot* **92**, 761–767, doi:10.3732/ajb.92.4.761 (2005).
64. Neme, R. & Tautz, D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics* **14** (2013).
65. Rivals, I., Personnaz, L., Taing, L. & Potier, M. C. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* **23**, 401–407 (2007).
66. Du, Z., Zhou, X., Ling, Y., Zhang, Z. H. & Su, Z. agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res* **38**, W64–W70 (2010).

Acknowledgements

We thank Dr. Richard M. Clark for comments on the manuscript. We also thank Dr. Rätsch Gunnar for his contribution to processing the RNA-seq data and getting the gene expression estimates. This work was supported by the National Science Foundation (NSF) (0929262 to C. T. and Richard M. Clark) and Kansas Institutional Development Award (IDeA) Network of Biomedical Research Excellence (K-INBRE) Postdoc Award (to L.L) from the National Institute of General Medical Sciences of the National Institutes of Health (NIH) under grant number P20 GM103418. Contribution number 16–246-J from the Kansas Agricultural Experiment Station.

Author Contributions

L.L. and C.T. conceived the study, managed its organization and edited the draft versions of the manuscript. L.L. performed research (all the different steps), analyzed data and wrote the first draft of the manuscript. E.J.O. performed research (expression data analysis), J.G.S. performed research (producing the RNA-seq data). L.L., E.J.O., J.G.S. and C.T. edited the manuscript. All authors discussed the results, commented on the manuscript, read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-07866-6

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.