Mass-spectrometry-based draft of the *Arabidopsis* proteome

https://doi.org/10.1038/s41586-020-2094-2

Received: 4 June 2019

Accepted: 17 January 2020

Julia Mergner[1], Martin Frejno[1], Markus List[2], Michael Papacek[3], Xia Chen[4], Ajeet Chaudhary[4], Patroklos Samaras[1], Sandra Richter[5], Hiromasa Shikata[6,7], Maxim Messerer[8], Daniel Lang[8], Stefan Altmann[9], Philipp Cyprys[10], Daniel P. Zolg[1], Toby Mathieson[11], Marcus Bantscheff[11], Rashmi R. Hazarika[12,13], Tobias Schmidt[1], Corinna Dawid[14], Andreas Dunkel[14], Thomas Hofmann[14], Stefanie Sprunck[10], Pascal Falter-Braun[9,15], Frank Johannes[12,13], Klaus F. X. Mayer[8,16], Gerd Jürgens[5], Mathias Wilhelm[1], Jan Baumbach[2], Erwin Grill[3], Kay Schneitz[4], Claus Schwechheimer[6] & Bernhard Kuster[1,13,17] ✉

Plants are essential for life and are extremely diverse organisms with unique molecular capabilities[1]. Here we present a quantitative atlas of the transcriptomes, proteomes and phosphoproteomes of 30 tissues of the model plant *Arabidopsis thaliana*. Our analysis provides initial answers to how many genes exist as proteins (more than 18,000), where they are expressed, in which approximate quantities (a dynamic range of more than six orders of magnitude) and to what extent they are phosphorylated (over 43,000 sites). We present examples of how the data may be used, such as to discover proteins that are translated from short open-reading frames, to uncover sequence motifs that are involved in the regulation of protein production, and to identify tissue-specific protein complexes or phosphorylation-mediated signalling events. Interactive access to this resource for the plant community is provided by the ProteomicsDB and ATHENA databases, which include powerful bioinformatics tools to explore and characterize *Arabidopsis* proteins, their modifications and interactions.

The plant model organism *A. thaliana* has revolutionized our understanding of plant biology and influenced many other areas of the life sciences[1]. Knowledge derived from *Arabidopsis* has also provided mechanistic understanding of important agronomic traits in crop species[2]. The *Arabidopsis* genome was sequenced 20 years ago, and hundreds of natural variants have since been analysed at the genome and epigenome level[3,4]. By contrast, the *Arabidopsis* proteome as the main executer of most biological processes is far less comprehensively characterized. To address this gap, we used state-of-the-art mass spectrometry and RNA sequencing (RNA-seq) analyses to provide the first, to our knowledge, integrated proteomic, phosphoproteomic and transcriptomic atlas of *Arabidopsis*. We use select examples to show how this rich molecular resource can be used to explore the function of single proteins or entire pathways across several omics levels.

Multi-omics atlas of *Arabidopsis*

We generated an expression atlas covering, on average, 17,603 ± 1,317 (mean ± s.d.) transcripts, 14,430 ± 911 proteins and 14,689 ± 2,509 phosphorylation sites (P-sites) per tissue, using a reproducible biochemical and analytical approach (Fig. 1a, b, Extended Data Fig. 1a–c, Supplementary Data 1, 2). In total, the protein abundance data cover 18,210 out of the 27,655 protein-coding genes (66%) annotated in Araport11[5]. This is a substantial increase compared with the percentage of genes with evidence at the protein level reported in UniProt (27%)[6], and more than double the number of proteins identified in an earlier tissue proteome analysis[7] (Fig. 1c, Extended Data Fig. 1d–f). In addition, we report tissue-resolved quantitative evidence for a total of 43,903 P-sites, making this study one of the most comprehensive single *Arabidopsis* phosphoproteomes published so far (Fig. 1c). We found that 47% of the expressed proteome was phosphorylated in at least one instance, confirming earlier analyses of individual phosphoproteomes or information assembled in the PhosPhAT database[8,9] (Extended Data Fig. 1g, h). The actual figures in plantae are likely to be considerably higher as there is limited overlap between these data sources. This can be readily explained by technical factors (such as insufficient sequence coverage) and, more importantly, differences in plant genotypes or stimuli that influence phosphorylation. The above figures are similar to those

---

[1]Chair of Proteomics and Bioanalytics, Technical University of Munich (TUM), Freising, Germany. [2]Chair of Experimental Bioinformatics, Technical University of Munich (TUM), Freising, Germany. [3]Chair of Botany, Technical University of Munich (TUM), Freising, Germany. [4]Plant Developmental Biology, Technical University of Munich (TUM), Freising, Germany. [5]Center for Plant Molecular Biology, University of Tübingen, Tübingen, Germany. [6]Chair of Plant Systems Biology, Technical University of Munich (TUM), Freising, Germany. [7]Devision of Plant Environmental Responses, National Institute for Basic Biology, Okazaki, Japan. [8]Plant Genome and Systems Biology, Helmholtz Center Munich, German Research Center for Environmental Health, Munich-Neuherberg, Germany. [9]Institute of Network Biology (INET), Helmholtz Center Munich, German Research Center for Environmental Health, Munich-Neuherberg, Germany. [10]Cell Biology and Plant Biochemistry, University of Regensburg, Regensburg, Germany. [11]Cellzome GmbH, Heidelberg, Germany. [12]Population Epigenetics and Epigenomics, Technical University of Munich (TUM), Freising, Germany. [13]Institute of Advanced Study (IAS), Technical University of Munich (TUM), Freising, Germany. [14]Chair of Food Chemistry and Molecular Sensory Science, Technical University of Munich (TUM), Freising, Germany. [15]Chair of Microbe-Host Interactions, Ludwigs-Maximilians-University (LMU), Munich, Germany. [16]Plant Genome Biology, Technical University of Munich (TUM), Freising, Germany. [17]Bavarian Biomolecular Mass Spectrometry Center (BayBioMS), Technical University of Munich (TUM), Freising, Germany. ✉e-mail: kuster@tum.de

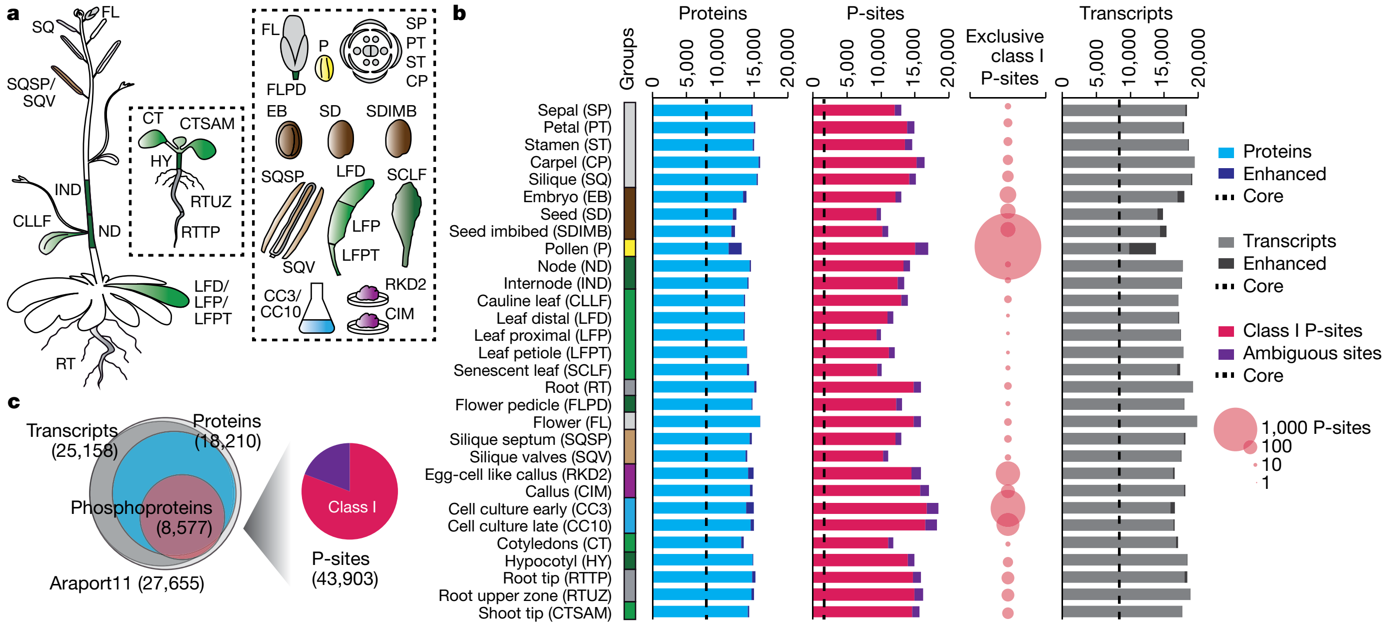


Fig. 1 | Tissue map and multi-omics dataset. a, Schematic of tissue samples analysed, coloured according to morphology group (abbreviations defined in **b**): flower (light grey); seed (dark brown); pollen (yellow); stem (dark green); leaf (light green); root (dark grey); fruit (light brown); callus (magenta); cell culture (blue). **b**, Number of identifications at the protein, P-site and transcript levels for all tissues ($n = 1$ measurement per tissue). Dashed lines indicate the number of core proteins, P-sites or transcripts detected in all tissues. Tissue-enhanced proteins or transcripts are marked by a darker colour. P-sites with high-confidence amino acid localization (class I sites; more than 0.75 localization probability) are shown in pink; ambiguous site localizations are in purple. The number of P-sites exclusively detected in one tissue is shown by circles. **c**, Total number and overlap of identified gene loci in the transcriptome, proteome and phosphoproteome datasets compared with Araport11 (left), and the total number of identified P-sites and the proportion of class I sites (right).

reported for mammalian systems, which underscores the strong conservation of post-translational protein modifications as a means to create functional diversity from a limited set of genes and proteins[10].

The atlas can be explored using the web portals ATHENA (Arabidopsis THaliana ExpressioN Atlas; athena.proteomics.wzw.tum.de) and ProteomicsDB (www.proteomicsdb.org)[11]. Analysis and visualization options include co-expression analysis, the exploration of tissue-specific interaction networks and a tool for predicting tandem mass spectra of peptide sequences, which can be used to validate reported peptide identifications[12] (Fig. 2a, b, Supplementary Table 1). ProteomicsDB further provides spectral libraries for more than 328,000 unmodified and more than 43,000 phosphorylated peptides. These can facilitate the development of quantitative mass-spectrometry-based protein assays, which is particularly useful for research in *Arabidopsis*, for which comparatively few antibodies exist.

Proteomic annotation of the genome

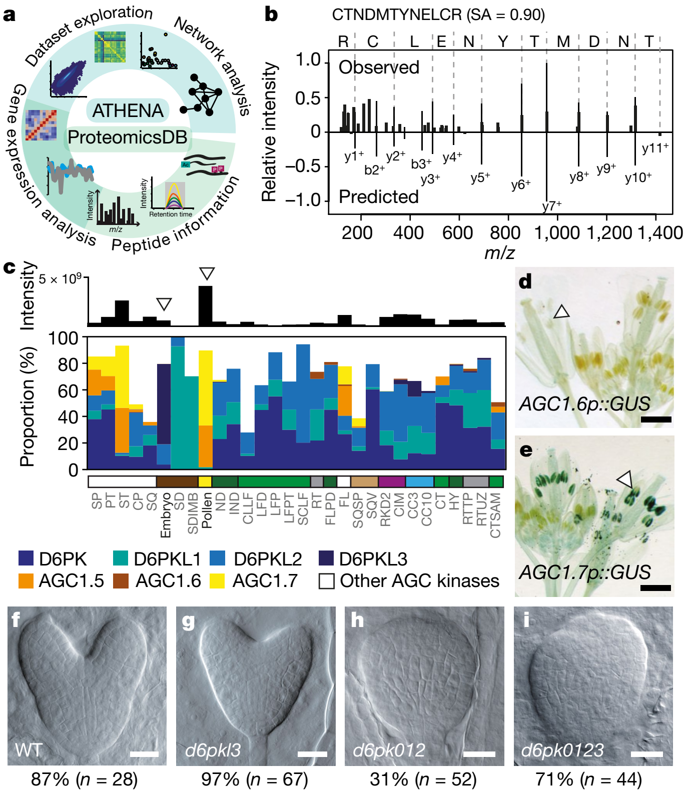
Using Araport11, the proteomic data corroborated a substantial number of annotated open-reading frame borders based on the detection of 2,776 N-terminal and 2,656 C-terminal peptides (Extended Data Fig. 2a). As expected, N-terminal peptides often showed cleavage of the initiator methionine and N-terminal acetylation that was strongly dependent on the amino acid adjacent to the initiator methionine[13] (Extended Data Fig. 2b, c). The mass spectrometry data covered, on average, 44% of each protein sequence, enabling the detection of unique peptides for 14,115 protein isoforms. In 297 cases, the data also distinguished 2 or 3 splice variants (Extended Data Fig. 2d–f, Supplementary Data 3). A selection of these isoform-specific peptides was validated using synthetic peptides (confirmation rate 80%) (Extended Data Fig. 2g). To investigate the potential translation of predicted short open-reading frames (sORFs), we searched the proteomic data against a published and an in-house-compiled sORF collection (ARA-PEP[14]; ATSO), which led to the identification of 51 distinct sORFs that were subsequently confirmed using synthetic peptides (Extended Data Fig. 2g, h, Supplementary Data 3). These results demonstrate the potential of the atlas to refine *Arabidopsis* gene models and protein sequences.

Quantitative expression landscapes

The dynamic range of protein and transcript expression spanned six and four orders of magnitude, respectively (Extended Data Fig. 2i, j). As previously described, protein evidence was underrepresented for low-abundance transcripts[15], which indicates that proteome coverage was not yet exhaustive. Low-abundance transcripts were enriched for Gene Ontology (GO) terms such as cell signalling or regulation of gene expression (Supplementary Data 4). Still, the protein data cover approximately 50% of all annotated transcription factors and 75% of all transcriptional regulators, kinases and phosphatases, many of which are themselves regulated post-translationally[16,17] (Extended Data Fig. 2k). Notably, protein phosphorylation was detected across the entire dynamic range of protein abundance, which highlights the efficiency of the phosphopeptide enrichment used (Extended Data Fig. 2j).

A breakdown of the molecular data for individual tissues or morphologically similar groups (for example, leaf, seed or root) showed that only few transcripts or proteins are expressed in a tissue-specific manner (Extended Data Fig. 3a). Instead, tissue types are characterized by distinct quantitative abundance patterns. For example, the floral organs express a similar set of genes but at levels that are characteristic for each flower compartment. Here, the MADS-domain transcription factors showed the expected differential expression, thus recapitulating established models of floral organ identity[18] (Extended Data Fig. 3b, c).

The strong variation in quantitative gene expression between tissue groups was evident at both the protein and transcript levels (Extended Data Fig. 3d, e). Notably, we found little overlap in the rank order or even identity of the most abundant proteins and mRNAs within any



87% (*n* = 28) 97% (*n* = 67) 31% (*n* = 52) 71% (*n* = 44)

Fig. 2 | Data exploration in ATHENA and ProteomicsDB. a, Schematic of data analysis options in the ATHENA and ProteomicsDB databases. **b**, Mirror plot showing near identical observed and predicted tandem mass spectra ($n = 1$ acquired spectra) for a peptide from the 'uncertain' protein AT4G13955, thus validating its existence as a protein. 'SA' denotes the normalized spectral contrast angle—a measure for spectrum similarity. **c**, Top, summed absolute protein abundance of AGC1 and AGC3 subfamilies of AGCVIII kinases. Bottom, relative protein abundance profiles of D6PK family members and the closely related kinases AGC1.5, AGC1.6 and AGC1.7. Arrowheads indicate embryo and pollen tissue with high abundance of D6PKL3 or AGC1.5 and AGC1.7, respectively. **d**, **e**, Representative images of GUS-stained flowers expressing *AGC1.6* ($n = 20$ independent transgenic lines) and *AGC1.7* ($n = 12$ independent transgenic lines) promotor–GUS fusions. Arrowheads indicate anthers containing pollen grains. Scale bars, 1 mm. **f**–**i**, Representative images of embryo phenotypes (at heart stage or age-matched equivalent) and their respective frequency for wild-type (WT), *d6pkl3* single, *d6pk012* triple (*d6pk d6pkl1 d6pkl2*) and *d6pk0123* (*d6pk d6pkl1 d6pkl2 d6pkl3*) quadruple knockout mutants. Scale bars, 20 µm.

given tissue (Extended Data Fig. 3e, f). Extreme expression ranges were detected for seeds, for example, in which the storage protein CRA1 accounted for approximately 10% of the total protein and the ten most abundant gene products already represented more than 30% of the total amount of protein or transcript, respectively (Extended Data Fig. 3e, g). Similarly, the RuBisCO complex—renowned as the most abundant plant protein on the planet—accounted for 4–7% of the total leaf protein content[19] (Extended Data Fig. 3g). We note that despite the dominance of some proteins in certain tissues, we show that the proteomes of plant tissues can be probed very deeply using current technology, countering the long-standing notion that plant proteomics is particularly difficult[20]. As expected, photosynthetic activity was one of the main factors for variations in expression patterns between tissues, which can also be related to plastid localized proteins and their abundance (Extended Data Fig. 3h, i). The most divergent expression pattern at both the transcript and protein levels, however, was found for pollen (Extended Data Fig. 3h).

The comparison of expression patterns between tissues can be used to ascertain biological functions of individual proteins or to disentangle closely related protein family members. A closer look at the AGCVIII kinase family (23 members in *Arabidopsis*)[21] revealed high protein production of members of the D6 protein kinases (D6PK) subfamily in embryos as well as of AGC1.5 and AGC1.7 kinases in flowers and pollen (Fig. 2c). Consistent with these results, follow-up experiments showed promotor activity for *AGC1.5* and *AGC1.7* but not for the closely related *AGC1.6* gene in pollen, which strongly supports the contribution of the former two in pollen tube growth[22] (Fig. 2d, e). We further uncovered an important role for the D6PKs in embryo development, in which a combined knockout synergistically increased the occurrence of aberrant embryo phenotypes in loss-of-function mutants (Fig. 2f–i).

Regulation of protein amount in tissues

The amount of protein for a given gene is determined by a regulatory system comprising anabolic and catabolic transcriptional, translational and post-translational processes. This has been extensively studied in unicellular organisms and, more recently, in human[23]. Although this broad area of research cannot be covered in detail here, consistent observations were made. We found positive correlation between transcript and protein levels in most tissues (Pearson's correlation $r = 0.28$–0.7) (Extended Data Fig. 4a), with most high- or low-abundance transcripts resulting in high- or low-abundance proteins, respectively[24] (Fig. 3a). To identify further molecular determinants of protein abundance, we implemented a model selection approach and tested each of the selected features for the potential to explain protein-level variations on a tissue-specific or global level (Extended Data Fig. 4b–e, Supplementary Data 5). In addition to well-known predictors, such as transcript levels and codon usage[25], evolutionary conservation, mRNA sequence motifs and the number of protein interactions were important predictors of protein levels (Fig. 3b). However, a large proportion (48%) of variation in protein abundance still remains unexplained, which suggests that many other molecular factors with great individual but lower global effect may be at work that will require careful case-by-case experiments to resolve.

Protein-versus-transcript abundance plots clearly show that similarly abundant transcripts often lead to proteins with more than a 100-fold difference in abundance. These differences in the protein-to-mRNA ratio (PTR) of genes might be due to differences in translation efficiency of a given transcript, transcript stability and/or the stability of the respective protein[26] (Fig. 3c). High PTRs were detected for genes involved in photosynthesis or energy metabolism (such as *petA* and *rbcL*) (Supplementary Data 4). This indicates optimized translation and stability of those proteins (and transcripts) that are needed in large quantities. Conversely, genes with low PTRs were enriched in processes such as hormone-mediated signalling pathways, which, in plants, often involve protein degradation[27]. Members of the auxin-labile AUX/IAA protein family (such as IAA8 and IAA13) show low protein signals despite high transcript levels, and are known to be under tight control of auxin-dependent proteasomal degradation[28] (Fig. 3a).

We also observed inter-tissue variation of PTRs, especially for seed and pollen (Fig. 3d). Genes with low PTRs in seeds mainly show differential abundance at the protein level, and proteome-wide time-resolved measurements in germinating seeds in response to treatment with cycloheximide (CHX) (translation block) or MG132 (proteasome block) suggest that low PTR proteins are rapidly translated from stored mRNA after germination. Proteins with high PTRs do not show this behaviour, which indicates that they are stored in seed to be readily available for germination (Fig. 3e, f, Extended Data Fig. 4f, g). This is consistent with reports that seeds and pollen accumulate storage molecules including mRNAs and proteins to allow rapid development at the onset of germination, and may explain the observed uncoupling of transcription

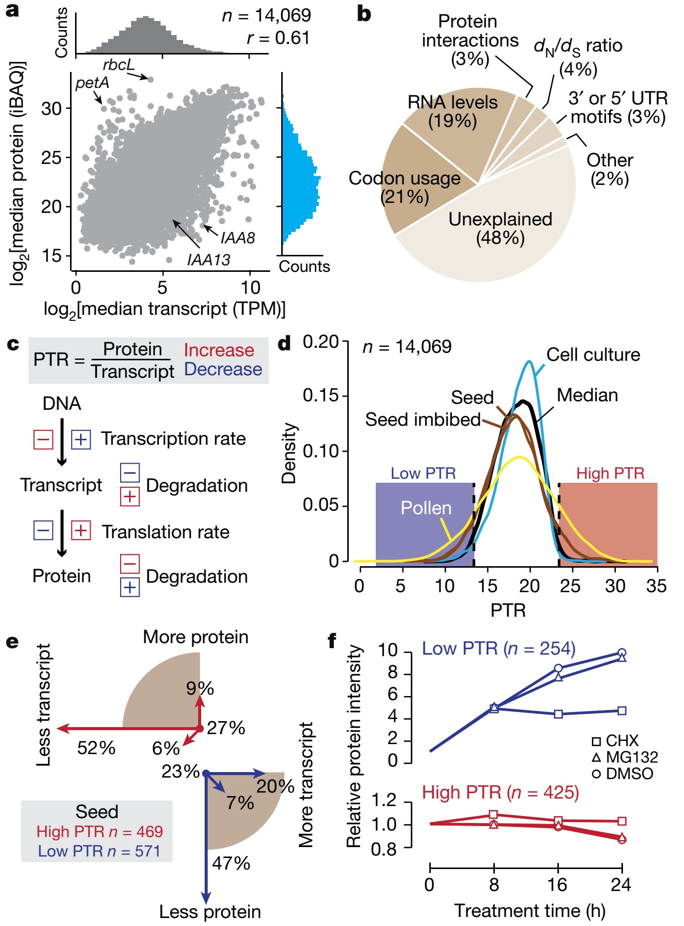


Fig. 3 | Protein and mRNA expression. a, Scatter plot of protein-versus-transcript abundance (median across 30 tissues). Arrows mark examples of genes with a high PTR (*rbcL* and *petA*) or low PTR (*IAA8* and *IAA13*). *r* denotes Pearson's correlation coefficient. TPM, transcripts per million. **b**, Relative contribution of the molecular features (averaged across tissues) that can predict protein abundance. d_N/d_S, ratio of non-synonymous-to-synonymous nucleotide substitutions; UTR, untranslated region. **c**, Schematic of the steps in which modulation of PTRs may occur. For example, the higher the transcription rate the lower the PTR. **d**, PTR distribution of genes for selected tissues. Genes with high and low PTRs are defined as being two standard deviations outside of the median PTR distribution. **e**, Analysis of the proportion of genes with high or low PTRs that are at least fourfold differently abundant at the transcript and/or protein level in seeds compared with all other tissues (depicted as arrows). The percentage of genes below the fourfold criterion is placed at the origin of the plot. *n* denotes the number of genes with high or low PTRs in seed as defined in **d**. **f**, Average relative changes in protein abundance for genes with high or low PTRs (from **e**) in seeds incubated for 8 h, 16 h or 24 h with 100 μM CHX (translation block), MG132 (proteasome block) or dimethylsulfoxide (DMSO; control).

and translation in these durable plant tissues[29]. Fluctuating PTRs of genes between tissues indicate tissue-specific regulation at either the transcript or the protein level. By contrast, genes with stable PTRs are probably under similar regulation in different tissues, and the extensive transcriptomic resources available for *Arabidopsis* may therefore be used to estimate relative protein abundances (Extended Data Fig. 5a, b). Similarly, stable ratios between the abundance of P-sites and corresponding protein levels indicates that, under steady-state conditions, most changes in the abundance of phosphorylation between tissues can be attributed to fluctuations in the underlying protein abundance rather than phosphorylation stoichiometry (Extended Data Fig. 5c, d).

Co-expression of paralogues and complexes

On the basis of the 'guilt-by-association' idea, gene co-expression analysis is widely used to ascribe common functionality between genes and can be performed at the transcript or protein level. We observed that co-expressed gene pairs also have high STRING[30] scores, indicative of physical or functional interactions (Extended Data Fig. 6a). To illustrate this point, we focused on co-expression analyses of gene paralogues and interacting proteins. Similar to most plants, *Arabidopsis* has undergone genome duplications and retains a considerable number of highly conserved paralogues in the genome[31]. Compared with randomly selected gene pairs, duplicated genes show a clear shift towards positive expression correlations (both transcript and protein). This indicator for redundant gene function is supported by the finding that knockout mutants of the more abundant paralogues are more likely to display a measureable plant phenotype[32] (Extended Data Fig. 6b–d, Supplementary Data 6). Such comparative expression analyses can be useful to prioritize knockout combinations for phenotype analyses of functionally redundant genes.

Analysing proteins that engage in physical interactions recorded in AtPIN[33] for co-expression across tissues suggests that approximately 26% of AtPIN pairs may be stable rather than transient complex partners (Extended Data Fig. 6e, f, Methods). As expected, protein-based correlations were consistently higher than transcript correlations. Because co-expression analyses tend to generate a large number of interaction candidates, we sought to prioritize these using data from size-exclusion chromatography–mass spectrometry (SEC–MS) experiments as an independent experimental approach[34] (Extended Data Fig. 7a–c, Supplementary Data 7, 8). We found that co-detection in SEC–MS, in addition to co-expression in the tissue atlas, substantially reduced the list of candidates, illustrated by the detection of the coatomer complex in the intersection of the SEC–MS and tissue atlas data (Fig. 4a, b). The identification of known protein complexes was consistently improved using this approach, which combines information from size-exclusion chromatography (SEC) and the tissue atlas datasets (Extended Data Fig. 7d, Supplementary Data 9, Methods).

Notably, in seed tissues, the ζ-subunit of the coatomer complex was almost exclusively provided by the ζ-1 paralogue, although this was almost absent in all other tissues (Extended Data Fig. 7e). Comparisons of subunit abundance also provided information about complex stoichiometry, and similar results were obtained for stable complexes when using the SEC–MS or the tissue atlas dataset (Fig. 4c, Extended Data Fig. 7f, g). We therefore propose that our tissue atlas resource can provide an initial approximation of the relative subunit frequency of known complexes with unresolved stoichiometry. This could be particularly useful for membrane- or cell-wall-associated complexes, which are difficult to study biochemically.

The *Arabidopsis* phosphoproteome

The number of detected kinases (642 ± 55) and phosphatases (119 ± 6) as well as the ratios of individual families were comparable across most tissues. Pollen and egg cell-like callus, however, stood out, which indicates extensive signalling activity in these tissues (Extended Data Fig. 8a–c). In contrast to the breadth of available phosphoproteomic data, information on kinase–substrate relationships, which is crucial for understanding signalling cascades or pathway activities, is still sparse. Because these relationships are difficult to discover experimentally, scientists often initially take a computational approach. We used the motif-X algorithm[35] to identify 266 phosphorylation motifs that grouped into 'proline-directed', 'acidic', 'basic' or 'other' motif classes[36] (Extended Data Fig. 8d–f, Supplementary Table 2, Supplementary Data 10). Together with information about the co-occurrence of kinases and P-sites in tissues, as well as external co-localization or interaction information, we anticipate that this dataset can help

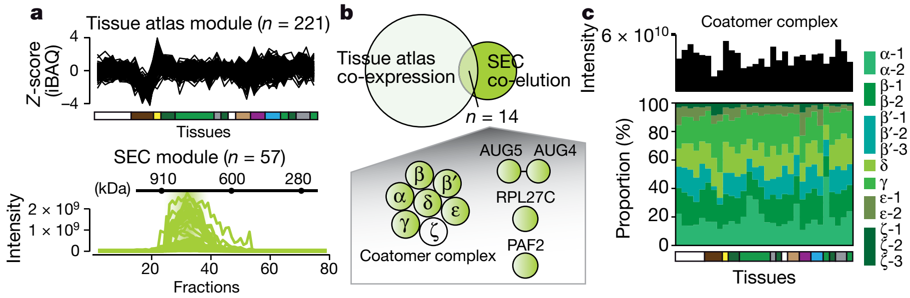


Fig. 4 | Characterization of protein complexes by protein co-expression and SEC–MS. a, Top, 221 proteins that form a module in the tissue atlas based on their co-expression across all tissues. Bottom, co-elution of 57 proteins in a SEC–MS experiment using flower tissue. **b**, Overlap of proteins ($n = 14$) between the groups of co-expressed and co-eluting proteins in **a**. Ten of the proteins in the intersection are subunits of the coatomer complex. **c**, Tissue-resolved absolute abundance (top) and relative proportions of coatomer complex subunits based on the tissue atlas protein abundance data (bottom) showing heterogeneous expression but a fairly homogeneous composition of the complex across tissues. Paralogues of the α, β, β′, ε and ζ coatomer subunits were combined for the analysis. Tissue groups are coloured as in Fig. 1.

untangle kinase–substrate and kinase–P-site relationships in the future.

Overall, the number of P-sites per protein showed vast variation. For example, members of the LEA protein family[37] are phosphorylated at almost every serine, threonine or tyrosine residue in their sequences (Fig. 5a, Extended Data Fig. 8g, h). Phosphorylation of these unstructured proteins involved in seed maturation and desiccation may be a mechanism to regulate their conformational state or phase transitions[38]. Conversely, other proteins showed clusters of P-sites in regulatory domains, including the juxtamembrane domain of receptor-like kinases[39] (Extended Data Fig. 8i), which indicates a role in recruiting interaction partners akin to human receptor tyrosine kinases.

The detection of a phosphorylation event does not directly indicate a functional consequence, but it does provide an important starting point[40]. For example, the abscisic acid (ABA) receptor RCAR10, which is involved in the ABA signalling cascade, was found to be phosphorylated at four sites. We generated phosphomimetic mutants (S113D and S32D) that displayed reduced ABA responses compared with the wild-type receptor (Fig. 5b). For Ser113, which is part of the binding surface for type 2C protein phosphatases (PP2C), this reduced response might be linked to an effect on ABA binding or PP2C interaction. In the case of Ser32, which is localized at the receptor periphery, the observed effect is more likely to be due to altered interactions with other regulatory proteins[41] (Extended Data Fig. 9a, b). Notably, both of the RCAR10 phosphomimetic variants altered the ABA response of RCAR10–PP2C co-receptor combinations, which suggests changes in the interaction profile and phosphorylation-based fine-tuning of ABA signalling during development and the stress response (Extended Data Fig. 9c, d).

As a second example for functional consequences of protein phosphorylation, we studied QKY (also known as MCTP15), a protein localized to plasmodesmata and involved in floral and silique morphogenesis[42]. Because one of its P-sites (Ser262) was detected in all assayed tissues and further P-sites were found between the first and second C2 domains in other MCTP family members, we speculated about a possible effect of phosphorylated Ser262 on QKY function (Extended Data Fig. 9e, f). To test this hypothesis, we generated plant lines that express phosphomimetic and phosphomutant *QKY* transgenes in the strong *qky-9* mutant background[43]. The phosphomimetic but not the phosphomutant protein constructs rescued both flower and silique mutant phenotypes, suggesting that phosphorylation of Ser262 is required for normal QKY function (Fig. 5c–f, Extended Data Fig. 9g–j).

Together, we have generated the most comprehensive, albeit still incomplete, draft of the *Arabidopsis* (phospho)proteome so far, and highlighted some of the many uses that can be envisioned for this resource. However, much remains to be accomplished, including a more systematic coverage of protein sequence variants or post-translational modifications. The tens of thousands of newly discovered phosphorylation sites that await functional characterization present a particular future challenge. The examples we provide show that investigating different levels of omics data can lead to new insights into biological processes. We therefore anticipate that this resource along with the provided online computational tools will enable the *Arabidopsis* research community to perform many types of systems-level analyses

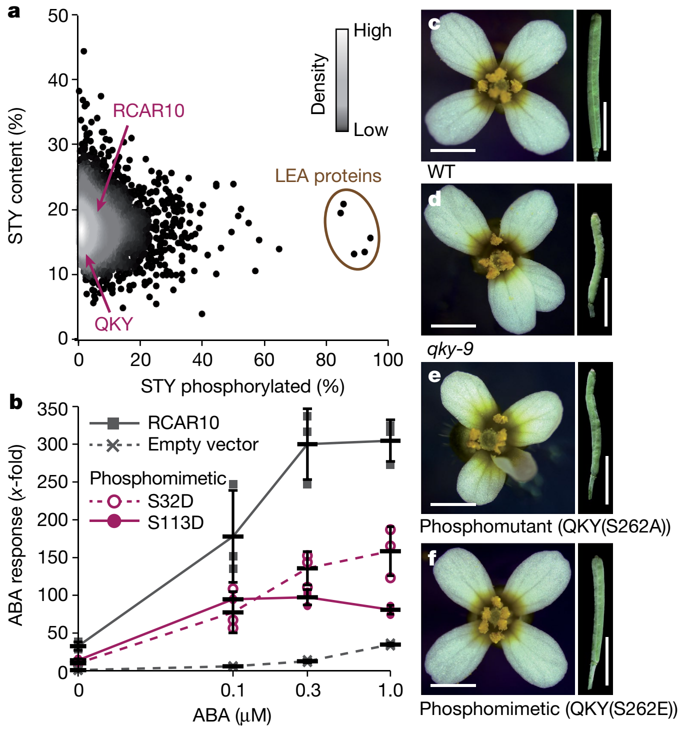


Fig. 5 | Ascribing function to protein phosphorylation. a, Scatter plot of the proportion of potential phosphorylation acceptor sites in a protein sequence (limited to Ser, Thr and Tyr) versus the proportion of Ser-Thr-Tyr (STY) residues found to be phosphorylated in this study. Although most proteins, including RCAR10 and QKY, show only a small proportion of phosphorylated residues, the LEA proteins are phosphorylated at almost every possible residue. **b**, ABA response (mean ± s.d.; $n = 3$) of ABA-treated protoplast cells transformed with wild-type or phosphomimetic variants of RCAR10 (S32D and S113D), which suggests that phosphorylation of Ser32 and Ser113 are functionally involved in RCAR10 activity. **c**–**f**, Representative images of flowers and siliques of wild-type, *qky-9* mutant and *qky-9* expressing phospho-dead QKY(S262A) or phosphomimetic QKY(S262E) constructs show that QKY(S262E) but not QKY(S262A) rescues the *qky-9* phenotype. Scale bars, 1 mm (flower) and 0.5 cm (silique).

not covered here. We also expect that the study will more broadly affect plant research, as the work shows that mass-spectrometry-based quantitative protein assays are a way to overcome the lack of antibodies for plant research.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2094-2.

1. Krämer, U. Planting molecular functions in an ecological context with *Arabidopsis thaliana*. *eLife* **4**, (2015).
2. Peng, J. et al. 'Green revolution' genes encode mutant gibberellin response modulators. *Nature* **400**, 256–261 (1999).
3. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).
4. Kawakatsu, T. et al. Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. *Cell* **166**, 492–505 (2016).
5. Cheng, C. Y. et al. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J*. **89**, 789–804 (2017).
6. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. **45** (D1), D158–D169 (2017).
7. Baerenfaller, K. et al. Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science* **320**, 938–941 (2008).
8. van Wijk, K. J., Friso, G., Walther, D. & Schulze, W. X. Meta-analysis of *Arabidopsis thaliana* phospho-proteomics data reveals compartmentalization of phosphorylation motifs. *Plant Cell* **26**, 2367–2389 (2014).
9. Durek, P. et al. PhosPhAt: the *Arabidopsis thaliana* phosphorylation site database. An update. *Nucleic Acids Res*. **38**, D828–D834 (2010).
10. Sharma, K. et al. Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep*. **8**, 1583–1594 (2014).
11. Schmidt, T. et al. ProteomicsDB. *Nucleic Acids Res*. **46** (D1), D1271–D1281 (2018).
12. Gessulat, S. et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).
13. Bienvenut, W. V. et al. Comparative large scale characterization of plant versus mammal proteins reveals similar and idiosyncratic *N*-α-acetylation features. *Mol. Cell. Proteomics* **11**, mcp.M111.015131 (2012).
14. Hazarika, R. R. et al. ARA-PEPs: a repository of putative sORF-encoded peptides in *Arabidopsis thaliana*. *BMC Bioinformatics* **18**, 37 (2017).
15. Wilhelm, M. et al. Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).
16. Zheng, Y. et al. iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* **9**, 1667–1670 (2016).
17. Yang, M. et al. A comprehensive analysis of protein phosphatases in rice and *Arabidopsis*. *Plant Syst. Evol*. **289**, 111–126 (2010).
18. Litt, A. & Kramer, E. M. The ABC model and the diversification of floral organ identity. *Semin. Cell Dev. Biol*. **21**, 129–137 (2010).
19. Bar-On, Y. M. & Milo, R. The global mass and average rate of rubisco. *Proc. Natl Acad. Sci. USA* **116**, 4738–4743 (2019).
20. Gupta, R. et al. Time to dig deep into the plant proteome: a hunt for low-abundance proteins. *Front Plant Sci* **6**, 22 (2015).
21. Galván-Ampudia, C. S. & Offringa, R. Plant evolution: AGC kinases tell the auxin tale. *Trends Plant Sci*. **12**, 541–547 (2007).
22. Zhang, Y., He, J. & McCormick, S. Two *Arabidopsis* AGC kinases are critical for the polarized growth of pollen tubes. *Plant J*. **58**, 474–484 (2009).
23. Eraslan, B. et al. Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. *Mol. Syst. Biol*. **15**, e8513 (2019).
24. Liu, Y., Beyer, A. & Aebersold, R. On the dependency of cellular protein levels on mRNA abundance. *Cell* **165**, 535–550 (2016).
25. Hanson, G. & Coller, J. Codon optimality, bias and usage in translation and mRNA decay. *Nat. Rev. Mol. Cell Biol*. **19**, 20–30 (2018).
26. Schwanhäusser, B. et al. Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
27. Santner, A. & Estelle, M. The ubiquitin-proteasome system regulates plant hormone signaling. *Plant J*. **61**, 1029–1040 (2010).
28. Luo, J., Zhou, J. J. & Zhang, J. Z. Aux/IAA gene family in plants: molecular structure, regulation, and function. *Int. J. Mol. Sci*. **19**, E259 (2018).
29. Bai, B. et al. Seed stored mRNAs that are specifically associated to monosome are translationally regulated during germination. *Plant Physiol*. **182**, 378–392 (2019).
30. Szklarczyk, D. et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res*. **45** (D1), D362–D368 (2017).
31. Wang, Y., Tan, X. & Paterson, A. H. Different patterns of gene structure divergence following gene duplication in *Arabidopsis*. *BMC Genomics* **14**, 652 (2013).
32. Lloyd, J. & Meinke, D. A comprehensive dataset of genes with a loss-of-function mutant phenotype in *Arabidopsis*. *Plant Physiol*. **158**, 1115–1129 (2012).
33. Brandão, M. M., Dantas, L. L. & Silva-Filho, M. C. AtPIN: *Arabidopsis thaliana* protein interaction network. *BMC Bioinformatics* **10**, 454 (2009).
34. Kristensen, A. R., Gsponer, J. & Foster, L. J. A high-throughput approach for measuring temporal changes in the interactome. *Nat. Methods* **9**, 907–909 (2012).
35. Schwartz, D. & Gygi, S. P. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat. Biotechnol*. **23**, 1391–1398 (2005).
36. Villén, J., Beausoleil, S. A., Gerber, S. A. & Gygi, S. P. Large-scale phosphorylation analysis of mouse liver. *Proc. Natl Acad. Sci. USA* **104**, 1488–1493 (2007).
37. Battaglia, M., Olvera-Carrillo, Y., Garciarrubio, A., Campos, F. & Covarrubias, A. A. The enigmatic LEA proteins and other hydrophilins. *Plant Physiol*. **148**, 6–24 (2008).
38. Bah, A. et al. Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch. *Nature* **519**, 106–109 (2015).
39. Mitra, S. K. et al. An autophosphorylation site database for leucine-rich repeat receptor-like kinases in *Arabidopsis thaliana*. *Plant J*. **82**, 1042–1060 (2015).
40. Landry, C. R., Levy, E. D. & Michnick, S. W. Weak functional constraints on phosphoproteomes. *Trends Genet*. **25**, 193–197 (2009).
41. Hauser, F., Li, Z., Waadt, R. & Schroeder, J. I. SnapShot: abscisic acid signaling. *Cell* **171**, 1708–1708 (2017).
42. Vaddepalli, P. et al. The C2-domain protein QUIRKY and the receptor-like kinase STRUBBELIG localize to plasmodesmata and mediate tissue morphogenesis in *Arabidopsis thaliana*. *Development* **141**, 4139–4148 (2014).
43. Fulton, L. et al. DETORQUEO, QUIRKY, and ZERZAUST represent novel components involved in organ development mediated by the receptor-like kinase STRUBBELIG in *Arabidopsis thaliana*. *PLoS Genet*. **5**, e1000355 (2009).

Methods

Data reporting

No statistical methods were used to predetermine sample size. Samples for mass spectrometry and RNA-seq measurements were prepared and measured in a random order. The investigators were not blinded to allocation during experiments and outcome assessment.

Sample preparation

Inflorescence and seed samples. *A. thaliana* wild-type Columbia-0 (Col-0) plants were grown under continuous white light conditions at 22 °C. Samples for flower parts and siliques were collected from mature plants. Stage 15 flowers[44] were dissected into petal, sepal, stamen and carpel. Fully grown green siliques were separated into septum, valves and green seeds (stage 10). Surface-sterilized mature dry seeds were stored for 2 days at 4 °C and subsequently imbibed for 24 h or kept dry. For pollen collection, plants were grown on soil under a long photoperiod (16 h of light, 21 °C, 65% humidity). Mature pollen was bulk-harvested from open flowers at developmental stage 13[44]. A vacuum cleaner was modified with three subsequent nylon meshes (80 µm, 35 µm, 10 µm mesh) for large-scale pollen isolation as previously described[45]. Pollen was collected in a 1.5-ml reaction tube, snap-frozen in liquid nitrogen and stored at −80 °C until further use.

Cell culture and callus samples. Cell culture samples (root cells, Col-0) were grown in medium composed of 4.3 g l^{-1} Murashige and Skoog basal medium (MS), 30 g l^{-1} sucrose and 0.33 mg ml^{-1} KH_2PO_4, supplemented with 2,4-dichlorophenoxyacetic acid (2,4-D) (final concentration 1 mg l^{-1}; pH adjusted to 5.8 with KOH) at 22 °C under continuous light and obtained either 3 or 10 days after sub-culturing. To generate callus-inducing medium (CIM) callus, root explants were collected from two-week-old seedlings (Col-0) grown in sterile culture on MS plates. Then, 5–10 mm long root segments were cultured on CIM medium composed of 1× Gamborg's B-5 salts, 20 g l^{-1} glucose, 0.5 g l^{-1} MES, 1× Gamborg's vitamin solution and 1% Phytoagar supplemented with 2,4-D (500 µg l^{-1}) and kinetin (50 µg l$_{-1}$). CIM calli appeared after 7–10 days and were propagated in 2-week subculture intervals on CIM medium. To generate the callus line with an egg-cell-like expression profile, the coding sequence of RKD2 (AT1G74480) was amplified from pistil cDNA using the primer pair RKD2fw and RKD2rev (Supplementary Data 11). Pistils from flowers at developmental stage 12[45] were collected to purify mRNA and generate cDNA as previously described[46]. The PCR fragment was cloned into pENTR/D-TOPO (Invitrogen) and subsequently transferred into the GATEWAY-compatible destination vector pH7FWG2.0[47] with LR clonase. The resulting expression vector 35Sp:RKD2-GFP was used for floral dip transformation of *Arabidopsis*[48]. Seeds of transformed plants were surface-sterilized and grown on half-strength MS medium with 2% (w/v) sucrose, 1% phytoagar and hygromycin (30 µg ml^{-1}). RKD2-induced calli had formed after 20–30 days and were propagated in two-week subculture intervals on half-strength MS, 2% (w/v) sucrose, 1% phytoagar and hygromycin (30 µg ml^{-1}). Calli were collected with a sterile scalpel blade and immediately frozen in liquid nitrogen.

Leaf and root samples. *Arabidopsis* plants (Col-0) were grown under continuous light conditions at 22 °C. Senescent rosette leaves were collected from 35-day-old plants. Samples for stem, first node, second internode and first cauline leaf were collected from 30-day-old plants. Rosette leaf sections (Leaf7: distal; proximal; petiole) were collected from 22-day-old plants before bolting. Seedlings were grown on half-strength MS plates under continuous light for 7 days and separated into cotyledons, hypocotyl, root tip, root maturation zone or seed apical meristem including cotyledons and first leaves. Whole roots were obtained from 22-day-old plants grown under continuous light on half-strength MS plates.

Classification of growth stage and plant section was done as previously described[44,49–51]. Obtained material from at least three individual plants was combined for each sample, frozen in liquid nitrogen and stored at −80 °C until further use.

Protein lysis and digest

Frozen plant material was homogenized with a tissue lyzer (Qiagen) or with mortar and pestle in liquid nitrogen. Proteins were precipitated overnight with 10% tricholoroacetic acid in acetone at −20 °C and subsequently washed twice with ice-cold acetone. Dry samples were incubated with urea digestion buffer (8 M urea, 50 mM Tris-HCl pH 7.5, 1 mM DTT, cOmplete EDTA-free protease inhibitor cocktail (PIC) (Roche), phosphatase inhibitor (PI-III; in-house, composition resembling phosphatase inhibitor cocktail 1, 2 and 3 from Sigma-Aldrich)) for 1 h. Protein concentration was determined with a Bradford assay[52]. Then, 1 mg of protein was reduced (10 mM dithiothreitol (DTT), 1 h, room temperature), alkylated (55 mM chloroacetamide, 30 min, room temperature) and subsequently diluted 1:8 with digestion buffer (50 mM Tris-HCl pH 8.0, 1 mM $CaCl_2$). In-solution pre-digestion with trypsin (Roche) was performed for 4 h at 37 °C (1:100 protease:protein ratio), followed by overnight digestion with trypsin (1:100 protease:protein ratio). Samples were acidified to pH 3 using trifluoroacetic acid (TFA) and centrifuged at 14,000*g* for 15 min at 4 °C. The supernatant was desalted on 50 mg SepPAC SPE catridges (Waters). Peptides were eluted with 0.1% TFA in 50% acetonitrile (ACN) and vacuum-dried in a Thermo Savant SPD SpeedVac (Thermo Fisher Scientific).

Peptide enrichment and off-line fractionation

Fe^{3+}-IMAC was performed as described previously with some adjustments[53]. In brief, desalted peptide samples were re-suspended in ice-cold IMAC loading buffer (0.1% TFA, 40% ACN). For quality control, 1.5 nmol of a synthetic library of phosphopeptides and their corresponding non-phosphorylated counterpart sequence (B2 and F1)[54] were spiked into each sample before loading onto a Fe^{3+}-IMAC column (Propac IMAC-10 4 × 50 mm, Thermo Fisher Scientific). The enrichment was performed with buffer A (0.07% TFA, 30% ACN) as wash buffer and buffer B (0.315% NH_4OH) as elution buffer. Collected full proteome and phosphopeptide fractions were vacuum-dried and stored at −80 °C until further use.

For the full proteome fraction, hydrophilic strong anion-exchange chromatography peptide separation was performed as described previously[15]. In brief, an equivalent of 300 µg protein digest was reconstituted in hydrophilic strong anion-exchange solvent A (5 mM Tris-HCl, pH 8.5) and separated using a Dionex Ultimate 3000 HPLC system (Dionex) equipped with an IonPac AG24 guard column (2×50 mm) and an IonPac AS24 stong anion exchange column (2×250 mm, Thermo Fisher Scientific). Fractions were collected in 96-well format and subsequently pooled to 24. Individual fractions were acidified with formic acid, desalted on self-packed StageTips (five disks, diameter 1.5 mm C18 material, 3 M Empore, elution solvent 0.1% formic acid in 50% ACN) and dried down before liquid chromatography–tandem mass spectrometry (LC–MS/MS) analysis. Phosphopeptide fractions were separated into four fractions using a StageTip (five disks, diameter 1.5 mm C18 material, 3 M Empore) based high pH reversed-phase protocol as previously described[55]. Phophosphopeptides were eluted with 2.5%, 7.5%, 12.5% and 50% ACN in 25 mM NH_4FA, pH 10. The flow through and 50% ACN fraction were combined and all fractions were dried down before LC–MS/MS analysis.

For the CHX and MG132 chase experiments, 300 µg peptides for each sample were reconstituted in high pH reversed-phase loading buffer (2.5 mM NH_4HCO_3, pH 8) and fractionated using a Dionex Ultimate 3000 HPLC system (Dionex) and a Waters XBridge column (BEH130 C18, 3.5 µm, 2.1 × 150 mm; Waters). Peptides were separated by a linear gradient from 4% buffer D to 32% D in 45 min, followed by a linear gradient from 32% D to 80% D in 6 min. The proportion of buffer A was kept at

10% during fractionation (buffer A: 25 mM NH_4HCO_3, pH 8; buffer C: 100% ACN; buffer D: 100% ultrapure water). Fractions were collected in 96-well format, subsequently pooled to 48, acidified and dried in a SpeedVac.

Seed treatment with CHX and MG132

Aliquots of surface-sterilized mature dry seeds (10 mg, Col-0) were stored for two days at 4 °C in the dark and subsequently imbibed with 2 ml liquid half-strength MS at 22 °C under constant light. After 4 h incubation, CHX or MG132 (*N*-(benzyloxycarbonyl)-Leu-Leu-Leu-al) was added to a concentration of 100 µM. Because both CHX and MG132 were dissolved at the same stock concentration in 100% DMSO, the respective volume of 100% DMSO was added to the control samples. As a baseline sample, one seed aliquot was dried and immediately frozen in liquid nitrogen after the 4 h incubation in half-strength MS medium. Seeds for CHX, MG132 and DMSO control were removed after 8 h, 16 h and 24 h treatment, dried and frozen in liquid nitrogen. Germination of treated seeds was visually checked after four days of incubation. Both compounds CHX and MG132 were active and led to complete or partial inhibition of seed germination as compared to a DMSO treated control sample (Extended Data Fig. 4i).

Protein extraction and digest was performed as described above. Peptide quantification before high pH reversed phase fractionation was done using the Pierce BCA protein assay kit (Thermo Fisher Scientific)[56].

SEC

Homogenized flower (stage 15), leaf (rosette leaf 7) and root (whole root) samples were lysed in ice-cold 50 mM Tris-HCl pH 7.4, 100 mM NaCl, 10% glycerin, PIC, PI-III and 0.1% Triton-X100. After filtering (diameter 0.2 µm), 0.25 ml lysate containing 1 mg of protein was injected on a Superose 6 10/30 GL column (GE Healthcare) and separated at a flow rate of 250 µl min^{-1} on an Äkta pure 25 (GE Healthcare). Molecular mass calibration of the column was performed with the high molecular mass gel filtration calibration kit (GE Healthcare). After the void volume, 0 fractions of 125 µl each were collected, vacuum-dried and re-solubilized in urea digestion buffer. In-solution digestion with trypsin and sample desalting on self-packed StageTips was performed as described above. For quality control, PROCAL peptide standard[57] was spiked into each SEC fraction before LC–MS/MS analysis.

LC–MS/MS analysis

Nanoflow LC-MS/MS was performed by coupling a Dionex 3000 (Thermo Fisher Scientific) to a QExactive Orbitrap HF (Thermo Fisher Scientific). Samples for the proteome or phosphoproteome analysis were re-suspended in loading buffer containing 0.1% formic acid or 50 mM citrate and 1% formic acid, respectively. Peptide loading and washing were done on a trap column (100 µm ID × 2 cm, packed in-house with Reprosil-Pur C18-GOLD, 5 µm resin, Dr. Maisch) at a flow rate of 5 µl min^{-1} in 100% loading buffer (0.1% formic acid) for 10 min. Peptide separation was performed on an analytical column (75 µm ID × 40 cm packed in-house with Reprosil-Pur C18, 3 µm resin, Dr. Maisch) at a flow rate of 300 nl min^{-1} using a 110 min gradient from 4% to 32% solvent B (solvent A: 0.1% formic acid, 5% DMSO in HPLC grade water; solvent B: 0.1% formic acid, 5% DMSO in acetonitrile) for the tissue map full proteome analysis and a two-step 110 min gradient from 2% to 27% solvent B for the phosphoproteome analysis[58]. Peptides were ionized using a spray voltage of 2.2 kV and a capillary temperature of 275 °C. The instrument was operated in data-dependent mode, automatically switching between MS and MS2 scans. For the full proteome samples of the tissue atlas experiment, full scan MS spectra (*m*/*z* 360–1,300) were acquired with a maximum injection time of 10 ms at 60,000 resolution and an automatic gain control (AGC) target value of 3×10^6 charges. For the top 12 precursor ions, high-resolution MS2 spectra were acquired in the orbitrap with a maximum injection time of 50 ms at 15,000 resolution (isolation window 1.7 *m*/*z*), an AGC target value of 2×10^5 and normalized collision energy of 25%. The underfill ratio was set to 1% with a dynamic exclusion of 30 s. Only precursors with charge states between 2 and 6 were selected for fragmentation. For the phosphoproteome analysis, the MS2 spectra were acquired with a resolution of 30,000 (isolation window 1.7 *m*/*z*) and a maximum injection time of 120 ms. Dynamic exclusion was set to 35 s.

SEC samples were measured using a 60 min 4% to 32% gradient. MS2 spectra were generated for the top 20 precursors, with a maximum injection time of 50 ms at 30,000 resolution (isolation window 1.7 *m*/*z*), a normalized collision energy of 25% and a dynamic exclusion of 20 s.

The seed samples treated with CHX or MG132 were analysed by a micro-flow LC system coupled online to an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific) as previously described[59]. The micro-flow LC system was built by combining a modified Vanquish pump with the auto-sampler of the Dionex UltiMate 3000 RSLCnano System. The sample was directly loaded onto a Thermo Fisher Scientific Acclaim PepMap 100 C18 LC column (2 µm particle size, 1 mm ID × 150 mm), the flow rate was 50 µl min^{-1} and column temperature was maintained at 55 °C. A 15-min linear gradient of 7–32% solvent B (solvent A: 0.1% formic acid, 3% DMSO in HPLC grade water; solvent B: 0.1% formic acid, 3% DMSO in ACN) was used to separate all the samples, followed by 1 min 95% solvent B washing and 1 min 0.5% solvent B equilibrium time at a flow rate of 100 µl min^{-1}. The Orbitrap Fusion Lumos mass spectrometer was operated with the Ion Max API Source installed with a HESI-II probe (50 µm ID). The detailed acquisition parameters are: Positive polarity; spray voltage 3.5 kV, funnel RF lens value at 40, capillary temperature of 325 °C, auxiliary gas heater temperature of 300 °C. The flow rates for sheath gas, aux gas and sweep gas were set to 32, 5 and 0, respectively. Full scan MS spectra (*m*/*z* 360–1,300) were acquired in the orbitrap with a maximum injection time of 10 ms at 120,000 and an AGC target of 4×10^5. MS2 spectra were acquired in the linear ion trap (rapid scan mode) after collision-induced dissociation fragmentation with collision energy set at 35%, an AGC target of 1×10^4 and maximum IT of 10 ms. The isolation window was set to 0.4 *m*/*z* and scans were recorded with a maximum duty cycle of 0.6 s and the option 'inject ions for all parallelizable time' enabled. Only precursors with charge states between 2 and 6 were selected for fragmentation and the lowest scan range of MS2 was fixed at 100 *m*/*z*. The intensity threshold was set to 5×10^3 and the dynamic exclusion to 12 s.

Peptide and protein identification and quantification

Raw data files were searched with MaxQuant software (v.1.5.8.3) using standard settings unless otherwise described[60]. MS/MS spectra were searched against Araport11[5] protein coding genes (Araport11_genes.201606.pep.fasta; downloaded June 2016), with trypsin as protease and up to two allowed missed cleavages. Carbamidomethylation of cysteines was set as fixed modification and oxidation of methionines and N-terminal acetylation as variable modifications. The 'match-between-runs' function was enabled for corresponding fractions within one parameter set, where applicable.

For the tissue map analysis, full proteome and phosphoproteome samples were processed together as two separate parameter groups, with phosphorylation of serine, threonine or tyrosine defined as variable modification only for the phosphoproteome group. Here we also added the spike-in phosphopeptide library sequences[54] to the database search space.

For sORF identification, two sORF databases (ATSO and ARA-PEP) were added to the search space in addition to Araport11. ARA-PEP is a previously described repository of putative sORF-encoded peptides in *Arabidopsis*[14]. ATSO (*A. thaliana* sORFs) was generated by using sORFfinder[61] and our RNA-seq analysis data to identify putative new sORFs in non-coding intergenic regions (ATSO.non_coding.pep.nr.cd-hit_aS1_aL0.3_c1.out.nr.fasta). Three targets were used as input for sORFfinder. Sequences of intergenic and non-coding regions relative to the Araport11 annotation, and sequences resulting from a Trinity[62] de novo transcriptome assembly that are non-overlapping with the Araport11 annotations. Afterwards, the sORF nucleotide sequences were translated to amino acid sequences. CD-HIT[63] was used with the parameters –aS 1 –aL 0.3 –c 1 to reduce sequence redundancy.

SEC experiment raw files for flower, leaf and root were all searched together, with each SEC fraction designated as individual experiment and the 'match-between-runs' function enabled.

The seed experiment raw files were searched using standard settings as described above.

Peak list files generated previously[7] were downloaded from Pride (PRD000044) and searched with the Mascot[64] search engine (v.2.4.1, Matrix science) against the Araport11 database. The target-decoy option of Mascot was enabled and search parameters included a precursor tolerance of 1.3 Da and a fragment ion tolerance 0.45 Da. Enzyme specificity was set to trypsin and up to two missed cleavages were allowed. The Mascot ^{13}C option, which accounts for the misassignment of the monoisotopic precursor peak, was set to 1 and oxidation of methionine and carbamidomethylation of cysteines were included as variable and fixed modifications, respectively. The isobarQuant workflow was used to generate protein identification files[65].

Spectra validation

sORF peptides identified by database searching that could not be mapped to an existing Araport11 gene model were synthethized at JPT Peptide Technologies using Fmoc-based SPOT synthesis on membranes[57] and measured on the same mass spectrometer that was used for data acquisition (tissue atlas full proteome method; see 'LC–MS/MS analysis' section). Experimental and synthetic peptide spectra were extracted from the raw files and used for similarity calculation without prior processing. Normalized spectral contrast angle comparison between spectra of the tissue sample and synthetic peptides was performed using in-house R scripts[66] factoring in peaks which are either shared between spectra or exclusive to the synthetic peptide spectra. Peaks exclusive to experimental spectra were ignored. Candidates were selected with spectral contrast angle cut-off values larger than 0.7 and BLASTed against the UniProt database for further validation (Supplementary Data 3).

Protein identifications for the uncertain evidence category of UniProt were validated by comparing the experimental spectra to in-silico predicted fragment spectra[12] (Supplementary Table 1). For this, a spectral library was obtained from Prosit (https://www.proteomicsDB.org/prosit)[12] by uploading all sequence-charge combinations of peptides that were identified for uncertain proteins. To obtain the best matching spectra, the collision energy of Prosit was calibrated using a standard quality control run from the same mass spectrometer that was used for data acquisition. The resulting predicted spectral library was visualized and compared to the experimentally acquired data analogue to the sORF candidate peptides. Spectral contrast angle values larger than 0.7 were used as cut-off values.

RNA sequencing

Total RNA was isolated with the NucleoSpin RNA Plant kit (Macherey-Nagel) according to the manufacturer's instructions. DNA was removed by on-column treatment with rDNase (Macherey-Nagel). For recalcitrant samples (seed, silique, pollen), a LiCl-based protocol was adopted with minor modifications[67]. After LiCl precipitation, the RNA pellet was dissolved in rDNase buffer and treated with rDNase (Macherey-Nagel) at 37 °C for 10 min. The final pellet was re-suspended in 35 µl DEPC-treated water.

RNA was quantified (Nanodrop, Thermo Fisher Scientific) and quality checked with a Bioanalyzer 2100 (Agilent Technologies). RNA integrity number values between 6.4 and 10 were accepted for further analysis. cDNA libraries were prepared using the TruSeq Stranded mRNA Sample Preparation kit (Illumina) according to the instructions. Clusters were generated in two batches and sequenced on a high-throughput flow cell with the HiSeq 2500 platform (Illumina) to a depth of 36 million reads per sample. Quality assessment of raw and trimmed 75-bp paired RNA-seq reads was performed with FastQC. Raw RNA-seq reads were trimmed to remove adaptor contamination and poor-quality base calls using Trimmomatic v.0.35 with parameters (ILLUMINACLIP: Illumina-PE.fasta:2:30:10; LEADING:3; TRAILING:3; SLIDINGWINDOW:4:20; MINLEN:36)[68]. Trimmed RNA-seq reads were mapped to the Araport11[4] transcriptome with Kallisto v.0.43.1 (default parameters)[69].

Data processing

MaxQuant output tables were filtered for non-plant contaminants, reversed sequences and proteins that were only identified based on modified peptides. Protein abundance estimation was based on either intensity-based absolute quantification (iBAQ)[26], or top-3 quantification[70], depending on the analysis. MaxQuant ProteinGroups containing several gene loci were filtered out to retain only unambiguously identified gene loci for further analyses. In case multiple protein isoforms were identification in distinct ProteinGroups, only the isoform with the higher number of razor + unique peptides was retained.

mRNA quantities are displayed as TPM and a cut-off value of 1 TPM was used as lower limit of detection across all samples. GO[71] term analysis for the transcript abundance range was performed using the 1D enrichment function from Perseus[72,73] with Benjamini–Hochberg false discovery rate (FDR) threshold set to 0.01. In this function, a two-sided Wilcoxon–Mann–Whitney test is used to test for systematically larger or smaller transcript abundance within a GO category as compared to the global distribution of all values (Supplementary Data 4).

For further qualitative and quantitative analyses, all transcript or protein isoform information was collapsed onto gene level. Unless otherwise stated, displayed abundances for protein, transcript and P-site were $\log_2$-transformed. Protein, peptide and transcript datasets for the tissue map and seed treatment experiments were median-centred to the overall median of the respective dataset. No normalization was performed for the P-site dataset, because variations in the intensities of total P-sites between tissues are also due to biological sample differences. Instead, the spike-in phosphopeptide library was used, to assess reproducible enrichment efficiency and MS measurement quality of phosphoproteome samples[54].

P-sites that were reported in the MaxQuant output table (Phospho(STY)sites.txt file) with a 'localization probability' larger than 0.75 were designated as 'high confidence localizations' or class I sites[74]. P-sites were considered exclusive if they were only detected in a single tissue. The number and identity of P-sites and phosphoproteins detected in this study were compared to datasets available through PhosPhAT4.0 and a published meta-analysis[8,9] (Extended Data Fig. 1h).

Data analysis

Genome annotations. Chromosome information contained in the *Arabidopsis* locus identifiers (AGI codes: AT (*A. thaliana*); 1, 2, 3, 4, 5, M, C (chromosome number, M for mitochondrial, C for chloroplast); G (gene); 12300 (five-digit code for position on chromosome)), was used to assign genes to their respective chromosomes (Extended Data Fig. 1c). Araport11 gene identifiers (*Arabidopsis* gene identifiers (AGI)) were mapped to the UniProt *A. thaliana* reference proteome (taxon identifier 3702; UP000006548; downloaded November 2018) based on protein sequence. Swiss-Prot, TremBL as well as protein existence criteria annotations (level 1: evidence on protein level; level 2: evidence on transcript level; level 3: inferred from homology, level 4: predicted, level 5: uncertain) were subsequently obtained from UniProt (Extended Data Fig. 1d).

N- and C-terminal peptide sequences were extracted from the MaxQuant peptides.txt file and filtered for zero missed cleavages (n = 2,776) (Extended Data Fig. 2a). N-terminal peptides were divided into groups with (n = 1,707) or without (n = 1,069) cleavage of the initiator methionine. The frequency of the 20 genetically encoded amino acids at the position after the start codon was subsequently calculated separately for both groups and displayed as a pie chart. For N-terminal peptides with the same amino acids at the position after the start codon, the percentage of peptides with N-terminal acetylation was also calculated and displayed as bar plot. This analysis was also performed separately for the groups with or without cleavage of the initiator methionine (Extended Data Fig. 2b, c).

Gene isoforms annotated in Araport11 were considered as 'identified', if they were detected with a TPM intensity larger than our limit of detection cut-off value (1 TPM; transcript level) or with an isoform-specific peptide (protein level) in at least one tissue sample (Extended Data Fig. 2f, Supplementary Data 3). A selection of isoform-specific peptides was synthesized together with the sORF peptide collection at JPT Peptide Technologies and the synthetic peptides were used for validation of the experimental spectra identifications as described above.

Transcription factor, transcription regulator, kinase and phosphatase family annotations and classifications are based on previous reports[16,17]. Proportional coverage of these families within our dataset was calculated by counting how many of them could be identified on transcript, protein or phosphoprotein level, respectively (Extended Data Fig. 2k).

Tissue groups and tissue characteristics. Tissue samples were assigned to tissue groups as follows, based on their origin or common morphology: flower (sepal, petal, stamen, carpel, silique, flower); fruit (silique septum, silique valves); seed (embryo, seed, seed imbibed); pollen (pollen); leaf (cauline leaf, leaf distal, leaf proximal, leaf petiole, senescent leaf, cotyledons, shoot tip); stem (node, internode, flower pedicle, hypocotyl); root (root, root tip, root upper zone); callus (egg-cell like callus, callus); and cell culture (cell culture early, cell culture late) (Fig. 1a, b).

To provide a measure of tissue similarity, Pearson's correlation coefficient (*r*) was calculated for the gene expression of all pairwise tissue combinations. Pearson's correlation coefficient is a measure of the linear correlation between two variables, in this case the expression levels in two tissue samples. Correlations were computed both on protein level and transcript level and displayed as separate heat maps for both omics levels (Extended Data Fig. 1a). For three examples of morphologically highly similar tissue pairs, the gene expression levels were also displayed as scatter plots using protein abundance or transcript abundance measurements, respectively (Extended Data Fig. 1b).

Tissue or tissue-group specificity of genes was calculated based on iBAQ or TPM abundance values, respectively[75]. Genes were assigned to categories in the following order: 'tissue-specific', 'tissue-enhanced', 'group-specific', 'group-enhanced', 'shared' and 'mixed'. Genes were considered 'specific', if they were only detected in a particular tissue or tissue-group and 'enhanced' if their abundance was at least fivefold higher in a particular tissue or tissue-group as compared to the average levels in all other tissues. Genes that were detected in all 30 tissues but did not show enhanced expression in a tissue or tissue group were classified as shared. All remaining genes are contained in the mixed category (Extended Data Fig. 3a).

Genes detected in all tissues on either protein or transcript level were assigned to the core datasets (core_transcript $n = 8{,}405$; core_protein $n = 7{,}734$; core_intersection $n = 5{,}043$). This classification does not consider gene expression levels and should not be confused with the tissue-specificity classification described in the previous paragraph.

Hierarchical clustering analysis for tissues representing flower parts or the whole flower was performed on $\log_2$-transformed, *z*-scored protein intensities (iBAQ values) in Perseus using Euclidean distance and average linkage (Extended Data Fig. 3b). The flower organ identity model was restricted to the simplified ABC model, thus only displaying homeotic genes for the A, B and C classes[76], namely the expression of MADS-box transcription factors *APETALA1* (*AP1*) for class A, *APETALA3* (*AP3*) and *PISTILLATA* (*PI*) for class B, and *AGAMOUS* (*AG*) for class C. Class A expression specifies sepal formation, the combination of class A and B specifies petal formation, the combination of class B and C specifies stamen formation, and class C expression specifies carpel formation in developing flowers.

For the cumulative abundance calculation of five representative tissues (flower; pollen; root tip; leaf proximal; seed), proteins (iBAQ) and transcripts (TPM) were first ranked from highest to lowest individual intensity (Extended Data Fig. 3e). The running total was then plotted against the rank order. The names or identifiers of the five most abundant transcripts or proteins (rank 1 to 5) are listed in descending order for the respective tissue. Shared IDs among the top 100 most abundant transcripts (TPM) and proteins (iBAQ) were calculated for each individual tissue (Extended Data Fig. 3f). For the genes representing the highest abundant protein in at least one tissue, the contribution to the total protein amount in an individual tissue was calculated by dividing the iBAQ intensity (not log-transformed) of the respective protein by the summed total iBAQ intensity (not log-transformed) in this tissue (Extended Data Fig. 3g).

Principal component analysis of proteome and transcriptome data was performed in Perseus for the intersection of the core datasets ($n = 5{,}043$) on $\log_2$-transformed and *z*-scored iBAQ and TPM intensities (Extended Data Fig. 3h). Subcellular localization information was downloaded from SUBA (downloaded November 2016; unambiguous localizations $n = 3{,}506$)[77] (Supplementary Data 2) and used to calculate the intensity contribution of proteins assigned to a specific compartment for the different tissue groups. For this, the percentage of proteins with the same SUBA annotation was calculated by dividing the summed protein intensity (iBAQ, not log-transformed) of each SUBA category by the total summed iBAQ intensity (not log-transformed) within each tissue. To compare tissue groups, the respective protein intensity proportion was then averaged for all tissues within one group. These averaged proportions were subsequently plotted for each subcellular compartment (Extended Data Fig. 3i).

Protein/mRNA relation. The Pearson's correlation value (*r*) between median protein (median iBAQ of 30 tissues) and median transcript (median TPM of 30 tissues) abundances was calculated using the set of genes with abundance measurements on both protein and transcript level in more than ten tissues (more than 10 pairwise complete observations; $n = 14{,}069$). The scatterplot of median transcript (TPM) and protein abundance (iBAQ) was displayed together with their marginal histograms (Fig. 2a). Tissue-specific Pearson's correlation values between protein and transcript abundance were calculated using all genes from the core proteome and core transcriptome intersection dataset ($n = 5{,}043$) (Extended Data Fig. 4a). The core dataset was used here to base the Pearson's correlation value comparison between tissues on the same set of genes in all tissues.

The data subset with more than 10 pairwise complete observations on protein and transcript level ($n = 14{,}069$) was also used for all further calculations involving PTR (Fig. 3d, Extended Data Fig. 5). PTR values are calculated by building the ratio between protein abundance (iBAQ) and the corresponding transcript abundance (TPM) for each gene and were calculated separately for each tissue. The tissue PTR values were then used to calculate the median and median absolute deviation (MAD) of PTR values across all 30 tissues (Extended Data Fig. 5). The variation of PTR values across tissues indicates, whether protein and transcript levels of a given gene are regulated in a similar (small PTR variation) or different way (high PTR variation) between individual tissues. To estimate the proportion of genes with rather stable or variable PTR values, the PTR MAD values were distributed into five equal parts, so that 20% of all genes from this analysis are contained in one part (5 MAD quantiles: Q1–Q5). For each gene, the median PTR values were plotted against their MAD across tissues and the MAD quantiles indicated in the plot (Extended Data Fig. 5a). GO term analysis for the median PTR distribution (median PTR across all tissues) was performed using the 1D enrichment function from Perseus[72] with Benjamini–Hochberg FDR threshold set to 0.01 (Supplementary

Data 4). In this function, a two-sided Wilcoxon–Mann–Whitney test is used to test for systematically larger or smaller PTR values within a GO category as compared to the global distribution of all values.

In a similar way, the variation of phospho ratios across tissues indicates, whether P-site levels are driven by protein abundance (small phospho ratio variation) or P-site stoichiometry (high phospho ratio variation) changes, respectively. Phospho ratios were calculated by building the ratio between the abundance of a P-site (intensity) and the abundance of its corresponding protein (iBAQ). For this, we used the set of P-sites, with abundance measurements on both P-site and protein level in more than 10 tissues (more than 10 pairwise complete observations; $n = 13{,}793$). Phospho ratios were calculated separately for each tissue and subsequently combined to calculate the median phospho ratio and phospho ratio MAD across all 30 tissues. To estimate the proportion of P-sites that closely resemble the protein abundance profile, the phospho ratio MAD values were again distributed into five equal parts (five MAD quantiles: Q1–Q5), and the median phospho ratio values plotted against their MAD across tissues each P-site. MAD quantiles were indicated in the plot (Extended Data Fig. 5c).

To provide further information about inter-tissue variability, MAD quantiles were also calculated and displayed for the expression levels on protein (iBAQ; $n = 14{,}069$), transcript (TPM; $n = 14{,}069$) and P-site (peptide intensity; $n = 13{,}793$) levels, respectively (Extended Data Fig. 5b, d).

Features for protein level prediction models. As a substantial proportion of variation in protein levels remained unexplained when using transcript level information alone (Pearson's correlation between protein and transcript abundance in tissues $r = 0.28$–0.79), additional molecular features, that could explain protein abundance variations, were tested for their influence on protein abundance in a model selection approach. Predictors selected for this analysis were: mRNA levels, codon usage, non-synonymous-to-synonymous substitution (d_N/d_S) ratios, which are a measure of evolutionary conservation, gene/coding sequence (CDS) length, exon number, gene position on the chromosome, cytosine methylation, the number of putative protein interactions and mRNA sequence motifs (*k*-mers of size 3 to 7 nucleotides):

Codon usage statistics for the *A. thaliana* genome were obtained from Kazusa (www.kazusa.or.jp/codon/current/species/3702) and parsed to extract NCBI gene identifiers. These identifiers were mapped to their corresponding UniProt entries and Ensembl/TAIR10 annotations using the UniProt Retrieve/ID mapping tool. The extracted TAIR10 annotation was merged with the Kazusa codon usage dataset. Codon frequencies were calculated for each gene by dividing the count (×3) of a given codon by the full length of the coding sequence.

The d_N/d_S substitution rates were calculated from CDS pairs between *A. thaliana* and its closest relative *Arabidopsis lyrata*. Reciprocal best BLAST with an *E*-value cut-off of $\leq 1 \times 10^{-8}$ was used to identify orthologue sequences. Individual CDS pairs were aligned using PRANK[78] and Gblocks[79] was applied to eliminate poorly aligned positions in an alignment with a cut-off value of eight contiguous non-conserved positions and without allowing for gap positions. The yn00 package in the program PAML[80] for pairwise sequence comparison was used to estimate substitution rates, d_N and d_S, respectively.

The total number of exons and the total gene lengths were obtained from Araport11. The distance of each gene from the chromosomal centromeres was calculated to capture potential position-specific effects.

A. thaliana (Col-0) whole-genome bisulfite sequencing data were obtained from a previous publication[81]. For each gene, methylation levels were calculated for contexts CG, CHG and CHG (in which H denotes adenine, cytosine or thymine) separately. Per gene methylation levels were defined as:

$$g_i = \frac{1}{\max(j)} x \sum \frac{Nm_j}{N_j}$$

where max(*j*) is the total number of cytosines, Nm_j is the number of methylated reads, and N_j the total number of reads at the *j*th cytosine.

Arabidopsis protein–protein interactions were downloaded from STRING[30] and the number of protein interaction partners extracted for each gene.

All of these features have previously been associated with transcriptional activity and/or protein abundance levels in *Arabidopsis* and/or other species[82–87]. For a detailed feature set description, see Supplementary Data 5.

De novo motif identification. Motifs in the mRNA CDS, 3′ or 5′UTR region were identified as previously described[23]. In brief, protein abundance levels were $\log_{10}$-transformed and median centred. For genes with two or more transcript isoforms, the transcript isoform reported to have the largest summed iBAQ values across all tissues was defined as the major transcript isoform per gene and used to compute all sequence feature and mRNA levels. Raw RNA-seq reads were trimmed to remove adaptor contaminations and poor-quality base calls using Trimmomatic 0.39 with parameters (ILLUMINACLIP:Illumina-PE.fasta:2:30:10; LEADING:3; TRAILING:3; SLIDINGWINDOW:4:20; MINLEN:36). Trimmed RNA-seq reads were mapped to the Araport11 transcriptome annotation with Kallisto v.0.46.0 using default parameters. To estimate the levels of mature mRNA, the number of reads mapping to exonic and intronic regions of the transcript were counted separately and then normalized by the total exonic and intronic region lengths, respectively. Normalized intronic counts were subtracted from the normalized exonic counts to obtain the mature mRNA counts. The resulting normalized exonic counts per sample were corrected by the DESeq2[88] library size factor and $\log_{10}$-transformed. Transcripts with 10 reads per 1 kb were treated as transcribed. Tissue-specific PTRs were computed using the normalized protein and transcript levels. GEMMA software[89] was used to identify de novo motifs in 5′ UTR, CDS and 3′ UTR regions using the tissue-specific PTRs as response variables. GEMMA uses a linear mixed model, in which the effect of each individual *k*-mer on the median PTR across tissues is assessed while controlling for the effect of other *k*-mers (random effects), region length and GC percentage (fixed effects). The motif search was performed for *k*-mers ranging from 3 to 7 nucleotides. Obtained *P* values were adjusted for multiple testing with Benjamini–Hochberg's FDR and jointly computed across the *P* values of all tissues. Gemma was run using the median PTR with FDR < 0.1 and covariates set to 'false'. In total, 82 significant putative motifs were obtained based on their sequence (5′ UTR $n = 32$; 3′ UTR $n = 38$; CDS $n = 12$) and sub-sequence (initial, all, end) region. Then, 39 motifs that lie in the 'initial' and 'end' sub-sequences were further selected. The presence or absence of each enriched motif with respect to each gene was extracted in form of a binary matrix and used for downstream multivariate feature selection analysis.

Model-based feature selection. Tissue-specific protein abundance data were merged to the feature matrix. Preliminary pairwise correlation analysis showed only weak to moderate correlation between individual features. In addition, variance inflation factors were calculated for each feature using the fmsb R package[90]. The result showed low variance inflation factors, suggesting that multicollinearity was not an issue for downstream analyses.

Two model selection approaches were implemented—stepwise regression and Lasso regression[91,92]. To select the most predictive features for protein abundance in each tissue, we used a forward–backward model selection approach in a multiple regression framework. The method was implemented using stepwiseAIC() function in R[93], which compares the fit of nested models. To ensure that the comparison of model AICs was not affected by unequal sample size, missing data were removed before the analysis.

Because stepwise regression can occasionally lead to over-fitting[94], a Lasso regression was implemented as an alternative model selection procedure. Lasso regression performs L1 regularization, which adds a penalty equivalent to the absolute of the magnitude of regression coefficients and tries to minimize them. The strength of the penalty was controlled via the tuning parameter λ[92]. Lasso was implemented using the glmnet package[95] in R. Model training was performed on a random selection of genes (50% of the dataset) and implemented using the cv.glmnet() function. The optimal value for λ was extracted and used to re-build the model using the glmnet() function. Finally, the fitted model was used to obtain predictions in the remaining 50% of the genes.

For each tissue, stepwise and Lasso regression models were compared. Stepwise regression yielded only slightly higher coefficients of determination (R^2) values compared to Lasso, suggesting over-fitting was not an issue (Supplementary Data 5). As stepwise regression yielded more parsimonious models in general, we used this approach for further analysis. Selected features here could explain on average 52% of protein abundance variation (37% in pollen, 62% in cell culture early) (Supplementary Data 5). For each tissue, features from the best fitting models were summarized in an incidence matrix along with the effect direction (positive or negative effect on protein levels). To determine the importance of each feature to the overall model fits, R^2 variance decomposition was performed using the 'genici' metric which is implemented in the relaimpo R package[96] (Extended Data Fig. 4c). Relative feature contributions were averaged across all tissues (Fig. 3b).

Because many of the detected motifs appear in a tissue-specific manner (predictive only in tissue subsets), we clustered tissues according to the presence or absence status of 5′ UTR motifs in each tissue model (Extended Data Fig. 4d). Seed and pollen tissues form a distinct subcluster, which might be connected to the increased/decreased PTR levels in these tissues compared to the other tissues (Fig. 3c, Extended Data Fig. 4a). In contrast to the motif analysis, d_N/d_S ratios were consistently (and positively) associated with protein abundance in all 30 tissues (Supplementary Data 5). To explore the relationship between d_N/d_S ratios and protein abundance in more detail, groups of genes with low or high (bottom 5% or top 5% of d_N/d_S distribution) were compared to each other with regard to protein levels (Extended Data Fig. 4e).

Seed tissue PTR regulation. The PTR value distribution was plotted for the median PTR (median across all tissues, see 'Protein/mRNA relation') and selected tissues (seed, seed imbibed, pollen, cell culture early). Seed and pollen tissues show particularly low Pearson's correlation values between protein and transcript abundance levels (Extended Data Fig. 4a). Cell culture early in comparison shows the highest Pearson's correlation among the analysed tissues. Tissue-specific 'outliers' with especially high or low PTR as compared to all other tissues were selected if their PTR value was outside two standard deviations of the mean of the median PTR distribution. For seed, this resulted in 469 genes with 'high' and 571 genes with 'low' PTR. To interpret whether this change in seed PTR values was caused by differential regulation on transcript and/or protein level, the fold change between seed protein and transcript abundance and the median protein and transcript abundance of all tissues was calculated. Genes with at a least fourfold change in either protein or transcript abundance were considered regulated (less/more transcript and/or less/more protein in seed compared to all tissues). The percentage of genes with high or low PTR levels in the seed tissue, which are regulated in the same manner (less/more transcript and/or protein) was calculated and displayed as an arrow plot (Fig. 3e). About one-quarter of each gene set (high PTR 27%; low PTR 24%) showed no regulation based on our fourfold-change cut-off.

For the low PTR gene set, a high proportion of genes with lower protein abundance in the seed tissue was observed. A protein-level chase experiment was performed, to investigate whether this reduction was due to translational inhibition or increased protein degradation in seeds. Protein level (iBAQ, not log-transformed) fold changes in CHX-, MG132- and DMSO-treated samples were calculated relative to the baseline protein abundance at 0 h (control sample). The average protein-level fold change for the subset of genes with either high or low PTR in seeds (see previous paragraph) was subsequently determined for each time point and plotted separately for 'high' and 'low' PTR genes (Fig. 3f). To test for significant fold change differences between treatment time points, an analysis of variance (ANOVA) and post hoc Tukey honestly significant difference (HSD) test was performed using the stats package in R (Extended Data Fig. 4f). For this, the data were normalized for changes in the proteome caused by seed germination, by calculating the $\log_2$-transformed fold change in protein abundance (iBAQ) between seed samples treated with CHX or MG132 and the time-point-matched DMSO control. Outlier values were removed for the box plot visualization but are included in the ANOVA and post hoc Tukey analysis.

Co-expression analysis. To compare gene co-expression information in our dataset to prior knowledge about various associations between genes pairs, the Pearson's correlations between all pairwise gene combinations on both transcript and protein level (core dataset intersection; $n = 5{,}043$) was calculated. The correlation coefficient value for a given gene pair on transcript level was then plotted against the correlation coefficient value of the same pair on protein level (Extended Data Fig. 6a). Protein–protein association data were downloaded from STRING[30] (www.string-db.org; 3702.protein.links.detailed.v10.5.txt.gz; download July 2018), the experimental STRING subscores calculated for each gene pair and the information merged to the correlation value matrix. The scatter plot of transcript level versus protein level correlations was then divided into 50 × 50 bins along the x and y axis and the mean experimental STRING subscore for all gene pairs within a bin was calculated. Mean subscores were $\log_{10}$-transformed and visualized as a heat map together with their marginal histograms (Extended Data Fig. 6a).

For a subset of genes, pairwise co-expression was further investigated. Pairs used in this analysis were genes that are either annotated as duplicated[31] or as protein interactors (AtPIN[33], downloaded August 2017), respectively. The Pearson's correlation between the expression levels of gene pairs across tissues was calculated for gene pairs with abundance measurements in more than 10 matching tissues on both protein and transcript level (more than 10 pairwise complete observations). The duplicated gene set was further filtered for three unique peptide identification of each paralogue. Top-3 intensities were used as protein abundance measure for the duplicate gene set as the subsequent ratio comparison was also performed with top-3 intensities. Paralogue genes often have high sequence identity on the amino acid level, which can lead to distorted ratios owing to uneven assignment of shared (razor) peptides to only one of the paralogues. Duplicated genes ($n = 3{,}612$) were divided into three subsets: duplications caused by a whole-genome duplication event ($n = 2{,}104$), local duplication (local, $n = 408$) or transposon-mediated duplication (transposed, $n = 1{,}100$). For each subset, the distribution of Pearson's correlation values between duplicates on either transcript or protein levels was plotted (Extended Data Fig. 6b). A set of random gene pairs ($n = 27{,}353$) was generated for comparison, filtered like the duplicated gene set (number of pairwise complete observation, three unique peptides) and the Pearson's correlation value distribution displayed in the same way. To compare relative protein abundance between paralogues, the ratio between the protein abundance (top-3) of one paralogue and the protein abundance (top-3) of the other paralogue was calculated (Extended Data Fig. 6c). As an example for the intensity proportion of paralogue genes in different tissues, the top-3 intensity (not log-transformed) of MAC5A and MAC5B was plotted as proportion of the summed top-3 intensity (not log-transformed) of both paralogues (Extended Data Fig. 6c). For 57 paralogue gene pairs with phenotypic information about asymmetric loss-of-function mutant combinations[32], we build the average top-3 intensity proportion of the first paralogue (dupl1/(dupl1 + dupl2)) across 30 tissues and plotted them in descending order together with the phenotype information (Extended Data Fig. 6d).

A set of well-studied, stable protein complexes (26S proteasome, COP9 signalosome, chaperonin, cellulose synthase) was selected to estimate the protein-level (iBAQ) co-expression expected for stable interaction partners. For this, the Pearson's correlation coefficient (*r*) values between all pairwise subunit combinations of an individual complex were calculated and displayed in a density plot (Extended Data Fig. 6e). On the basis of this analysis, a Pearson correlation coefficient cut-off value of >0.5 was subsequently used as indication for stable protein interactions. The protein interactor gene set (AtPIN) was used to test the recovery of annotated protein–protein interactions based on co-expression data from this study. Again, the distribution of Pearson's correlation values (*r*) between gene pairs (here interactors) on either transcript or protein level (iBAQ) was plotted. This was done for the whole AtPIN dataset ($n = 57{,}152$) and the following subsets: interactions identified by yeast-two hybrid experiments ($n = 7{,}621$), by affinity-purification mass spectrometry ($n = 17{,}982$) or by both yeast two-hybrid and affinity-purification mass spectrometry ($n = 829$).

SEC and complex analysis. Reproducibility between the three SEC experiments was tested by comparing the elution behaviour of the same proteins. The coefficient of determination between protein peaks in the different experiments was above 0.7 (flower–root $R^2 = 0.82$; flower–leaf $R^2 = 0.78$; root–leaf $R^2 = 0.7$). Protein peak elution between the samples was also consistent across the whole SEC gradient range (Extended Data Fig. 7b).

To assign proteins to potential complexes, peak correlation profiling was performed. For this, the R package CCprofiler v.0.1 was used to analyse the peptides.txt output table from MaxQuant[61,97]. Peptide table entries were restricted to the maximum molecular mass per gene name for the quantification of protein groups. This step removed peptides mapping to smaller isoforms of each gene. The restricted proteinGroups.txt served as the trace annotation input table for CCprofiler. The calibration table of the SEC column was based on the high molecular mass gel filtration calibration kit (GE Healthcare). The peptides table was split into the different tissues and the resulting tables converted into the input format for CCprofiler. All subsequent steps were performed separately for each tissue table.

The peptides table was imported, converted into a traces object, where a trace object is the SEC-elution profile of a specific peptide and annotated with the corresponding fractions and additional external information (for example, molecular mass of the respective fraction, molecular mass of the respective protein) based on the aforementioned calibration and restricted proteinGroups tables. Traces not containing at least three consecutive peptide identifications across fractions, as well as traces with a sibling peptide correlation of less than 0.2 between peptides originating from the same protein, were filtered out. On the basis of the remaining traces, protein features (elution peak positional information) consisting of highly correlated peptides (correlation higher than 0.9) were detected using CCprofiler's sliding window algorithm. Peptides were randomly assigned to proteins to control the false discovery rate of protein features at 5% (random decoy model). Subsequently, protein traces were calculated by summing up the intensities of the top two most abundant peptide traces.

Next, complex hypotheses consisting of target and decoy protein complexes were constructed based on a previously published mapping of *Arabidopsis* orthologues to mammalian protein complexes from CORUM[98,99]. Proteins were only mapped to the same decoy complex if they were not interacting with each other directly (minimum path length of two). These target and decoy complexes were used during the detection of complex features consisting of highly correlated proteins using the same sliding window algorithm as above to control the false discovery rate of complex features at 5% (target-decoy model).

Both protein features (elution peak positional information) and protein traces (quantitative information) were then used to generate an input matrix for weighted gene correlation network analysis (WGCNA)[100], which was used to find clusters of proteins with correlated SEC elution profiles. Protein traces were restricted to the ones, which were part of at least one protein feature. Each protein trace was duplicated for each protein feature it was part of and all intensities outside of the respective protein feature were set to zero. With this, protein traces were effectively split into separate elution peaks corresponding to distinct (sub)complexes. Afterwards, elution peaks, which met either of the following criteria were filtered out: (1) the absolute difference between the molecular mass at the apex of the elution peak and the monomer molecular mass of the corresponding protein was smaller than twice the monomer molecular mass. (2) The absolute difference between the molecular mass at the apex of the elution peak and the monomer molecular weight of the corresponding protein was less than 200 kDa. (3) The apex of the elution peak was in fraction 77 (void fraction). (4) The apex of the elution peak was in fraction 4 (monomer range).

The remaining elution peaks were restricted to the intersection with the core protein dataset across tissues described above (see 'Tissue groups and tissue characteristics'), generating the SEC WGCNA input dataset. At the same time, an additional WGCNA input dataset was generated based on the core protein dataset across tissues and restricted to the intersection with the SEC WGCNA input dataset. Proteins with several SEC elution peaks were duplicated to match the SEC WGCNA input dataset. WGCNA was carried out separately for each of these datasets. Signed co-expression similarities were calculated between all pairs of proteins with at least seven pairwise-complete observations using the following formula:

$$s_{ij} = \frac{1 + \mathrm{cor}(x_i, x_j)}{2}$$

where s_{ij} is the signed co-expression similarity between two proteins x_i and x_j (based on their Pearson's correlation coefficient (*r*) across tissues). Adjacency matrices were calculated with $A = [s_{ij}]$ for several values of $\beta \in [1,30]$.

$$a_{ij} = s_{ij}^{\beta}$$

where a_{ij} is the adjacency between two proteins x_i and x_j. The adjacency function parameter β was selected to be 30 for the construction of signed protein co-expression networks. The topological overlap matrix $\Omega = [\omega_{ij}]$ of the two networks was calculated with:

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}}$$

where $l_{ij} = \sum_u a_{iu}a_{uj}$ and $k_i = \sum_u a_{iu}$. Hierarchical clustering was performed on the topological-overlap-based dissimilarity matrices $D = [1 - \omega_{ij}]$ using average linkage. Correlation clusters ('Modules') were detected using adaptive branch pruning[101] using the 'Dynamic Hybrid' method set to respect the dendrogram topology during PAM operations and requiring at least 30 members per module. Similar clusters in the SEC or tissue dataset were merged, if the dissimilarity of their module eigengenes was smaller than 1 − MPC or 1 − MTC, respectively. MPC is the median peak correlation of the FDR-filtered complex features from CCprofiler (known complexes), and MTC is the median tissue correlation of the very same complex features (pairwise correlation between proteins mapping to a specific complex feature). The resulting modules for both datasets were then mapped to each other using a combination of Fisher's exact test and manual curation. Enrichment of functional annotations from Corum[99], GO[72], KEGG[102] and Reactome[103] were calculated in each cluster using Fisher's exact test. All *P* values were adjusted for multiple-testing by calculating FDRs[104] (Supplementary Data 8).

To quantify how well complexes can be detected using the tissue atlas or SEC WGCNA output, we calculated a summary statistic termed 'complex index'. Protein interactor information was downloaded from

UniProt[6] (https://uniprot.org; December 2018) and manually curated for the presence of large (>4 subunits) and small (≤4 subunits) complexes (Extended Data Fig. 7d, Supplementary Data 9). The complex index (C) was calculated using the following formula:

$$C = \frac{S_m}{S} \times \frac{S_m}{M}$$

where S_m is the number of subunits detected as present in the same module. S is the total number of subunits identified in the restricted datasets for either SEC or tissue atlas experiments and M is the total number of entries in the respective module. The complex index is a measure of how well complex subunits co-occur (either by co-elution or by co-expression) and thus are detected in the same module and also quantifies the resolution of the detection method (module size). The complex index is 1 when all subunits of a complex are identified in the same module and no other proteins are contained in the module.

Complex stoichiometry analysis. Absolute SEC protein elution traces of selected complexes (chaperonin, 26S proteasome, CSN, CESA) were plotted using top-3 intensities (not log-transformed). Relative subunit proportions or stoichiometry for selected protein complexes within individual tissues of the tissue atlas dataset were calculated using top-3 intensities (based on the three most abundant unique peptides). Top-3 intensity rather than iBAQ intensity was used, to avoid distorted ratios caused by shared (razor) peptide assignment. In the case of paralogue genes with high protein sequence similarity, peptides were not required to be unique for one gene, but rather for one 'paralogue group'. The three most abundant peptides within a paralogue group were then used for the top-3 calculations.

For the example of the coatomer complex, the relative proportion of paralogue genes in different tissues was calculated by plotting the top-3 intensity (not log-transformed, unique for gene) of each paralogue as proportion of the summed top-3 intensity (not log-transformed) of all paralogues (Extended Data Fig. 7e). For the calculation of subunit stoichiometry ratios of selected complexes (chaperonin, 26S proteasome, CSN, CESA), the top-3 intensity (not log-transformed) of a subunit was divided by the average top-3 intensity (not log-transformed) of all complex subunits. Subunit ratios were first calculated for each tissue and subsequently averaged across all 30 tissues (chaperonin, 26S proteasome, CSN) or only across tissues where the respective subunits are mainly expressed (CESA4/7/8: node, internode, silique septum, silique valves; CESA1/3/6: all other) (Extended Data Fig. 7g).

Subnetwork extraction. De novo network enrichment was performed to identify tissue- or tissue-group-specific subnetworks and to link them to a molecular function. In contrast to gene set overrepresentation or gene set enrichment analysis[105], this approach is suited to identify previously uncharacterized functions from large-scale molecular interaction networks. KeyPathwayMiner[106] was used to extract tissue- and tissue-group-specific subnetworks from transcriptomics and proteomics datasets. To enrich subnetworks with proteins that deviate in their expression from other tissues, we used z-scored expression values. These reflect how many standard deviations the expression in a given tissue is away from the expression range found in all other tissues:

$$z_{x,g} = \frac{x - \mu_{X\backslash x}}{\sigma_{X\backslash x}}$$

where x is a specific tissue or tissue group of interest, g a gene, μ the mean, and σ the standard deviation. On the basis of the following rule, a one-column indicator matrix $I(x,g)$ was constructed as input for KeyPathwayMiner for each tissue or tissue group x:

$$I(x,g) = \begin{cases} 1 \text{ if } |z_{x,g}| \rangle 2 \\ 0 \text{ else} \end{cases}$$

Here each row corresponds to a gene g. STRING[30] *A. thaliana* network (v.10.5, downloaded October 2018) was used as the molecular interaction network, considering only high-confidence interactions with a score >900. Subnetworks extracted via KeyPathwayMiner were made available for individual tissues or tissue groups as part of the ATHENA web application.

Phosphorylation sites and motif analysis. Serine, threonine and tyrosine content and the P-site number of individual proteins was calculated based on the longest sequence in case of ambiguous isoform identifications (Fig. 5a, Extended Data Fig. 8g). Similarly, the schematic of phosphorylation (P-site) localization in LEA and receptor-like kinase proteins was based on the P-site localization in the longest isoform sequence (Extended Data Fig. 8h, i). Functional domain assignment for receptor-like protein kinase sequences was done using SMART (http://smart.embl-heidelberg.de/smart/batch.pl)[107] and P-sites were manually assigned to specific domains based on their localization in the protein (Extended Data Fig. 8i).

For motif analysis within our dataset, sequence windows of 15 amino acids centred on the identified class I P-sites (15 mers) were assigned to a motif class of proline-directed, acidic, basic or other by following a binary decision tree[36] (Extended Data Fig. 8d). Assignment to these categories was done sequentially as follows: Pro amino acid at position −1 (proline-directed), 5 or more Asp or Glu at position +1 to +7 (acidic), Arg or Lys at position −3 (basic), Asp or Glu at position +1, +2 or +3 (acidic), 2 or more Arg or Lys at position −6 to −1 (basic) and otherwise (other).

P-sites with high-confident localization scores (class I) were divided into Ser, Thr and Tyr P-site datasets. Motif extraction was performed separately for each motif class category (proline-directed, acidic, basic or other) using the motif-X algorithm, implemented in the rmotifx R package[35] (Supplementary Data 10). Cut-off settings were: min-seq. = 30, pval.cutoff = 1×10^{-6} (Ser and Thr datasets); min-seq. = 10, pval.cutoff = 1×10^{-5} (Tyr dataset). Motif-X calculates the fold-increase of a respective motif in the 'foreground' (P-site dataset) compared to the 'background' dataset (non-phosphorylated peptides). To avoid mass spectrometry-based residue biases, all peptides from our dataset that contained an STY amino acid and had only been identified in non-phosphorylated form were used as the background for all motif-X analyses (258,395 sequences). These peptide sequences were centred on STY and extended where necessary along the N- and C-terminal window to generate 15 mers. Position weight matrix sequence logos were drawn using the 'bits' method for amino acid sequences in the ggseqlogo R package[108] (Extended Data Fig. 8e, Supplementary Table 2). The motif-X fold increase for S motifs was plotted for motifs with 2, 3 and 4 fixed amino acid positions (Extended Data Fig. 8f).

Published P-site motifs were retrieved from PhosPhAT4.0[9] and the Human Protein Reference Database[109] (HPRD; www.hprd.org) and divided into groups depending on the number of fixed amino acid positions within the motif sequence. These motifs were subsequently matched to the P-site or background dataset to calculate detection frequency and fold increase (P-site/background dataset) of a reported motif (Supplementary Data 10).

AGCVIII protein abundance and mutant phenotypic characterization

For the relative expression analysis of AGC kinases across tissues, the AGC1 and AGC3 subfamilies of *Arabidopsis* AGCVIII kinases were selected[21]. The summed total intensity of AGC1 and AGC3 subfamily kinases (top-3, not log-transformed) was calculated for each tissue. Relative protein amounts of individual kinases in different tissues were calculated as the top-3 intensity (not log-transformed, unique for gene) of each kinase in proportion to the summed top-3 intensity (not log-transformed) of all AGC1 and AGC3 kinases (Fig. 2c). For clear visualization only the D6PK family (D6PK, D6PKL1, D6PKL2 and D6PKL3), AGC1.5, AGC1.6 and AGC1.7 were coloured in the plot.

Mutant alleles of the *D6PK* family kinase genes—*d6pk-1*, *d6pkl1-1*, *d6pkl2-2* and *d6pkl3-2*—and their combinations were previously described[110]. Embryos were prepared omitting fixation with ethanol and acetic acid as previously described[111]. Embryos were analysed with a Zeiss Axio Imager.M2, Axiocam 512 camera and 20×/0.8 Plan Apochromat objective using differential interference contrast.

For promoter:GUS constructs of *AGC1.5*, *AGC1.6* or *AGC1.7*, 2,916-bp, 1,207-bp or 2,214-bp fragments upstream from the respective start codon were cloned into the SalI and EcoRI sites of pCAMBIA1391Z. Flowers from transgenic plants, obtained by the floral-dip method[48], were first incubated in 90% cold acetone for 15 min followed by a β-glucuronidase (GUS)-staining solution (50 mM sodium phosphate pH 7.0, 10 mM EDTA, 2 mM potassium ferricyanide, 2 mM potassium ferrocyanide, 0.1% Triton X-100, 0.5 mg ml^{-1} 5-bromo-4-chloro-3-indolyl-β-D-glucuronide) overnight in the dark. After a wash in 70% ethanol, flowers were incubated in ethanol/acetate (6:1) for removing of chlorophyll and then rehydrated in a graded ethanol series (90%, 70%, 50% and 30%). Samples were mounted in chloral hydrate/glycerol/water (8 g:1 ml: 2 ml) for imaging. The experiment was performed twice and GUS signal was observed for 12/18 *AGC1.5p*::*GUS*, 0/20 *AGC1.6p*::*GUS* and 12/18 *AGC1.7p*::*GUS* independent lines.

ABA response of RCAR phosphomutants

Preparation and analysis of *Arabidopsis* protoplasts was performed as previously described[112]. In brief, protoplasts from 3-week-old Col-0 plants (10^5 protoplasts; 0.1 ml) were transfected with 5 μg of reporter construct (*pRD29B::LUC*), 3 μg of *p35S::GUS* plasmid as internal control and 3 μg of effector plasmid. The effector plasmids drive expression of respective RCAR and PP2C cDNAs under control of the 35S promoter[113]. RCAR10 phosphomimetic reporter constructs (RCAR10 mutations S32D and S113D) were obtained by site-directed mutagenesis using primer pairs S32D_F/S32D_R and S113D_F/S113D_R, respectively (Supplementary Data 11). The protoplast suspension was incubated at various levels of ABA as indicated and reporter expression was determined after 18 h of incubation at 25 °C. Three biological replicates per data point were performed for each assay. All data points were normalized to the empty vector control. Structural modelling for RCAR10 was performed with SWISS-MODEL[114] using RCAR11 as template model (PDB code 3K3K[115,116]) and modified with UCSF CHIMERA[117].

Phenotypic characterization of QKY phosphomutants

The pCAMBIA2300-based pQKY::mCherry:QKY construct was described previously[42]. The pQKY::mCherry:QKY(S262A) and pQKY::mCherry:QKY(S262E) plasmids were obtained using the Q5 site-directed mutagenesis kit (NEB, E0554S) according to the manufacturer's recommendation. Primers are Q5SDM S262A_F and Q5SDM S262A_R for pQKY::mCherry:QKY(S262A) and Q5SDM S262E_F and Q5SDM S262E_R for pQKY::mCherry:QKY(S262E) (Supplementary Data 11). All PCR-based constructs were confirmed by sequencing. Floral tissue for quantitative PCR (qPCR) was collected from plants grown under long day conditions. With minor changes, tissue collection, RNA extraction and qPCR analysis were performed as previously described[118,119]. For detection of *mCherry:QKY* expression, qPCR analysis was performed with primers mCherry-qRT-For and mCherry-qRT-Rev (Supplementary Data 11). Average *mCherry:QKY* expression was calculated relative to the expression of three control genes (*AT4G33380*, *AT2G28390* and *AT5G46630*) (Supplementary Data 11) and normalized to the wild-type control for visualization. *A. thaliana* (L) Heynh. var. Landsberg *erecta* (L*er*) was used as wild-type strain. The likely null allele *qky*-9 (L*er*) has previously been described[43]. *qky-9* mutant plants were transformed using *Agrobacterium* strain GV3101/pMP90[120] and the floral dip method[48]. Transgenic T_1 plants were selected on kanamycin (50 μg ml^{-1}) and transferred to soil for further inspection. Plants were grown as previously described[43]. In total, 4 out of 17 independent T_1 transformants showed a wild-type phenotype for the S262A transgenic line, whereas 13 out of 17 T_1 transformants displayed no (7 out of 17) or partial (6 out of 17) rescue, a notable decrease in functionality compared to the wild-type construct (14 out of 17 rescue, 2 out of 17 partial rescue, 1 out of 17 no rescue). For the S262E transgene, 11 out of 13 independent T_1 transformants showed a wild-type phenotype (1 out of 13 partial rescue, 1 out of 13 no rescue).

Floral organs were imaged using a Leica SAPO stereo microscope equipped with a digital MC 170 HD camera (Leica Microsystems GmbH). Images were adjusted for colour and contrast using ImageJ/Fiji software[121]. Confocal laser scanning microscopy of *qky-9 pQKY::mCHerry:QKY*, *qky-9 pQKY::mCHerry:QKY(S262A)* and *qky-9 pQKY::mCHerry:QKY(S262E)* six-day-old seedling roots was performed with an Olympus FV1000 setup using an inverted IX81 stand and FluoView software (FV10-ASW v.01.04.00.09) (Olympus Europa) equipped with a water-corrected 40× objective (NA 0.9) at 3× digital zoom. Confocal high sensitivity detection was used involving two gallium arsenide phosphide photomultipliers mounted equidistantly to the probe. The experiment was performed twice independently with 2 and 12 roots for *qky-9 pQKY::mCHerry:QKY*, 6 and 15 roots for *qky-9 pQKY::mCHerry:QKY(S262A)* and 15 roots for *qky-9 pQKY::mCHerry:QKY(S262E)*.

Locus identifiers

Gene locus identifiers are listed for all gene names mentioned in the manuscript (Supplementary Data 11).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The data supporting the findings of this study are available within the paper, the Supplementary Information and the public repositories. Source Data for Figs. 1–5 and Extended Data Figs. 1–9 are included with the paper. Transcriptome sequencing and quantification data are available at ArrayExpress (www.ebi.ac.uk/arrayexpress) under the identifier E-MTAB-7978. The raw mass spectrometric data and MaxQuant result files have been deposited to the ProteomeXchange Consortium via PRIDE[122], with the dataset identifier PXD013868.

44. Smyth, D. R., Bowman, J. L. & Meyerowitz, E. M. Early flower development in *Arabidopsis*. *Plant Cell* **2**, 755–767 (1990).
45. Johnson-Brousseau, S. A. & McCormick, S. A compendium of methods useful for characterizing *Arabidopsis* pollen mutants and gametophytically-expressed genes. *Plant J*. **39**, 761–775 (2004).
46. Sprunck, S. et al. Egg cell-secreted EC1 triggers sperm cell activation during double fertilization. *Science* **338**, 1093–1097 (2012).
47. Karimi, M., Inzé, D. & Depicker, A. GATEWAY vectors for *Agrobacterium*-mediated plant transformation. *Trends Plant Sci*. **7**, 193–195 (2002).
48. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J*. **16**, 735–743 (1998).
49. Schmid, M. et al. A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet*. **37**, 501–506 (2005).
50. Boyes, D. C. et al. Growth stage-based phenotypic analysis of *Arabidopsis*: a model for high throughput functional genomics in plants. *Plant Cell* **13**, 1499–1510 (2001).
51. Bowman, J. L. Arabidopsis*: an Atlas of Morphology and Development* (Springer-Verlag, 1994).
52. Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem*. **72**, 248–254 (1976).
53. Ruprecht, B. et al. Optimized enrichment of phosphoproteomes by Fe-IMAC column chromatography. *Methods Mol. Biol*. **1550**, 47–60 (2017).
54. Marx, H. et al. A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nat. Biotechnol*. **31**, 557–564 (2013).
55. Ruprecht, B., Zecha, J., Zolg, D. P. & Kuster, B. High pH reversed-phase micro-columns for simple, sensitive, and efficient fractionation of proteome and (TMT labeled) phosphoproteome digests. *Methods Mol. Biol*. **1550**, 83–98 (2017).
56. Smith, P. K. et al. Measurement of protein using bicinchoninic acid. *Anal. Biochem*. **150**, 76–85 (1985).
57. Zolg, D. P. et al. PROCAL: a set of 40 peptide standards for retention time indexing, column performance monitoring, and collision energy calibration. *Proteomics* **17**, (2017).

58. Hahne, H. et al. DMSO enhances electrospray response, boosting sensitivity of proteomic experiments. *Nat. Methods* **10**, 989–991 (2013).
59. Bian, Y. et al. Robust, reproducible and quantitative analysis of thousands of proteomes by micro-flow LC-MS/MS. *Nat. Commun*. **11**, 157 (2020).
60. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protocols* **11**, 2301–2319 (2016).
61. Hanada, K. et al. sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics* **26**, 399–400 (2010).
62. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol*. **29**, 644–652 (2011).
63. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282–283 (2001).
64. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
65. Franken, H. et al. Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry. *Nat. Protocols* **10**, 1567–1593 (2015).
66. Toprak, U. H. et al. Conserved peptide fragmentation as a benchmarking tool for mass spectrometers and a discriminating feature for targeted proteomics. *Mol. Cell. Proteomics* **13**, 2056–2071 (2014).
67. Oñate-Sánchez, L. & Vicente-Carbajosa, J. DNA-free RNA isolation protocols for *Arabidopsis thaliana*, including seeds and siliques. *BMC Res. Notes* **1**, 93 (2008).
68. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
69. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol*. **34**, 525–527 (2016).
70. Silva, J. C., Gorenstein, M. V., Li, G. Z., Vissers, J. P. & Geromanos, S. J. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol. Cell. Proteomics* **5**, 144–156 (2006).
71. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res*. **45** (D1), D331–D338 (2017).
72. Cox, J. & Mann, M. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics* **13** (Suppl. 16), S12 (2012).
73. Tyanova, S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
74. Olsen, J. V. et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**, 635–648 (2006).
75. Uhlén, M. et al. Transcriptomics resources of human tissues and organs. *Mol. Syst. Biol*. **12**, 862 (2016).
76. Rijpkema, A. S., Vandenbussche, M., Koes, R., Heijmans, K. & Gerats, T. Variations on a theme: changes in the floral ABCs in angiosperms. *Semin. Cell Dev. Biol*. **21**, 100–107 (2010).
77. Heazlewood, J. L., Verboom, R. E., Tonti-Filippini, J., Small, I. & Millar, A. H. SUBA: the Arabidopsis Subcellular Database. *Nucleic Acids Res*. **35**, D213–D218 (2007).
78. Löytynoja, A. Phylogeny-aware alignment with PRANK. *Methods Mol. Biol*. **1079**, 155–170 (2014).
79. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol*. **17**, 540–552 (2000).
80. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol*. **24**, 1586–1591 (2007).
81. van der Graaf, A. et al. Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. *Proc. Natl Acad. Sci. USA* **112**, 6676–6681 (2015).
82. Gebert, D., Jehn, J. & Rosenkranz, D. Widespread selection for extremely high and low levels of secondary structure in coding sequences across all domains of life. *Open Biol*. **9**, 190020 (2019).
83. Camiolo, S., Melito, S. & Porceddu, A. New insights into the interplay between codon bias determinants in plants. *DNA Res*. **22**, 461–470 (2015).
84. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA* **102**, 14338–14343 (2005).
85. Das, S. & Bansal, M. Variation of gene expression in plants is influenced by gene architecture and structural properties of promoters. *PLoS ONE* **14**, e0212678 (2019).
86. Celaj, A. et al. Quantitative analysis of protein interaction network dynamics in yeast. *Mol. Syst. Biol*. **13**, 934 (2017).
87. Niederhuth, C. E. et al. Widespread natural variation of DNA methylation within angiosperms. *Genome Biol*. **17**, 194 (2016).
88. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. **15**, 550 (2014).
89. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet*. **44**, 821–824 (2012).
90. Nakazawa, N. fmsb: functions for medical statistics book with some demographic data. R package v.0.6.3; https://CRAN.R-project.org/package=fmsb (2018).
91. Zhang, Z. Variable selection with stepwise and best subset approaches. *Ann. Transl. Med*. **4**, 136 (2016).
92. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. A Stat. Soc*. **58**, 267–288 (1996).
93. R Core Team. R: A language and environment for statistical computing. https://www.R-project.org/ (R Foundation for Statistical Computing, 2014).
94. Knecht, W. *Pilot Willingness to Take Off Into Marginal Weather, Part II: Antecedent Overfitting With Forward Stepwise Logistic Regression*. Final Report DOT/FAA/AM-05/15 (Federal Aviation Administration, 2005).
95. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw*. **33**, 1–22 (2010).
96. Groemping, U. Relative importance for linear regression in R: the package relaimpo. *J. Stat. Softw*. **17**, 1–27 (2007).
97. Heusel, M. et al. Complex-centric proteome profiling by SEC-SWATH-MS. *Mol. Syst. Biol*. **15**, e8438 (2019).
98. McBride, Z., Chen, D., Reick, C., Xie, J. & Szymanski, D. B. Global analysis of membrane-associated protein oligomerization using protein correlation profiling. *Mol. Cell. Proteomics* **16**, 1972–1989 (2017).
99. Ruepp, A. et al. CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Res*. **38**, D497–D501 (2010).
100. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol*. **4**, Article17 (2005).
101. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008).
102. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. **45** (D1), D353–D361 (2017).
103. Fabregat, A. et al. The Reactome pathway Knowledgebase. *Nucleic Acids Res*. **44** (D1), D481–D487 (2016).
104. Hochberg, Y. B. Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. A Stat. Soc*. **57**, 289–300 (1995).
105. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
106. List, M. et al. KeyPathwayMinerWeb: online multi-omics network enrichment. *Nucleic Acids Res*. **44** (W1), W98–W104 (2016).
107. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res*. **46** (D1), D493–D496 (2018).
108. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645–3647 (2017).
109. Goel, R., Harsha, H. C., Pandey, A. & Prasad, T. S. Human Protein Reference Database and Human Proteinpedia as resources for phosphoproteome analysis. *Mol. Biosyst*. **8**, 453–463 (2012).
110. Zourelidou, M. et al. The polarly localized D6 PROTEIN KINASE is required for efficient auxin transport in *Arabidopsis thaliana*. *Development* **136**, 627–636 (2009).
111. Mayer, U. B. G. & Jurgens, G. Apical-basal pattern formation in the *Arabidopsis* embryo: studies on the role of the gnom gene. *Development* **177**, 149–162 (1993).
112. Moes, D., Himmelbach, A., Korte, A., Haberer, G. & Grill, E. Nuclear localization of the mutant protein phosphatase abi1 is required for insensitivity towards ABA responses in *Arabidopsis*. *Plant J*. **54**, 806–819 (2008).
113. Tischer, S. V. et al. Combinatorial interaction network of abscisic acid receptors and coreceptors from *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **114**, 10280–10285 (2017).
114. Waterhouse, A. et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. **46** (W1), W296–W303 (2018).
115. Nishimura, N. et al. Structural mechanism of abscisic acid binding and signaling by dimeric PYR1. *Science* **326**, 1373–1379 (2009).
116. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res*. **28**, 235–242 (2000).
117. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem*. **25**, 1605–1612 (2004).
118. Box, M. S., Coustham, V., Dean, C. & Mylne, J. S. Protocol: A simple phenol-based method for 96-well extraction of high quality RNA from *Arabidopsis*. *Plant Methods* **7**, 7 (2011).
119. Enugutti, B. et al. Regulation of planar growth by the *Arabidopsis* AGC protein kinase UNICORN. *Proc. Natl Acad. Sci. USA* **109**, 15060–15065 (2012).
120. Koncz, C. & Schell, J. The promoter of TL-DNA gene 5 controls the tissue-specific expression of chimaeric genes carried by a novel type of *Agrobacterium* binary vector. *Molecular and General Genetics MGG* **204**, 383–396 (1986).
121. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
122. Vizcaíno, J. A. et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res*. **44** (D1), D447–D456 (2016).
123. Kwok, S. F. et al. *Arabidopsis* homologs of a c-Jun coactivator are present both in monomeric form and in the COP9 complex, and their abundance is differentially affected by the pleiotropic cop/det/fus mutations. *Plant Cell* **10**, 1779–1790 (1998).

Acknowledgements We thank the NGS@tum core facility for RNA sequencing, R. Tofanelli for help with imaging the ovules, R. J. Schmitz for providing data access for the feature analysis and M. Reinecke, F. Bayer and S. Galinec for mass spectrometry measurements. This work was in part funded by the German Science Foundation (DFG, SFB924), a research fellowship to H.S. by the Japan Society for the Promotion of Sciences, and a research fellowship to X.C. by the Chinese Research Council.

Author contributions J.M. performed (phosho)proteomic and transcriptomic experiments under the supervision of B.K. S.R. and H.S. performed AGC kinase experiments in plants under the supervision of G.J. and C.S. M.P., A.C. and X.C. performed phosphomutant analysis under the supervision of E.G. and K.S. P.C. and S.S. generated and provided plant material. J.M., M.F., M.M., D.L., S.A., D.P.Z., T.M., C.D., A.D. and R.R.H. performed data analysis under the supervision of B.K., K.F.X.M., P.F., M.B., T.H. and F.J. M.L., P.S. and T.S. generated *Arabidopsis* resource databases under supervision of M.W. and J.B. J.M., C.S. and B.K. conceptualized the project and wrote the manuscript. All authors edited the manuscript.

Competing interests M.W. and B.K. are founders and shareholders of OmicScouts GmbH and msAId GmbH. They have no operational role in the companies. M.F. and D.P.Z. are founders and shareholders of msAId GmbH. T.M. and M.B. are employees and/or shareholders of Cellzome GmbH. The remaining authors declare no competing interests.

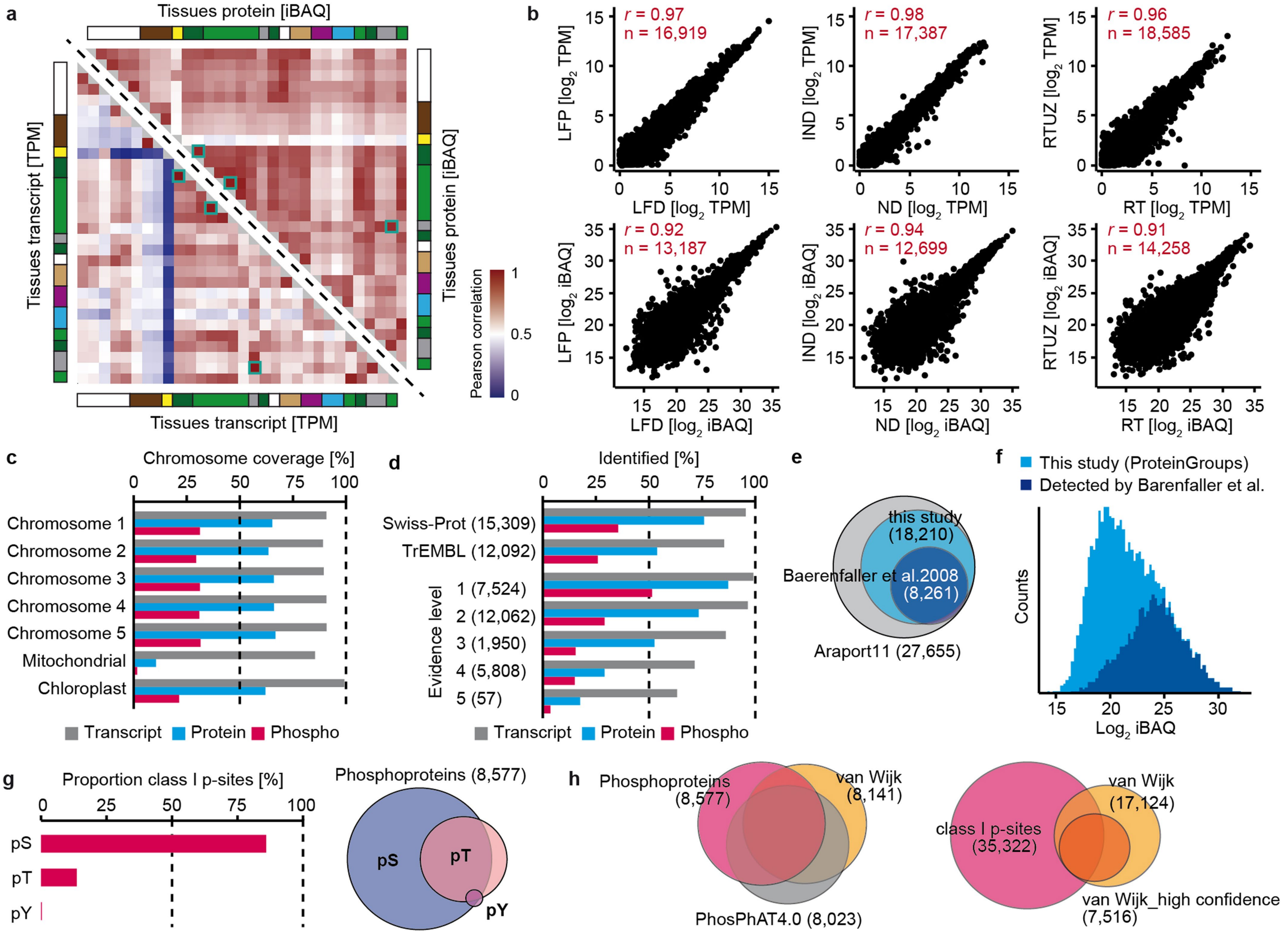
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2094-2.

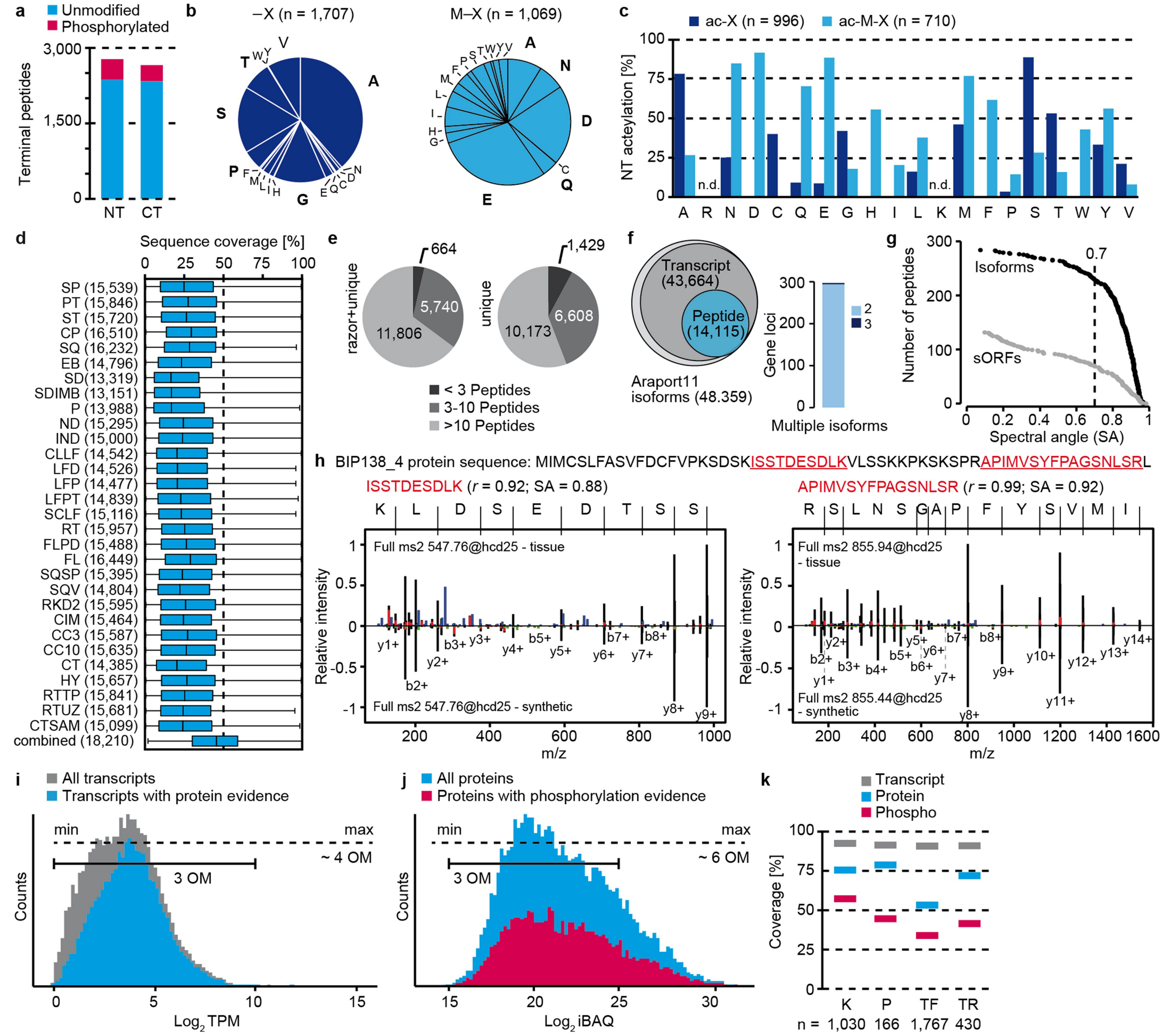
Correspondence and requests for materials should be addressed to B.K.

Peer review information *Nature* thanks José Dinneny, Paul Haynes and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

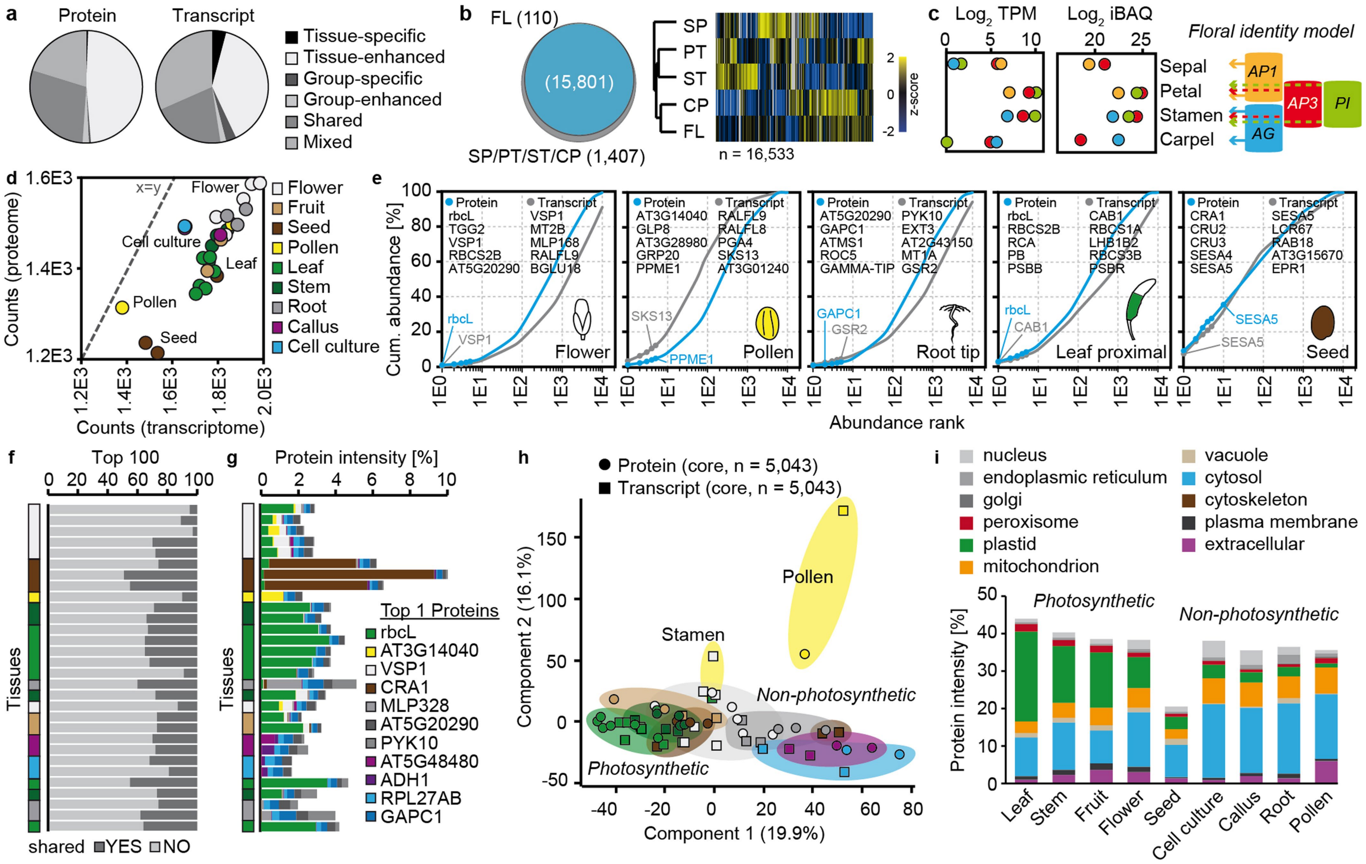
Reprints and permissions information is available at http://www.nature.com/reprints.



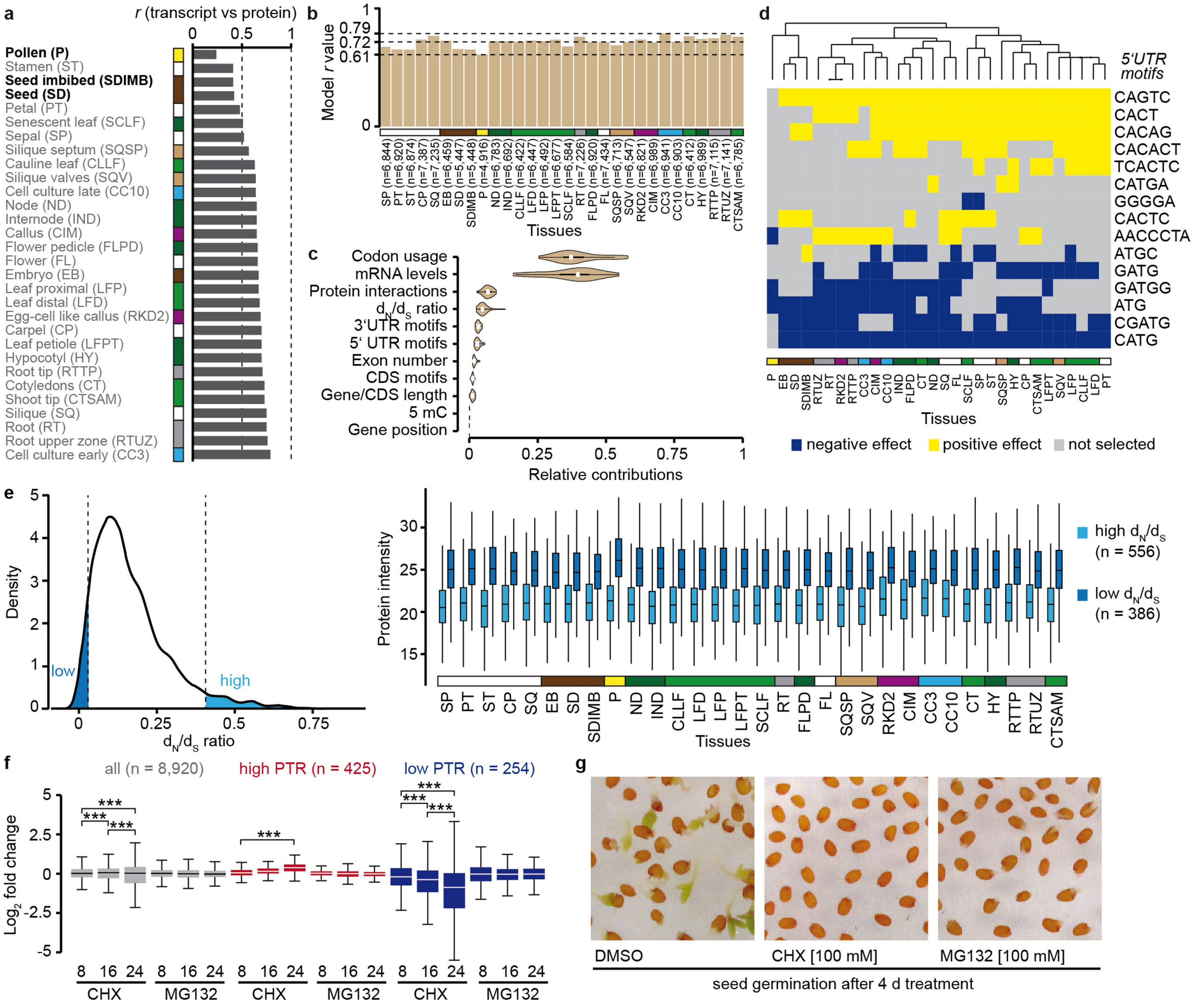
Extended Data Fig. 1 | Descriptive analysis of the multi-omic tissue atlas. **a**, Pairwise global Pearson's expression correlation analysis of all 30 tissues ($n = 1$ measurement per tissue) on the transcriptome level (bottom triangle) and proteome level (top triangle) using all identified gene loci. Proteins correlate more strongly between tissues than transcripts. Turquoise squares mark examples for morphologically highly similar tissues. Tissues are coloured as in Fig. 1. **b**, Scatter plots showing highly reproducible abundance measurements for transcript (top) and protein (bottom) in morphologically similar tissues that were marked in **a**; namely, node (ND) versus internode (IND), leaf distal (LFD) versus leaf proximal (LFP) and root (RT) versus root upper zone (RTUZ). *r* denotes the Pearson's correlation coefficient; *n* denotes the number of transcripts or proteins. **c**, Percentage of genes encoded by a specific chromosome that were identified at the transcriptome, proteome or phosphoproteome level. **d**, Percentage of Swiss-Prot and TrEMBL protein database entries as well as protein evidence categories from UniProt that were identified at the transcriptome, proteome or phosphoproteome level. Evidence level: (1) protein evidence; (2) transcript evidence; (3) homology; (4) predicted; and (5) uncertain. **e**, Comparison of protein identifications between an earlier *Arabidopsis* proteome study[7] based on 12 tissues, this study (30 tissues) and the number of protein-coding genes in Araport11. **f**, iBAQ intensity distribution of proteins identified in this study. Proteins also identified in a previous study[7] are projected into the same plot. **g**, Left, proportion of identified P-sites on S, T or Y residues with highly confident localization of the phosphorylation site within the identified peptide sequence (termed class I P-sites if the localization score is greater than 0.75). Right, distribution of proteins for which phosphorylated S, T or Y residues were identified. **h**, Left, Venn diagram comparing phosphoprotein datasets from a previous publication[8], PhosPhAT4.0 and this study. Right, Venn diagram comparing P-site localization confidence between class I sites identified in this study and the low and high confidence datasets reported in a previous publication[8].



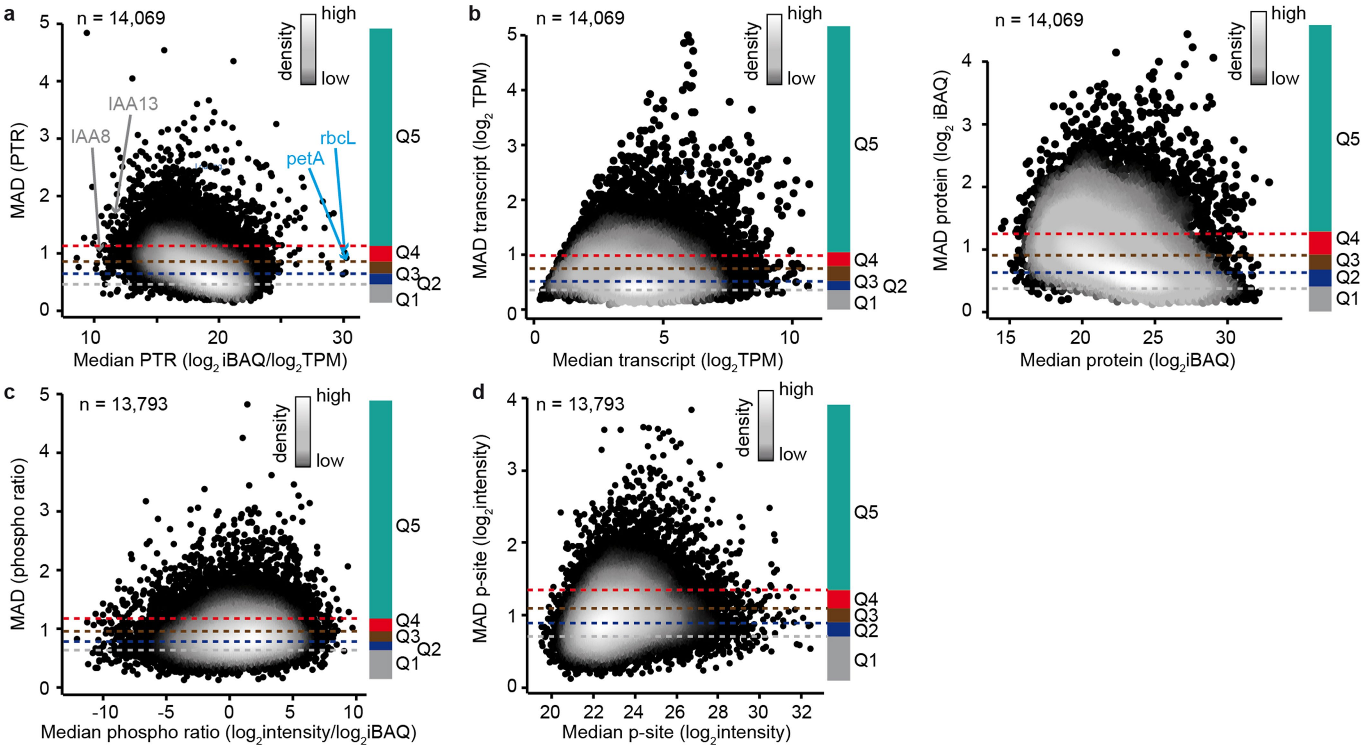
Extended Data Fig. 2 | Proteogenomics and dynamic range of transcript and protein expression. **a**, Number of identified N-terminal (NT) or C-terminal (CT) peptides of proteins in either unmodified or phosphorylated form. **b**, Frequency of amino acids following the initiator methionine in N-terminal peptides with (–X) or without (M–X) cleavage of the initiator methionine. X denotes the amino acid after the start codon. **c**, Frequency of protein N-terminal acetylation for amino acids in **b**. Because trypsin was used for protein digestion, the frequencies for Arg and Lys residues could not be determined (n.d.). **d**, Distribution of peptide-based sequence coverage of proteins in individual tissues and for the combined dataset (tissue abbreviations as in Fig. 1). Boxes contain 50% of the data and show the median as a black line. The top and bottom quartile ranges are shown as whiskers. The number of proteins is indicated for each tissue. **e**, Pie charts showing the percentage of proteins identified by <3, 3–10 or >10 peptides either allowing shared (razor) peptides or restricting to unique peptides only. **f**, Left, number of protein isoforms detected at the transcript and protein level compared with the number of all annotated isoforms in Araport11. Right, number of multiple isoforms of the same gene distinguished at the peptide level. **g**, Validation of protein isoform and sORF identification by comparing the tandem mass spectra from the tissue atlas to those of synthetic peptide reference standards. The normalized spectral contrast angle (SA) was used as a similarity metric (Methods). Candidate isoforms and sORFs were considered valid if the spectral contrast angle of the spectra was >0.7. These data are reported in Supplementary Data 3. **h**, Amino acid sequence and mirror plots of tandem mass spectra for two peptides of the sORF BIP138_4. The spectra pointing upwards were collected from tissue digests; those pointing downwards were collected from synthetic peptides. The normalized spectral contrast angle and Pearson's correlation coefficient (*r*) were used as similarity metrics (Methods) and indicate that both high-scoring spectra ($n = 1$ acquired spectra) are near identical, thus validating the identification of this sORF as an expressed protein. **i**, Dynamic range of transcript abundance (grey) and proportion of transcripts that were also identified at the protein level projected into this plot (blue). OM, orders of magnitude. Note that for lower abundance transcripts, fewer proteins were detected. **j**, Dynamic range of protein abundance and proportion of proteins with phosphorylation evidence. Protein abundance spans six orders of magnitude, whereas transcript abundance only spans four (**i**). In addition, note that phosphorylation was detected across the entire protein abundance range. **k**, Percentage of all annotated kinases (K), phosphatases (P), transcription factors (TF) and transcription regulators (TR) detected at the transcript, protein or phosphoprotein levels. Numbers below the *x* axis denote the number of genes for these protein classes in the *A. thaliana* genome.



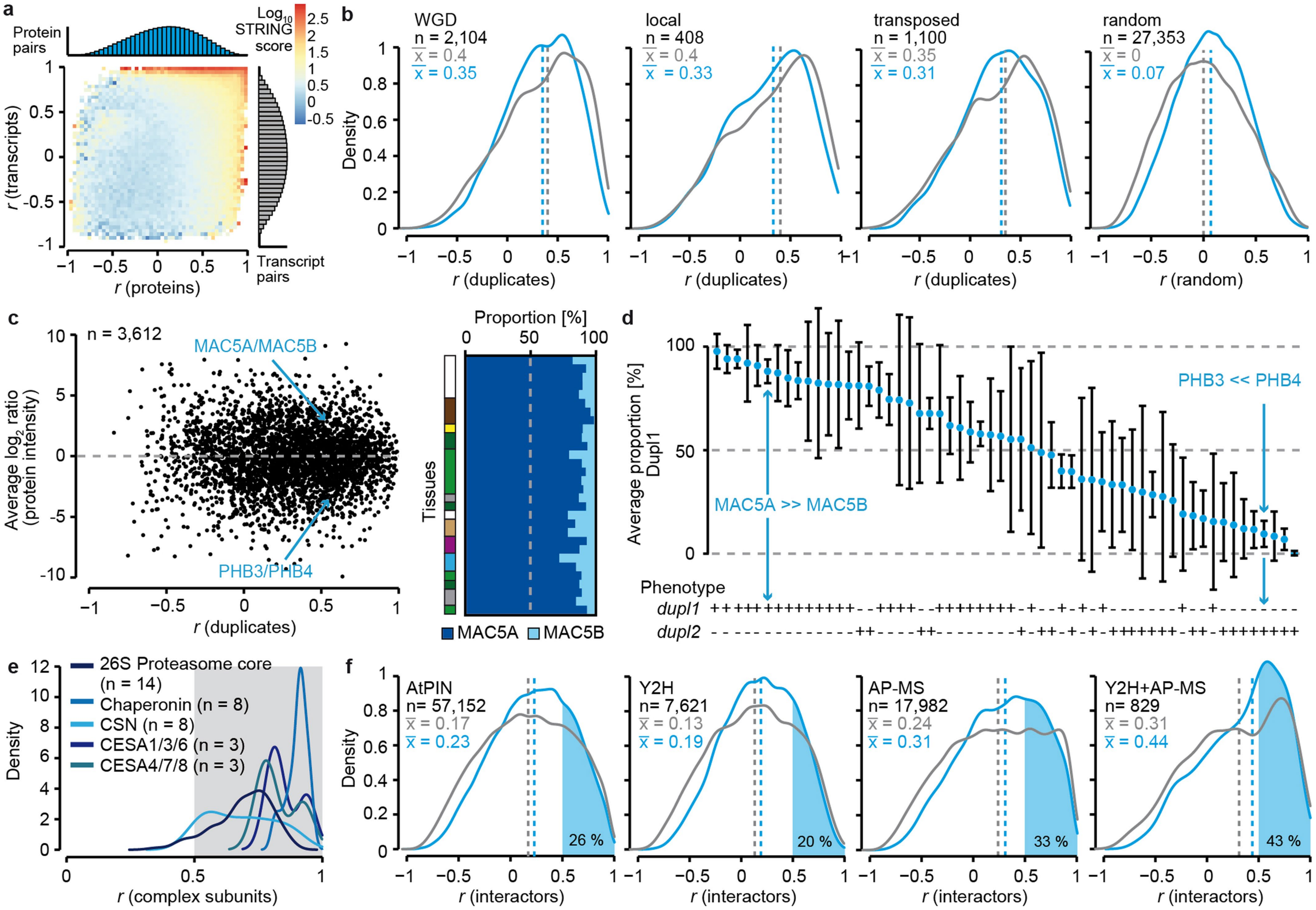
Extended Data Fig. 3 | Descriptive analysis of transcript and protein expression in tissues. a, Distribution of expression specificity categories for protein and transcript identifications. See Methods for the definition of these categories. In brief, there are very few transcripts and proteins that are only expressed in a single tissue. The quantities of the shared transcripts or proteins can differ vastly between tissues (**b**). **b**, Left, protein identifications shared between flower (FL) and flower organs showing an almost complete qualitative overlap of proteins. Sepal (SP), petal (PT), stamen (ST), carpel (CP). Right, clustering of *z*-scored protein intensities showing distinct quantitative expression differences between flower organs. **c**, Expression analysis of flower organ identity marker at the protein and transcript level. PISTILLATA (PI, green), APETALA3 (AP3, red), APETALA1 (AP1, orange), AGAMOUS (AG, blue). The expression of these markers is in line with the model of flower organ identity (AP1 expression marking sepal, AP1, AP3, PI marking petal, AG, AP3, PI marking stamen and AG marking carpel). **d**, Total number of transcripts plotted against the total number of proteins detected in each individual tissue ($n = 30$ tissues) showing that the more genes are expressed as mRNAs, the more proteins can be detected in a tissue (Pearson's correlation $r = 0.79$). Tissues are coloured according to tissue groups as in Fig. 1. **e**, Cumulative abundance plots of intensity-ranked identifications of transcripts and proteins for five representative tissues. The five most abundant transcripts and proteins are listed in descending order for each tissue. These are generally not the same. In addition, note that the characteristics of the plots are not the same for all tissues. In flower, the protein line rises more quickly than the transcript line. The opposite is true for pollen and a more even characteristic is observed in seed. **f**, Distribution of shared and unique identifications among the 100 most abundant transcripts and proteins in each tissue. Relatively few proteins and transcripts are found together on the list of the 100 most abundant transcripts and proteins. This demonstrates that the quantitative differences in transcript and protein expression are more important in defining a tissue than the qualitative expression of transcripts or proteins. **g**, List of 11 proteins that were found as the most abundant protein (in at least one tissue) and their proportion of the total iBAQ intensity in each tissue. Individual proteins can represent up to 9% of the total protein in a given tissue. **h**, Principal component analysis (PCA) of the core tissue proteomes and transcriptomes (that is, the proteins and transcripts that were identified in every tissue) using *z*-scored abundances. Only about 30% of all protein and 20% of all mRNAs were detected in every of the 30 tissues despite the fact that all tissues were deeply profiled at both protein and transcript level. This shows that strong qualitative and quantitative expression differences exist between tissues. The PCA separates tissues into photosynthetically active versus inactive tissues (component 1) and separates pollen from all other tissues (component 2), indicating that the molecular composition of pollen is particularly different from all other tissues. **i**, Proportion of the total summed protein intensity for genes with specific subcellular compartment annotation (from SUBA[77]; Methods) in the different tissue groups. The comparison of photosynthetically active and inactive tissues shows that most of the protein content in photosynthetically active tissues is contained in the plastids, whereas most protein is found in the cytosol for photosynthetically inactive tissues. Proteins with only one single subcellular compartment annotation were selected for the plot and the proportion of their iBAQ intensities were averaged for each tissue group. Nucleus ($n = 1,393$), endoplasmic reticulum ($n = 58$), Golgi ($n = 68$), peroxisome ($n = 67$), plastid ($n = 525$), mitochondrion ($n = 317$), vacuole ($n = 71$), cytosol ($n = 385$), cytoskeleton ($n = 1$), plasma membrane ($n = 268$), extracellular ($n = 351$).



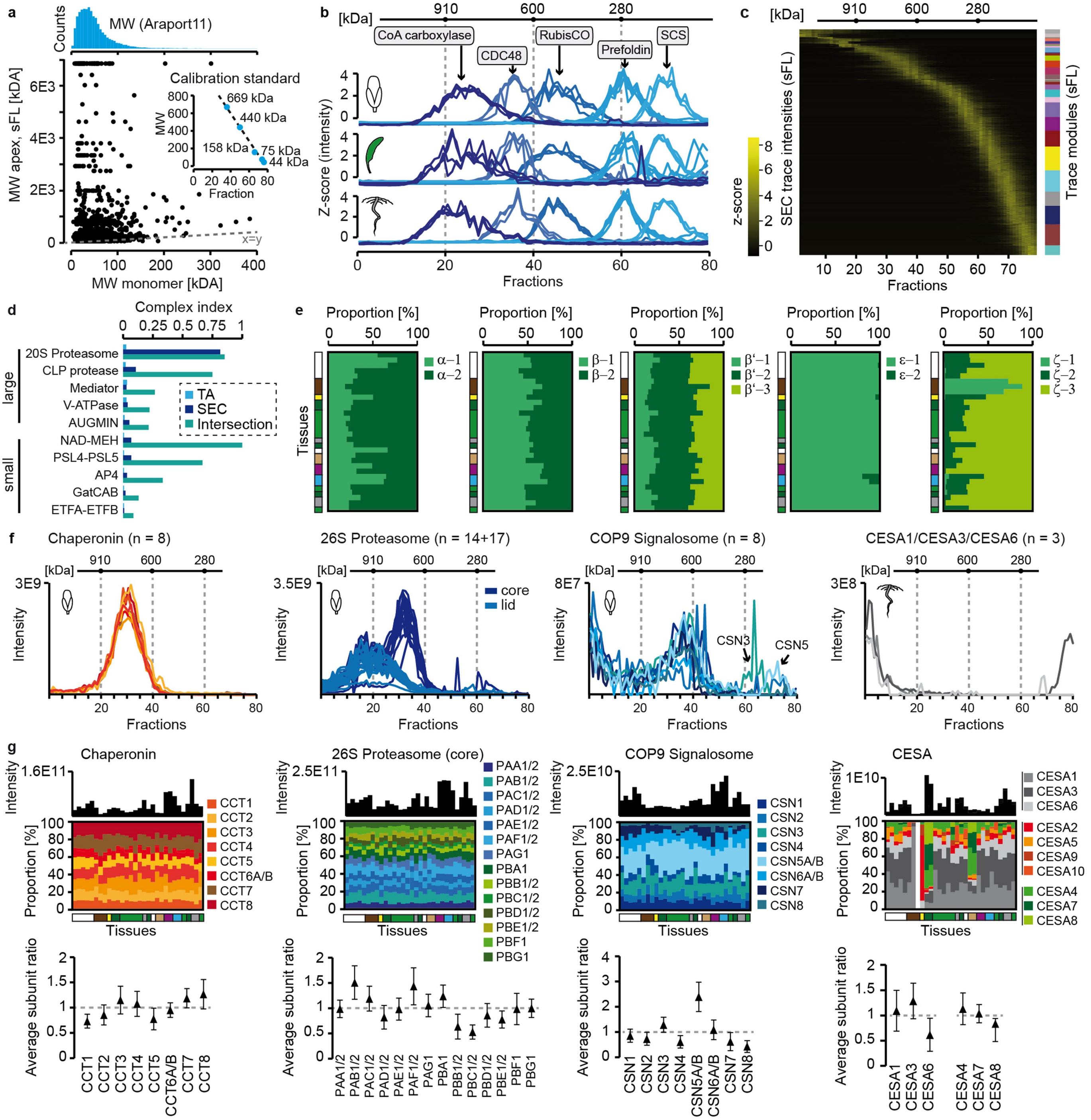
Extended Data Fig. 4 | Relationships between transcript and protein levels. **a**, Pearson's correlation (*r*) of transcriptome and proteome expression (core datasets; *n* = 5,043) for each tissue. **b**, Pearson's correlation between measured and predicted protein abundance levels in all tissues. Predicted protein abundance levels were obtained from the best fitting feature selection model for each tissue (Methods). The number of genes used for the correlation analysis is indicated for each tissue. **c**, Violin plots showing the spread in relative contribution of selected features to the prediction of gene-level protein abundance across tissues (*n* = 30 tissues) using our model. Violin shapes show the kernel density estimation of the data distribution and the median as white dot. Thick black bars denote the interquartile range. **d**, Specific nucleotide sequence motifs in 5′ UTRs of mRNAs contribute to the prediction of protein levels in a subset of tissues. Clustering tissues based on the presence or absence of detected 5′ UTR motifs shows that several features are repeatedly selected for inclusion in the model while others appear to be more tissue-specific. **e**, On the basis of the observation that the d_N/d_S ratio between orthologous of *A. thaliana* and *A. lyrata* contributed to the prediction of protein levels (**c**), we analysed this feature in more detail. Left, distribution of the d_N/d_S ratio for orthologous genes in *A. thaliana* and *A. lyrata*. The distribution is plotted for the example of 'leaf distal' (*n* = 6,447 genes). To compare evolutionarily conserved genes (defined by low d_N/d_S ratios) and genes that evolve neutrally or are under positive selection (high d_N/d_S ratios), we selected the bottom 5% and top 5% of the d_N/d_S ratio distribution, respectively. Right, evolutionarily conserved genes (low d_N/d_S ratio) show 10–20 times higher protein abundance than genes under evolutionary pressure. Boxes contain 50% of the data and show the median as a black line. Whiskers denote 1.5 times the interquartile range. Outliers were omitted from the plot for clarity. **f**, Time-course analysis of median protein abundance changes after treatment with CHX (translation block) or MG132 (proteasome block) versus time-matched DMSO control samples (Methods). Boxes contain 50% of the data and show medians as black lines. Whiskers denote 1.5 times the interquartile range. Outliers were omitted from the plot for clarity but were included in the statistical tests below. All proteins in the experiment (*n* = 8,920, grey), proteins that have a high PTR in seed (*n* = 425, red) or a low PTR in seed (*n* = 254, blue) (defined as in Fig. 3d) are shown. Differences between time points were tested for significance within each subset (all; high PTR; low PTR) using one-way ANOVA and the post hoc Tukey HSD test. ***$P < 0.001$ (all_CHX8–CHX16: $P < 1 \times 10^{-7}$; all_CHX8–CHX24: $P < 1 \times 10^{-7}$; all_CHX16–CHX24: $P = 0.0002$; highPTR_CHX8–CHX24: $P = 0.0003$; lowPTR_CHX8–CHX16: $P = 0.0000004$; lowPTR_CHX8–CHX24: $P < 1 \times 10^{-7}$; lowPTR_CHX16–CHX24: $P < 1 \times 10^{-7}$). **g**, Representative images of seeds after 4 days of incubation with CHX, MG132 or DMSO control medium (*n* = 1). Germination was completely inhibited by CHX and partially inhibited by MG132, showing that the drug treatments were effective.



Extended Data Fig. 5 | Correlations between transcriptomes, proteomes and phosphoproteomes. **a**, Median PTRs across tissues plotted against the inter-tissue variation of these PTRs (expressed as MAD; proteins and transcripts had to be detected in at least 10 matching tissues to be included in the analysis). Arrows denote examples of genes with high PTRs (*rbcL* and *petA*) and low PTRs (*IAA8* and *IAA13*). Bar plot shows the MAD range segmented into five quantiles, each containing the same number of genes (coloured bars and dashed lines). Most genes have reasonably stable PTRs across tissues. **b**, As in **a** (dataset *n* = 14,069) but for transcript (left) and protein (right) measurements. There is more variation in protein levels across tissues than there is mRNA variation (80% of all transcripts show a MAD of <1; 80% of all proteins show a MAD of 1.2). There is also more variation in the protein levels across tissues for low abundant proteins. This may in part be due to technical limitations as low abundance proteins can generally be less accurately quantified. **c**, As in **a** but for the ratio of phosphorylation site versus protein abundance. P-sites and proteins had to be detected in at least 10 matching tissues to be included in the analysis (*n* = 13,793). **d**, As in **b** (dataset *n* = 13,793) but for P-site abundance. P-site abundance shows greater variation across tissues than protein abundance (60% of all P-sites show MAD <1 compared with 80% of all proteins; see **b**). Again, this may in part be due to technical limitations as P-site quantification is performed on a peptide level and does not benefit from aggregating multiple peptide quantifications into one value for protein quantification.

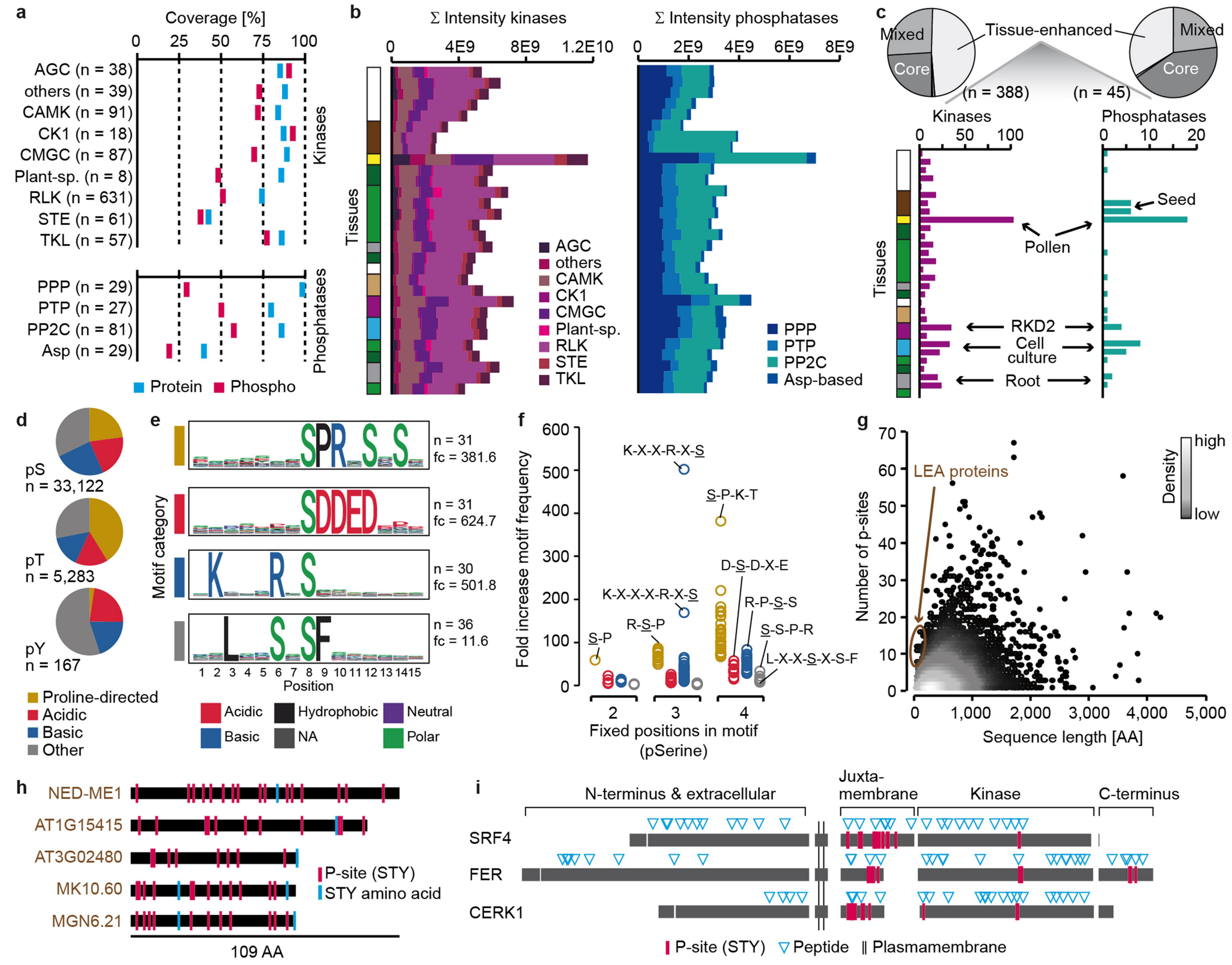


Extended Data Fig. 6 | Inferring redundant gene function and physical interactions from co-expression analysis. a, Scatter plot of Pearson's correlation coefficients (*r*) as a measure for co-expression across tissues for all pairs of proteins (*x* axis) and all pairs of transcripts (*y* axis) (core dataset only, $n = 5{,}043$) along with their marginal histograms. Colours denote the $\log_{10}$-normalized STRING scores of individual gene pairs as a measure of known or predicted direct (physical) or indirect (functional) associations. Strong co-expression of transcripts or proteins or both are more strongly related (physically or functionally) than transcripts and proteins that are not. **b**, Co-expression analysis of duplicated genes (pairs had to be detected in at least 10 matching tissues to be included in the analysis). The density plots show the distribution of Pearson's correlation coefficients (*r*) of co-expressed transcripts (grey) or proteins (blue) for genes that arose by whole-genome duplications (WGD), local duplications or transposon-mediated duplications. Randomly selected gene pairs are shown as a control. Medians are given and displayed as dotted lines. There is substantial co-expression of duplicated genes, indicating that these genes probably have redundant functions. **c**, Left, protein-level Pearson's correlation coefficient (*r*) values (from **b**) for all duplicate gene pairs (WGD, local, transposed) plotted against the protein abundance ratio of each pair (average across 30 tissues) (Methods). Blue arrows denote an example of a high or low ratio of protein production for the duplicated genes. Right, example for tissue-resolved protein intensity proportions (top-3) (Methods) for the duplicate pair *MAC5A* and *MAC5B*. Irrespective of the tissue, *MAC5A* is always much higher expressed than *MAC5B*. Tissues are coloured as in Fig. 1. **d**, Top, ranked protein abundance ratio for selected duplicate pairs (mean ± s.d.; $n = 30$) and annotated for phenotypic effects (bottom) in the loss-of-function mutant for either duplicate 1 or duplicate 2 (+). Minus symbols denotes absence of a phenotypic effect. Asymmetric protein production within duplicate pairs can be associated with the occurrence of a phenotype in the loss-of-function mutant of the higher expressed duplicate protein, indicating a dominant functional role of the more highly expressed protein. Blue arrows highlight MAC5A–MAC5B and PHB3–PHB4 as examples. **e**, Inference of physical protein–protein interactions from co-expression data. Distribution of pairwise Pearson's correlation coefficients (*r*) of co-expressed proteins across (at least 10) tissues that are subunits of selected protein complexes. $r > 0.5$ (shaded in grey) was chosen as a cut-off for the selection of proteins for subsequent analysis to make sure that proteins present in well-characterized protein complexes are retained. CONSTITUTIVE PHOTOMORPHOGENESIS9 SIGNALOSOME (CSN), CELLULOSE SYNTHASE (CESA). **f**, Recovery of annotated protein–protein interactions by co-expression analysis. Distribution of Pearson's correlation coefficients (*r*) of pairs of transcripts (grey) or protein (blue) that are annotated to interact physically in the AtPIN database[33] (pairs had to be detected in at least 10 matching tissues to be included in the analysis). Subsets of the AtPIN database, namely interactions detected by the yeast two-hybrid (Y2H) method, by affinity purification–mass spectrometry (AP–MS) or both. $r > 0.5$ are shaded in blue (protein). Dotted lines denote median values. Co-expression only recovers a minority of annotated physical interactions and interactions supported by more than one line of experimental evidence also tend to show stronger co-expression.

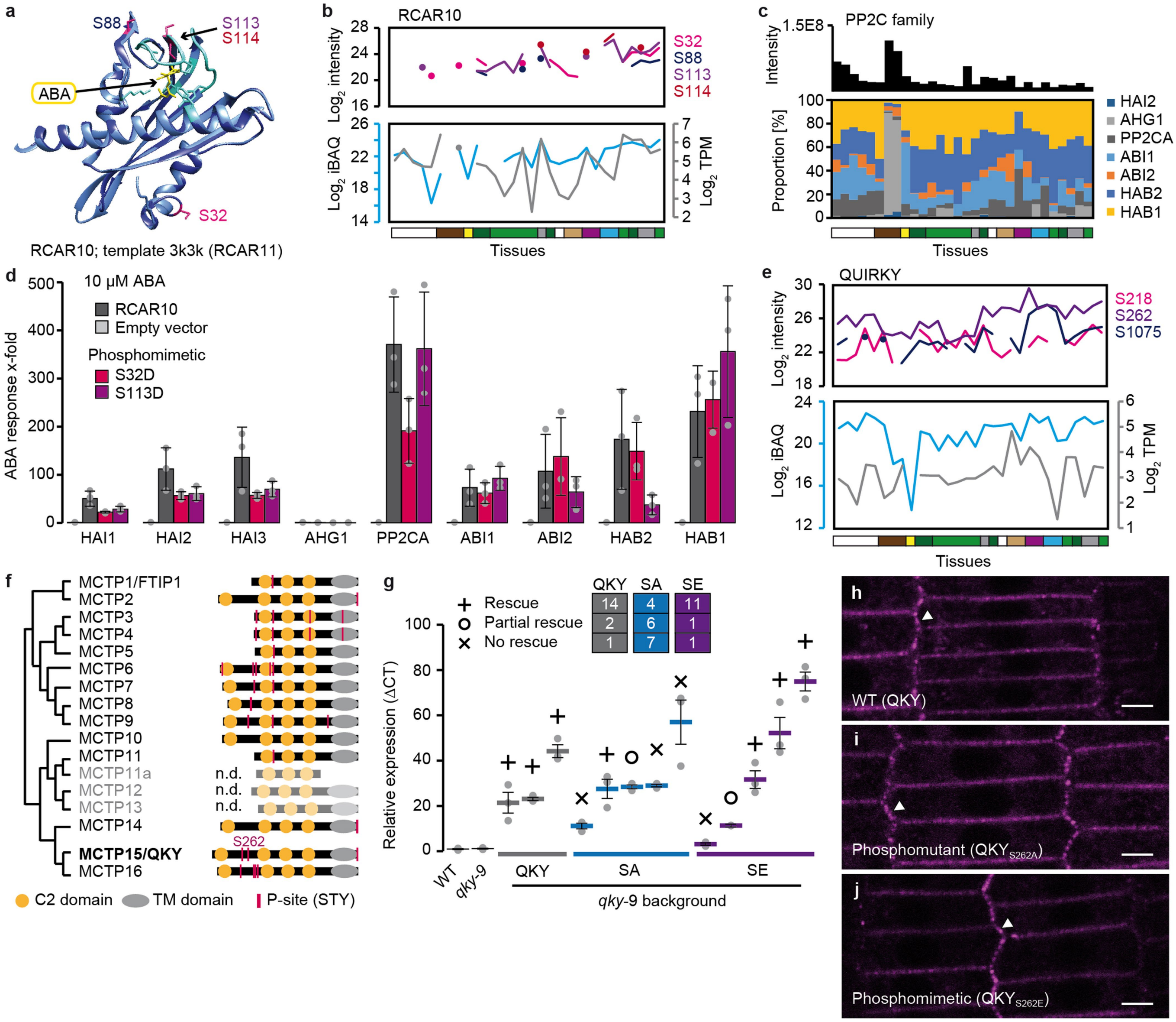


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Inferring protein complexes and subunit stoichiometry from proteome correlation profiling using SEC–MS. **a**, Molecular mass (MW) of monomeric proteins (determined from sequence) plotted against the mass determined from the apex of the elution profile for proteins identified by SEC–MS fractions of flower tissue (sFL). Inset shows the molecular mass calibration of the SEC column using a protein calibration standard (mass between 44 and 690 kDa). The distribution of proteins annotated in Araport11 is shown at the top. Many proteins show a much higher apparent molecular mass than would be expected from their sequences (data points above the $x = y$ line). This suggests that these proteins engage in physical protein interactions that are sufficiently stable during SEC separation. **b**, SEC traces of proteins from five well-characterized protein complexes for flower, leaf and root tissue. Although the resolution of SEC separations is not very high, the complex subunits show very strong co-elution behaviour and the SEC separations of the five complexes are reproducible between tissues. CoA carboxylase $n = 4$ proteins; CDC48 $n = 3$ proteins; RubisCO $n = 4$ proteins; prefoldin $n = 6$ proteins; SCS $n = 3$ proteins. **c**, Intensity-normalized SEC elution profile of proteins for flower tissue. Proteins are ordered based on the SEC fraction in which their intensity peaks and the data are displayed as a heat map ($n = 2{,}485$ protein traces). Co-eluting proteins were grouped into 'trace modules' (Methods). Proteins in trace modules may represent members of protein complexes and thus serve as candidates for further experimental validation. **d**, To quantify how well protein complexes can be detected using co-expression analysis from data in the tissue atlas (TA) or by SEC–MS, a summary statistic termed 'complex index' was calculated (Methods). The complex index is 1 when all subunits of a complex are identified in the same module and no other proteins are contained in the module. Bar plots show examples for complex indices obtained from the different datasets and are divided into large (>4 subunits) and small (≤4 subunits) protein complexes (according to UniProt). Co-expression alone generates many candidates of interactors, but combining co-expression and SEC–MS analysis is an efficient way to prioritize candidates for follow-up experiments. **e**, Subunit heterogeneity within the coatomer complex. The coatomer complex consists of seven subunits, five of which (α, β, β′, ε and ζ) can be provided by twelve paralogues of these five genes. Plots show the protein proportions of these paralogues in all 30 tissues (data from tissue atlas). The coatomer complex has a similar composition in most tissues. A notable exception is seed tissues, in which production of subunit ζ-1 dominates over the two other paralogous proteins, suggesting that the coatomer complex in seed tissue also preferentially contains the ζ-1 subunit. Tissues are coloured as in Fig. 1. **f**, Absolute SEC intensity traces of individual complex subunits for determining subunit stoichiometry. Examples from left to right: the chaperonin complex (flower, 8 proteins, ratio of all subunits: 1:1), the 26S proteasome core and lid (flower, 14+17 proteins, ratio of all subunits: 1:1), the COP9 signalosome (flower, CSN; 8 proteins, ratio of all subunits: 1:1) and the CESA1–CESA3–CESA6 complex (root, 3 proteins, ratio of all subunits: 1:1). CSN3 and CSN5 were detected both as part of the CSN complex and in monomeric form. **g**, Top, total intensity of protein complex subunits across all tissues for the complexes shown in **f** (subunit intensities from the tissue atlas). Middle, relative proportion (mean ± s.d.; $n = 30$ tissues) of subunits across tissues (Methods). For the CESA complex, ratios were calculated for the subunit combinations CESA1–CESA3–CESA6 and CESA4–CESA7–CESA8. The stoichiometries determined from the tissue expression data are generally well-aligned with the expected 1:1 ratio of subunits in these complexes. As noted in **f**, a substantial amount of CSN5 was detected as a monomer in the SEC analysis, and the tissue expression atlas also shows higher relative expression of this protein compared with all other complex partners. This suggests that the protein is produced in excess over what is required for the COP9 complex (as observed previously[123]), and may therefore indicate an additional function within the cell.



Extended Data Fig. 8 | Kinases, phosphatases and phosphorylation motifs. **a**, Percentage of annotated kinases and phosphatases family members detected at the protein or phosphoprotein level. Parentheses denote the number of genes in each family in the *Arabidopsis* genome. **b**, Tissue-resolved combined intensity (that is, protein abundance) of families of kinases (left) and phosphatases (right). Tissues are coloured as in Fig. 1. Several tissues (notably pollen) stand out in terms of the expression of kinases and phosphatases, which indicates that these tissues are particularly active in phosphorylation-mediated dynamic signalling. **c**, Top, pie chart of specificity categories for kinases and phosphatases (see Methods for definition). Bottom, distribution of tissue-enhanced kinases and phosphatases across the 30 tissues. Several tissues (such as pollen) stand out in terms of the expression of certain kinases and phosphatases, which indicates tissue-specific signalling. **d**, Pie charts showing the proportion of proline-directed, acidic, basic and other motif categories for phosphorylated Ser (pS), Thr (pT) and Tyr (pY) residues. Only class I P-sites (localization score > 0.75; Methods) were considered in this analysis. **e**, Example motif logo plots for motifs such as proline-directed, acidic and basic. P-site motifs were identified using the motif-X algorithm (see Supplementary Table 2 for all 266 motifs). *n* denotes the number of phosphorylation sites that contain the respective motif; 'fc' denotes the fold change (that is, enrichment) of the motif in phosphorylated versus unmodified peptides (Methods). **f**, Enrichment of proline-directed (yellow), acidic (red), basic (blue) and other (grey) sequence motifs (circles) in the serine P-site dataset versus the same motifs detected in the background dataset of unmodified peptides (Methods). Motifs are shown for two, three and four fixed amino acid positions. The P-site in each motif example is underlined. 'X' denotes any amino acid. **g**, Number of identified P-sites for a given protein plotted against the sequence lengths of the same protein. LEA proteins are shown. **h**, Schematic of the LEA protein sequences (black bars). Pink denotes phosphorylated and blue denotes unphosphorylated STY residues. Almost all STY residues in LEA proteins can be phosphorylated. **i**, Schematic of the sequences and domain topology of the receptor-like kinases SRF4, FER and CERK1. P-sites often preferentially occur in specific domains, notably the juxtamembrane domain. Protein sequence regions covered by identified peptides are marked in blue, and P-sites are marked in pink.



Extended Data Fig. 9 | Functional analysis of phosphorylation mutants of RCAR10 and QKY. **a**, P-site localization within the structure of RCAR10. The RCAR10 structure (blue) was modelled using the RCAR11 protein crystal structure (cornflower blue) as a template[115]. ABA-binding loops are shown in turquoise, P-sites in pink, and ABA ligand in yellow. **b**, RCAR10 expression across tissues at the protein (blue, iBAQ), transcript (grey, TPM) and P-site (pink, intensity) level. **c**, Tissue-resolved total protein intensity and relative proportions of the members of the PP2C co-receptor family. Seed tissues stand out in terms of overall expression as well as the dominance of AHG1 in these tissues. **d**, Measurement of the ABA response after expression of RCAR10 or phosphomimetic mutant variants in combination with different PP2C co-receptors in protoplasts (Methods). Columns display the average ABA response (mean ± s.d., $n = 3$) and grey dots indicate individual measurements. Co-expression of the phosphatases HAI1–HAI3 leads to similar responses in both phosphomimetic mutants, whereas other co-expressed phosphatases show diverse responses. **e**, QKY expression across tissues at the protein (blue, iBAQ), transcript (grey, TPM) and P-site (pink, intensity) level. **f**, Members of the MCTP family clustered by sequence similarity (left) and schematic of their domain structures along with detected P-sites (right). MCTP11a, MCTP12 and MCTP13 were not detected (n.d.) in this study. MCTP15 (also known as QKY) is in bold. **g**, Number of independent transgenic plant lines (*qky-9* mutant background) transformed with wild-type QKY, phosphomutant (S262A, SA; blue) or phosphomimetic (S262E, SE; purple) constructs that show complete, partial or no rescue of the mutant phenotype. qPCR results (mean ± s.d., $n = 3$; individual data points as grey dots) show the relative transgene expression in wild-type, *qky-9* mutant and selected transgenic lines. **h**–**j**, Representative confocal images of six-day-old *qky-9 pQKY::mCherry:QKY* (WT QKY; $n = 14$ roots), *qky-9 pQKY::mCherry:QKY(S262A)* (phosphomutant; $n = 21$ roots), *qky-9 pQKY::mCherry:QKY(S262E)* (phosphomimetic; $n = 15$ roots) root epidermal cells of the meristematic zone. The punctate signal along the cell circumference shows the expected localization of the QKY protein. Arrows indicate punctate structures. Scale bars, 5 μm.

nature research

Corresponding author(s): Bernhard Kuster

Last updated by author(s): Jan 13, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	MS measurements for proteomic data recording were performed on a QExactive HF mass spectrometer. RNA sequencing was performed on a HiSeq 2500 platform. MS raw data files were searched with MaxQuant software (v.1.5.8.3) against protein coding genes annotated in Araport11 (Araport11_genes.201606.pep.fasta). RNAseq reads were processed with FastQC, Trimmomatic (v. 0.35) and mapped to the Araport11 transcriptome with Kallisto (v. 0.43.1). Luciferase reporter signals for transfected protoplasts were assayed on a lumimometer (Berthold Technologies). qRT-PCR for the measurement of QKY transgene expression was performed on a Roche LightCycler480. Embryo images were analyzed with a Zeiss Axio Imager.M2 and adjusted for color and contrast using Adobe Photoshop. Floral organs were imaged on a Leica SAPO stereo microscope and images adjusted for color and contrast using ImageJ/Fiji software. Confocal laser scanning microscopy of seedling roots was performed with an Olympus FV1000 and FluoView software (FV10-ASW v. 01.04.00.09).
Data analysis	Data analysis was performed in R, Perseus (v.1.5.5.3) and Excel (Microsoft Office Professional Plus 2013). Please see the methods section for details.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data supporting the findings of this study are available within the paper, the supplementary information and the public repositories. Transcriptome sequencing and quantification data are available at ArrayExpress (www.ebi.ac.uk/arrayexpress) under the identifier E-MTAB-7978. The raw mass spectrometric data and

MaxQuant result files have been deposited to the ProteomeXchange Consortium via PRIDE, with the dataset identifier PXD013868. Source data for main Figure 1-5 and Extended Data Figure 1-9 are included.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For the tissue atlas one sample per tissue was measured. Samples were chosen in a way that the atlas contains n≥2 samples for each tissue group (except Pollen n=1). Size-exclusion chromatography was performed with a total of 3 samples (flower, leaf, root). Seed chase experiments were performed once with three time points for each condition. For the D6PK mutant embryos analysis a total of 191 embryos was screened (Col n=28; d6pkl3 n=67; d6pk012 n=52; d6pk0123 n=44). The GUS staining of plants transformed with promoter-GUS constructs under the control of AGC1.6 and AGC1.7 promoter sequences was repeated twice and GUS signal was observed for 12 out of 18 independent lines (AGC1.7) and 0 out of 20 independent lines (AGC1.6). For the luciferase response assay of protoplast suspensions treated with ABA, three biological replicates were performed. For the flower and silique phenotype analysis of plants transformed with QKY constructs, a total of 47 plants were screened (QKY WT n=17; QKY SA n=17; QKY SE n=13). Confocal imaging of seedling roots transformed with QKY constructs was performed for a total of 48 plants (QKY WT n=14; QKY SA n=21; QKY SE n=15).
Data exclusions	All data were analyzed
Replication	At least two tissues were measured for each tissue group (except Pollen n=1). Size exclusion chromatography experiments were performed with three samples (n=1). Seed chase experiments were measured once but with three time point for each treatment. The D6PK mutant embryos analysis was performed one time with a total of 191 embryos. The GUS staining of plants transformed with promotor-GUS constructs under the control of AGC1.6 and AGC1.7 promoter sequences was repeated twice. For the luciferase response assay of protoplast suspensions treated with ABA, three biological replicates were performed. The flower and silique phenotype analysis of plants transformed with QKY constructs was performed one time with a total of 47 independent plant lines. Confocal imaging of seedling roots transformed with QKY constructs was performed twice for a total of 48 plants.
Randomization	Tissue atlas samples for mass spectrometry and RNAseq measurements were prepared and measured in random order. MS measurement of seed chase samples was performed in random order.
Blinding	Blinding was not performed in this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging