nature plants

Brief Communication

https://doi.org/10.1038/s41477-024-01738-4

Coexpression enhances cross-species integration of single-cell RNA sequencing across diverse plant species

Received: 9 May 2023

Accepted: 3 June 2024

Published online: 27 June 2024

Check for updates

Michael John Passalacqua [1] & Jesse Gillis [1,2]

Single-cell RNA sequencing is increasingly used to investigate cross-species differences driven by gene expression and cell-type composition in plants. However, the frequent expansion of plant gene families due to whole-genome duplications makes identification of one-to-one orthologues difficult, complicating integration. Here we demonstrate that coexpression can be used to trim many-to-many orthology families down to identify one-to-one gene pairs with proxy expression profiles, improving the performance of traditional integration methods and reducing barriers to integration across a diverse array of plant species.

Plants have a remarkably flexible cellular physiology, driving their adaptation into nearly every environment. Recently, the advent of single-cell RNA sequencing (scRNA-seq) has provided novel insights into the diversity of cell types underlying these adaptations[1,2]. The unique diversity in plants makes comparative assessments between species important but is complicated by uncertain homology relationships. Unlike in mammals, where homologous genes and structures can be easily identified, plant gene families frequently expand by whole-genome duplication, polyploidization and tandem gene duplication[3–5]. This scarcity of one-to-one gene pairs is a major barrier to defining a common gene space for the integration of single-cell data, a key step for successful cross-species comparative analysis or integration[6,7]. With vast amounts of plant scRNA-seq data becoming available[8], this study aims to address a critical gap in its analysis by using coexpression to identify pairs of genes that, while not exclusive orthologues, are functionally related enough to enable the integration of this high-dimensional data. By reducing barriers to integration, we prime the field for the discovery of novel, cell-type specific innovations that have been critical to plant adaptation and domestication.

While a given plant sample may have thousands of expressed genes, the expression patterns of these genes are not independent and are instead organized into the regulatory programs that underlie cell types. This coexpression generates the low-dimensional expression space that is foundational to the success of modern single-cell analysis[9]. We hypothesize that genes with highly similar expression profiles between two species can be used as reasonable proxies for integrating cell-type specific data, that we can identify such profiles using coexpression and that this will expand the shared gene space, improving our ability to compare cross-species data. The essence of the approach is to use meta-analysis from previous bulk RNA sequencing data to define cross-species gene pairs (coexpression proxies) that can be applied to more specific, but sparser, single-cell data. By utilizing robust coexpression networks built from over 16,000 publicly available RNA sequencing datasets, as well as gene phylogenies from OrthoDB v11 (a database of precomputed gene orthology relationships), we ensure that the coexpression proxies accurately reflect the underlying biology of each species pair they are drawn from[10,11]. We illustrate this approach, where coexpression data and gene phylogenies identify gene pairs that expand the one-to-one (1–1) gene space, improving data integration and alignment between known cell types and highlighting novel ones between species (Fig. 1a). While previous work has expanded the shared gene space through gene homology comparisons, our focus on coexpression uniquely captures both regulatory and functional shifts between species[12]. By improving integration, we enable researchers to identify new and conserved cell types in their scRNA-seq data. We validate the coexpression proxies with two test examples, highlighting their utility. In the first test, we show that coexpression proxies can accurately reintegrate a split dataset with no shared gene space. Second, we show that coexpression proxies improve the integration of real single-cell data between two species with complex genomes: maize and rice.

Our first test is an extreme one in which we generate and integrate two cross-species datasets with no one-to-one orthologues. This would

[1]Genomics Department, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. [2]Physiology Department and Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada. e-mail: passala@cshl.edu; jesse.gillis@utoronto.ca

https://doi.org/10.1038/s41477-024-01738-4

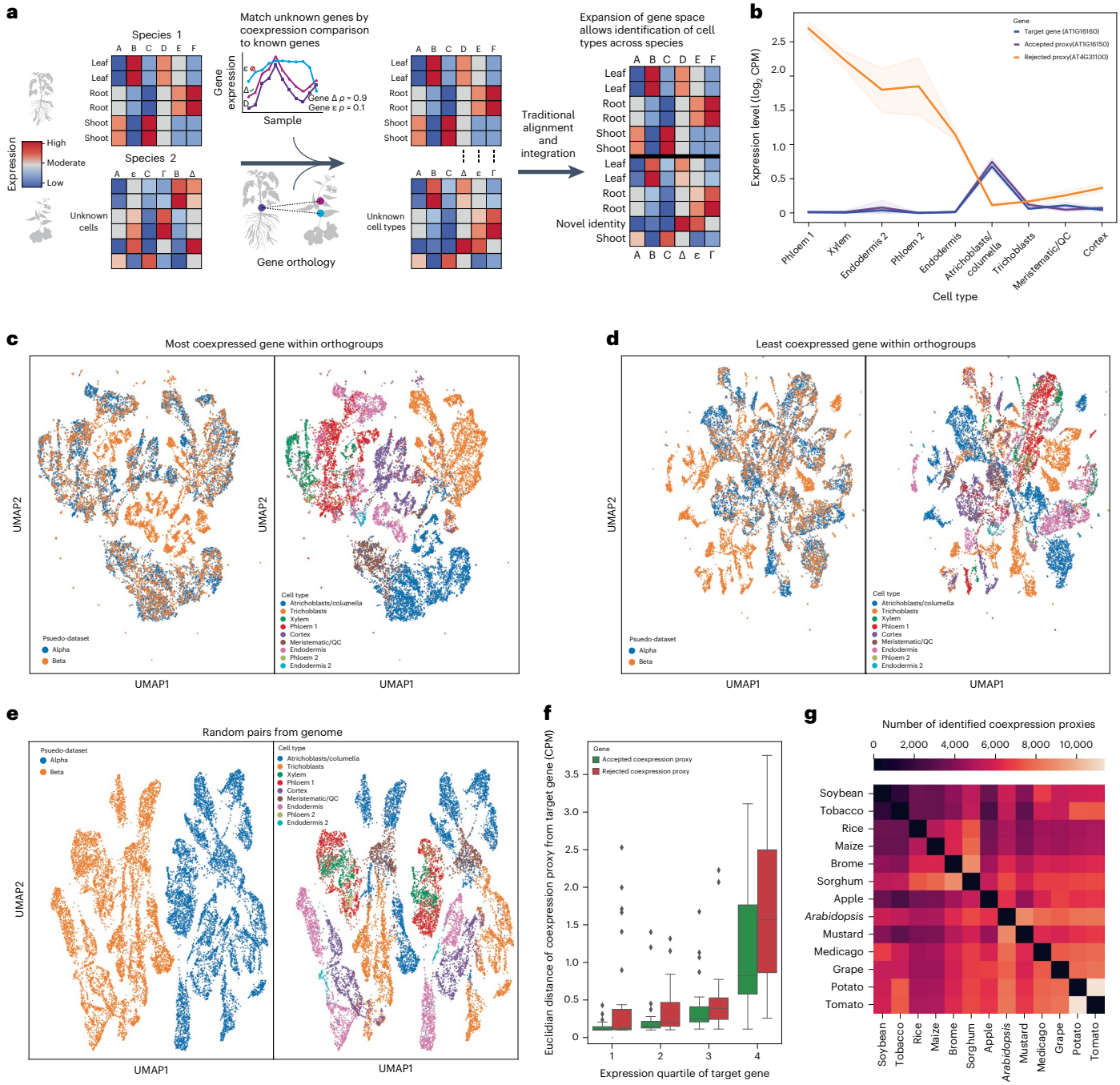

Fig. 1 | Coexpression proxies integrate a split dataset without shared genes. **a**, Schematic depicting the identification of coexpression proxies from gene orthology information and their use in expanding the gene space to enable integration followed by identification of novel and conserved cell types. **b**, Gene expression profile for target gene (*AT1G16160*) and two potential coexpression proxies (*AT1G16150*, *AT4G31100*). The gene with the more similar profile, *AT1G16150*, was identified as a coexpression proxy, while *AT4G31100* was rejected. The centre band is the mean counts per million (CPM) for each gene in the cell type in our single-cell dataset. The error bar is the 95% confidence interval. QC, quiescent center. **c**, UMAP showing integration of a split and dissociated *A. thaliana* dataset containing 16,636 cells using coexpression proxies. **d**, UMAP showing integration of the same dataset using the worst potential coexpression proxy from each gene family. **e**, UMAP showing the failed integration of the split and dissociated dataset using 1,900 random gene pairs. **f**, Euclidian distance from the expression profile of the target gene for $n = 117$ pairs of accepted coexpression proxies and rejected coexpression proxies in independent cell types, split by expression quartile of the target gene. The bottom of the box is the lower quartile, the top of the box is the upper quartile and the centre bar is the median. The whiskers are 1.5 times the interquartile range. **g**, Heat map showing the number of identified coexpression proxies between each species pair in the database.

be impossible with a traditional integration approach, which requires directly matched one-to-one orthology relationships between genes in each species for alignment before integration. To construct a case with a ground truth integration without using synthetic data, we split an existing *Arabidopsis* single-cell dataset into two pseudo-'species'. The first 'species' is generated by randomly selecting half of the cells as well as half the genome. For these cells, the second half of the genome is removed. We then take the remaining cells, which will become our second 'species', and remove the half of the genome present in the first set of cells (Supplementary Fig. 1). This provides two sets of cells with known,

https://doi.org/10.1038/s41477-024-01738-4

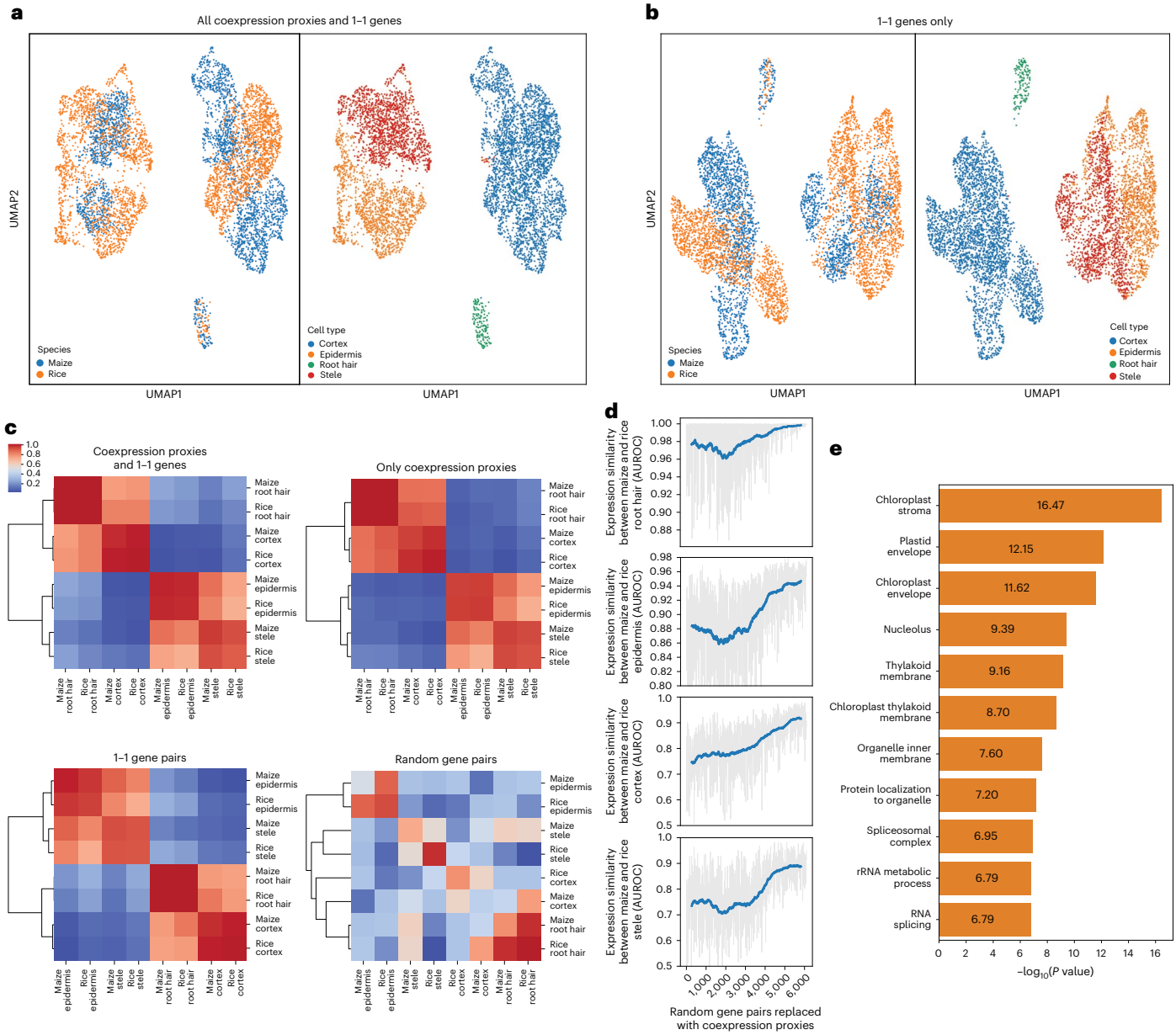

Fig. 2 | Integration of maize and rice scRNA-seq data using coexpression proxies. a, UMAP showing integration of 2,832 *Z. mays* cells and 3,500 *Oryza sativa* cells using coexpression proxies. **b**, UMAP showing integration of *Z. mays* and *O. sativa* using only 1–1 gene pairs from OrthoDB. **c**, MetaNeighbor plots showing post integration similarity between cell types using four different gene sets. **d**, Improvement in integration across 872 integration runs as random gene pairs are gradually swapped for coexpression proxies. **e**, Enriched Gene Ontology terms among rice–maize coexpression proxies. Enrichment was tested by a two-sided Fisher's exact test. It was then corrected using the Benjamini–Hochberg correction.

shared cell types and distinct genomes. We then identify coexpression proxies between the two subset genomes, finding pairs of genes with similar expression profiles. As an example, the selected coexpression proxy gene, *AT1G16150*, closely matches the expression profile of the target gene, *AT1G16160*. By contrast, *AT4G31100*, a rejected gene from the same orthologue family, has a distinct expression profile (Fig. 1b).

Next, we used these coexpression proxies to reintegrate the split *Arabidopsis* dataset. Highlighting that coexpression proxies smoothly integrate into existing workflows, we used Scanorama v1.7.1[13] to reintegrate and re-cluster the dataset, placing 82% of cells into a cluster with cells from both datasets (Fig. 1c). The reintegration was accurate, successfully matching cells of the same cell type across datasets 75% of the time. To evaluate how much of the gene proxies' success was dependent on information from the gene phylogenies and how much information was derived from the coexpression conservation profile, we attempted to integrate the datasets using the worst rejected proxy from within each orthologue group (that is, the proxy with the lowest coexpression). Performance was lower using these gene pairs, reducing the successful matching of cells to 65% (Fig. 1d). This moderate performance suggests that simple relaxation of orthology constraints is a substantial contributor to performance. However, coexpression provides a substantial overall signal boost. This was particularly clear for phloem, which was otherwise unintegrated or mixed with atrichoblasts and xylem. To determine whether sequence similarity alone would prove sufficient, we calculated the pairwise protein sequence similarity of every *Arabidopsis* gene and attempted to use this to identify gene proxies. While able to perform better than random, this metric was worse than coexpression at reintegrating the split dataset and completely failed to reintegrate certain clusters. Finally, we attempted integration using 1,900 random gene pairs and found that we were unable to

achieve any integration (Fig. 1e). To further evaluate our coexpression proxies, we assessed the degree to which rejected and selected gene pairs show the same expression across cell types on a per-gene basis (measured by Euclidean distance). We found that accepted coexpression proxies are much closer to the target's expression profile across cell types and that the rejected proxies are on average 83% further from the target's expression (Fig. 1f). This shows that the coexpression proxies are more similar in expression profile to their target genes than even other genes from the same orthogroup.

Given the success of our approach, we generated coexpression proxies between 13 plant species and identified an average of 5,750 gene pairs between species (Fig. 1g). The coexpression proxies are numerous enough to provide additional information across even highly diverged species and are well represented (4,899 pairs) even between *Zea mays* and *Arabidopsis thaliana*, which diverged 160 Ma. Importantly, although we used Scanorama, these coexpression proxies can be easily incorporated into any potential integration pipeline as they simply expand the shared feature space.

Having shown that coexpression proxies could integrate an otherwise uncorrectable dataset, we tested their ability to improve the integration of single-cell data across two different species. Using a supervised integration, we attempted the integration of two root datasets, one from maize and one from rice. We focused on four broad cell types for which author annotations directly aligned. Using coexpression proxies, we successfully integrated the maize and rice dataset, accurately integrating 36% of cells into clusters with cells from both datasets (Fig. 2a and Supplementary Fig. 2). The remaining cells were different enough to still appear as distinct sub-clusters across species. While this is far from 100%, real cross-species differences do exist, so it is not clear what the maximum plausible integration percentage is. Importantly, our integration is better than using only the 1–1 gene pairs, which integrated only 14% of the cells (Fig. 2b). Key cell types, such as epidermis and stele, are well integrated using coexpression and are less well integrated by 1–1 gene pairs, as evidenced by lack of species mixing within cell types and close proximity across cell types. Similarly, coexpression did not overfit away real differences, capturing the likely real difference between cortex cells where constitutive aerenchyma formation is critical to oxygen diffusion in partially submerged rice[14]. To evaluate the integration on a cell-type-by-cell-type basis, we used MetaNeighbor v3.19, which enables us to quantify the degree to which cell types replicate across datasets in a statistical framework[15,16]. We compare four integrations using scGen—utilizing coexpression proxies and 1–1 genes, using only coexpression proxies, using only 1–1 genes and using random genes (Fig. 2c). As can be seen, coexpression proxies alone, 1–1 pairs alone and the combination all accurately and similarly group cell types across species. While subtle for this broad classification, the full coexpression proxy set integrates better than either of its parts in all cell types when evaluated by MetaNeighbor (except cortex, where all methods are perfect), reflecting the additional information from the coexpression proxies. Because this is a validation focused on well-defined alignment, performances generally go from high to even higher (for example, stele goes from AUROC (area under the receiver operator curve) 0.93 to 0.973). To evaluate the utility of an increased known gene-pair space, as well as the robustness of the model, we swapped in coexpression proxies for random pairs and tracked performance improvement (Fig. 2d). Performance increases steadily to near 1 for most cell types, indicating that the typical number of 5,000 coexpression proxies is sufficient to integrate cross-species data. Further querying the coexpression proxies, we found they typically represented core conserved functions such as photosynthesis, mitochondrial proteins and ribosome metabolism (Fig. 2e).

Integrating cross-species single-cell data is an increasingly common goal in the fields of plant development, evolution and molecular biology. To facilitate this process, we have demonstrated that using coexpression proxies expands the gene space available for integration. To facilitate adoption of this approach by the community, we have generated pairwise coexpression proxies between 13 plant species at 3 thresholds. All coexpression proxy lists have been made available at https://gillislab.shinyapps.io/epiphites_v11/. In addition, we have provided a workflow for generating a coexpression network from scRNA-seq data and using it to identify coexpression proxies for integration (Supplementary Code), which additionally requires only gene phylogenies between the two species. We show that this approach generates networks similar to gold standard networks and enables similar integration (Supplementary Figs. 3 and 4). These proxy lists provide an important resource for improving the integration of single-cell data, accelerating the transfer of knowledge from well-studied model organisms to crop systems that are crucial to the global food supply.

Methods

Gene coexpression proxy identification

For each species pair, gene family orthology information was downloaded from OrthoDB V11[11]. Utilizing one-to-one gene pairs, coexpression conservation was calculated between all genes in each species[17]. Briefly, we compare each gene's top 10 coexpression partners across species. These top 10 are limited to genes that are one-to-one orthologues, although the matching of proxies is not limited in this way. Using one-to-ones as a basis set for comparison of other genes expands the range of potential proxies while still leaving it grounded in defined cross-species overlaps. We use the ranks from one species to predict the coexpression partners of the second species and then repeat this in the other direction, averaging the scores to generate the conservation of coexpression score, which is an AUROC. This resulted in a species A genes by species B genes matrix, filled with the AUROC score for each gene pair. For each gene family, the coexpression conservation matrix was filtered to every possible cross-species gene pair. Next, pairs in multigene groups were eliminated by thresholding in two steps. First, any gene pairs with scores below a quality threshold were discarded. Second, remaining pairs were required to be reciprocal best hits and to be higher than other potential options by a multi-pair threshold. For genes that were one-to-one matches, they were only discarded if below a lower single pair quality threshold. For the moderate filtering, the quality, multi-pair and single-pair junk thresholds were 0.85, 0.03 and 0.8. For lenient filtering, the thresholds were 0.8, 0.02 and 0.7, and for stringent filtering, these were 0.9, 0.035 and 0.85. The moderate threshold was chosen by evaluating the number of proxies identified at many thresholds and choosing the elbow, and lenient and stringent thresholds were picked to form a 0.1 range around this number.

Dataset integration and evaluation

To generate an integration task that was uncorrectable without a shared gene space, the *Arabidopsis* dataset was split into two sets of cells. Using Pandas DataFrame.sample, one half of the genome was randomly selected and assigned to the first set of cells, with other data being discarded. The second set of cells were assigned the second half of the genome, and the genes assigned to the first half were discarded. Utilizing the same method as above, coexpression proxies were identified between the two halves of the genome with the moderate threshold. Aligning the two gene spaces using these proxies, we performed integration using the Scanorama v1.7.1 Python package[13]. The scanorama.integrate function was used to integrate the two datasets into a shared low-dimensional space, and this was plotted using scanpy.pp.neighbors with the default parameters (15 nearest neighbours, 50 principal components) and the scanpy.tl.umap function (default parameters). For evaluation, we first clustered the integrated data using scanpy.tl.leiden at a resolution of 0.5. This provided an evaluation space that is based on the high-dimensional underlying data, instead of the 2D uniform manifold approximation and projection (UMAP), which can be misleading. Then, using these clusters, we defined a cluster of the same cell type

as one containing more than 60% of that cell type and a mixed cluster as one composed of between 30% and 70% of each starting dataset. In plotted boxplots, the centre line is the median, the box limits are the upper and lower quartiles, the whiskers are 1.5 times the interquartile range and the points are any datapoints beyond the whisker range.

As the cross-species integration scenario was more challenging, it was integrated utilizing scGEN v2.1.0[18]. The datasets were limited to 4 broad cell types for which author annotations clearly aligned, and the rice dataset was subset to 3,500 cells to match the maize dataset of 2,832 cells. The tissues were equally represented in each of the two datasets. Utilizing coexpression proxies between rice and maize at the moderate threshold, the two datasets were aligned. The two datasets were first aligned utilizing coexpression proxies between rice and maize at the moderate threshold. Next, the scGEN model was initialized using scgen.SCGEN and trained using scgen.model.train, using default parameters. Next, the integration was performed using scgen.model.batch_removal. To evaluate the integration beyond the low-dimensional representation, MetaNeighbor was used to compare the post-integration similarity of cell types[15]. To confirm the model was utilizing coexpression proxies and not relying on training information, the integration was run 872 times, starting with random gene pairs. Following each run, seven random pairs were replaced with seven coexpression proxies, until all were replaced. Gene Ontology term enrichment was performed using Fisher's exact test from scipy.stats.fisher_exact() to find terms over-represented in the coexpression proxies, utilizing all genes in the bulk network as the background gene set. Multiple hypothesis correction was performed using the Benjamini–Hochberg correction function from statsmodels.stats.multitest.multipletests() at alpha .05.

Reporting summary

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

Data availability

All analyses were performed in Python v3.9, Pandas v1.5, SCGEN v2.1.0, Statsmodels v0.14.1, Scipy v1.12.0, Scanorama v1.7.1 and SCANPY v1.9.1[19]. Aggregate coexpression networks were downloaded from CoCoCoNet[10]. *A. thaliana* single-cell RNA-seq expression data totalling 16,636 cells from 4 datasets were downloaded from the Gene Expression Omnibus (GEO IDs: GSE116614, GSE121619, GSE123818, GSE123013)[1,2,20,21]. Cluster assignments were downloaded from GEO for IDs GSE121619 and GSE123013 or provided by the authors for IDs GSE123981 and GSE116614. *O. sativa* single-cell RNA-seq expression data and accompanying cluster assignments were downloaded from GSE146035[22]. *Z. mays* single-cell RNA-seq expression data and accompanying cluster assignments were downloaded from GSE183171[23], and only nitrate-treated cells were used. Orthology information is from OrthoDB V11 (https://www.orthodb.org/).

Code availability

All code is available via our repository at https://github.com/gillislab/Coexpression_Proxies.

References

1. Denyer, T. et al. Spatiotemporal developmental trajectories in the *Arabidopsis* root revealed using high-throughput single-cell RNA sequencing. *Dev. Cell* **48**, 840–852.e5 (2019).
2. Ryu, K. H., Huang, L., Kang, H. M. & Schiefelbein, J. Single-cell RNA sequencing resolves molecular relationships among individual plant cells. *Plant Physiol.* **179**, 1444–1456 (2019).
3. Su, A. I. et al. Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA* **99**, 4465–4470 (2002).
4. Gharib, W. H. & Robinson-Rechavi, M. When orthologs diverge between human and mouse. *Brief. Bioinform.* **12**, 436–441 (2011).
5. Clark, J. W. & Donoghue, P. C. J. Whole-genome duplication and plant macroevolution. *Trends Plant Sci.* **23**, 933–945 (2018).
6. Bennetzen, J. L. Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. *Plant Cell* **12**, 1021–1029 (2000).
7. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
8. Chen, H. et al. PlantscRNAdb: a database for plant single-cell RNA analysis. *Mol. Plant* **14**, 855–857 (2021).
9. Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Syst.* **2**, 239–250 (2016).
10. Lee, J., Shah, M., Ballouz, S., Crow, M. & Gillis, J. CoCoCoNet: conserved and comparative co-expression across a diverse set of species. *Nucleic Acids Res.* **48**, W566–W571 (2020).
11. Kriventseva, E. V. et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2019).
12. Tarashansky, A. J. et al. Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *eLife* **10**, e66747 (2021).
13. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
14. Colmer, T. D. & Pedersen, O. Oxygen dynamics in submerged rice (*Oryza sativa*). *New Phytol.* **178**, 326–334 (2008).
15. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* **9**, 884 (2018).
16. Fischer, S., Crow, M., Harris, B. D. & Gillis, J. Scaling up reproducible research for single-cell transcriptomics using MetaNeighbor. *Nat. Protoc.* **16**, 4031–4067 (2021).
17. Crow, M., Suresh, H., Lee, J. & Gillis, J. Coexpression reveals conserved gene programs that co-vary with cell type across kingdoms. *Nucleic Acids Res.* **50**, 4302–4314 (2022).
18. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).
19. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
20. Shulse, C. N. et al. High-throughput single-cell transcriptome profiling of plant cell types. *Cell Rep.* **27**, 2241–2247.e4 (2019).
21. Jean-Baptiste, K. et al. dynamics of gene expression in single root cells of *Arabidopsis thaliana*. *Plant Cell* **31**, 993–1011 (2019).
22. Liu, Q. et al. Transcriptional landscape of rice roots at the single-cell resolution. *Mol. Plant* **14**, 384–394 (2021).
23. Li, X. et al. Single-cell RNA sequencing reveals the landscape of maize root tips and assists in identification of cell type-specific nitrate-response genes. *Crop J.* **10**, 1589–1600 (2022).

Acknowledgements

We thank K. Birnbaum and B. Guillotin for their discussion on cortex in monocots and their feedback on the manuscript. We thank D. Jackson and X. Xu for their inspiration of the project and their feedback on the manuscript. We thank J. Hover for his support in setting up the website. We acknowledge funding support from National Science Foundation IOS-1934388 and National Institutes of Health R01 MH113005. We acknowledge support from William Randolph Hearst Foundation and the Cold Spring Harbor Laboratory School for Biological Sciences.

Author contributions

M.J.P. and J.G. conceived and designed the project and test cases, analysed data, wrote the manuscript and reviewed and edited the

manuscript. M.P. performed experiments and produced the project website.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41477-024-01738-4.

Correspondence and requests for materials should be addressed to Michael John Passalacqua or Jesse Gillis.

Peer review information *Nature Plants* thanks Xuwu Sun and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2024

nature portfolio

Corresponding author(s): Jesse Gillis

Last updated by author(s): 5/3/2024

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies 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	No software was used for data collection
Data analysis	All code used is available on our repository at https://github.com/gillislab/Coexpression_Proxies All analyses were performed in Python 3.9, Pandas 1.5,SCGEN 2.1.0, Statsmodels 0.14.1, Scipy 1.12.0, and SCANPY version 1.9.117

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Aggregate coexpression networks were downloaded from CoCoCoNet10. Arabidopsis thaliana single-cell RNA-seq expression data totaling 16636 cells from 4 datasets were downloaded from the Gene Expression Omnibus GEO IDs: GSE116614, GSE121619, GSE123818, GSE123013)1,2,18,19. Cluster assignments were downloaded from GEO for IDs GSE121619 and GSE123013, or provided by the authors for IDs GSE123981 and GSE116614. Oryza sativa single-cell RNA-seq

expression data and accompanying cluster assignments were downloaded from GSE14603520. Zea mays single-cell RNA-seq expression data and accompanying cluster assignments were downloaded from GSE18317121, and only nitrate treated cells were used. Orthology information is from OrthoDBv11 - https://www.orthodb.org/

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	No human related research has been used in this manuscript
Reporting on race, ethnicity, or other socially relevant groupings	No human related research has been used in this manuscript
Population characteristics	No human related research has been used in this manuscript
Recruitment	No human related research has been used in this manuscript
Ethics oversight	No human related research has been used in this manuscript

Note that full information on the approval of the study protocol must also be provided in the manuscript.

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	No sample size calculation was performed. We utilized available, existing scRNA sequencing plant data sets for testing, using all cells available for integration - which is how our sample size was chosen. We show that sufficient genes are used in Figure 2D
Data exclusions	Cell types without clear matches between maize and rice were excluded from the maize rice integration. These were predetermined for exclusion based on annotation.
Replication	All experiments were repeated with multiple random seeds to confirm robustness and confirm initialization did not impact results. All replication was successful. All experiments with non-deterministic code were repeated at minimum of 3 times. The maize rice integration was also repeated 872 times to confirm robustness of results as the number of available genes changed.
Randomization	Arabidopsis cells were randomly assigned to each psuedo data set. No other randomization was relevant to the study as there were no direct experiments.
Blinding	Blinding was not possible as there was only evaluation of existing data to identify co-expression proxies. Existing data had known ground truth cell types, and we only evaluated how well we could restore this known ground truth. There was nothing possible for us to blind.

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 and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☐	☒ Plants

Methods

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

Plants

Seed stocks	No seed stocks were used, all data was from publicly available data sets
Novel plant genotypes	No new genotypes were used, all data was from publicly available data sets
Authentication	No authentication was performed, as no new data was generated