

Single-cell RNA-seq of *Lotus japonicus* provide insights into identification and function of root cell types of legume

Zhanmin Sun^{1†}, Sanjie Jiang^{2†}, Dan Wang^{3†}, Linxia Li¹, Boxin Liu², Qifan Ran⁴, Lizhen Hu⁵, Junbo Xiong⁶, Yixiong Tang¹, Xiaofeng Gu¹, Yanmin Wu^{1*}, Zhe Liang^{1*}

SUPPORTING INFORMATION

Material and Methods

L. japonicus growth conditions and protoplast isolation

L. japonicus MG-20 seeds were sterilized in 2% sodium hypochlorite for 10 min, washed 3–5 times using sterile water, and germinated on filter paper dampened with Murashige–Skoog (MS) liquid medium (Phyto Technology, USA), and vertically cultured in a growth chamber with 16 h of light and 8 h of darkness. Subsequently, 500 primary root tips (5 mm) were harvested and cut into small (1 mm) pieces, then incubated in digest solution (2% cellulase RS (Yakult, Japan), 0.4% macerozyme R10 (Yakult, Japan), 1% hemicellulase (Sigma, USA), 0.5% pectinase Y-23 (Yakult, Japan), 20 mM MES (2-(N-morpholino)ethanesulfonic acid) (Sigma, USA), 0.4 M mannitol (Sigma, USA), 10 mM CaCl₂ (Sigma, USA)) under vacuum (-60 kPa) at room temperature (RT) for 30 min. The mixture in a 50 ml tube was then spun at 60–70 rpm for 1 h to release root cell protoplasts. The protoplasts were then filtered through 80-mm nylon mesh and centrifuged at 130× g for 5 min, and

washed three times with 8% mannitol at RT. After discarding the supernatant, a small amount of 8% mannitol solution was added to resuspend the protoplasts. Trypan Blue (Sangon Biotech, China) staining was applied to confirm the quality (< 10% stained) of the protoplasts.

scRNA-seq library preparation and sequencing

The isolated protoplasts were adjusted to 1000–1200 cells/ml in 8% mannitol solution and then processed through the 10× Chromium 3' Single Cell Platform according to the manufacturer's instructions (10× Genomics). Briefly, 22,688 cells in one batch were loaded onto a single-cell chip for library construction using the Chromium Single Cell 30 Reagent Kit V3 (10× Genomics, USA). The protoplasts were partitioned into single-cell GEMs (gel beads in emulsion) followed by cell lysis and barcoded reverse transcription of RNA in the droplets. The quality of each cDNA library was analyzed using an Agilent 2100 Bioanalyzer, and sequencing was performed on the MGISEQ-2000 platform (MGI, China).

Cell-type clustering and identification of cell-type marker genes

The scRNA-seq data were aligned to our assembled *L. japonicus* reference genome, and counted using the Cell Ranger pipelines (version 6.0, 10x Genomics). The raw count matrix data were imported into R using the Seurat (V4.1.0) package for further data analysis (Butler et al., 2018). Cells with fewer than 200 detected genes and genes detected in fewer than 20 cells were removed from further analysis. Each individual dataset was scaled and normalized using the ScaleData and NormalizeData functions. The top 2000 genes with the most highly variable expression were subsequently used for principal component analysis (PCA) dimensionality reduction. The first 19 principal components (PCs) were selected according to the PCA elbow plot and used for clustering with resolution parameter 0.3. The clusters were visualized and explored using *t*-SNE (van der Maaten & Hinton, 2008) and

UMAP (Becht et al., 2018). Marker genes (cluster-enriched genes) were identified using the function FindAllMarkers in the Seurat package. Differential expression analysis was performed using the Wilcoxon Rank Sum test. The cluster-enriched genes were detected under the parameters of logfc.threshold=0.5. The final cluster annotation was validated using *in situ* hybridization experiments.

Bulk RNA-seq and analysis

Bulk RNA-seq was performed using RNA independently isolated from 5-day-old root tip (5 mm) cells or protoplasts from primary roots of *L. japonicus*, respectively. RNA-Seq was performed as previously described (Liang et al., 2018). Briefly, total RNA was extracted using the RNeasy Plus Mini Kit (Qiagen). The amount and quality of RNA were analyzed using a Qubit RNA Assay Kit (Life Technologies), by gel electrophoresis, and using an Agilent Bioanalyzer 2100 system. Oligo(dT)-linked magnetic beads were used to purify mRNA. The amplified library was sequenced on the MGISEQ 2000 platform. Sequence reads were trimmed using SOAPnuke (version 1.5.6) (Chen et al., 2018) and aligned to the *L. japonicus* reference genomes with HISAT2 (version 2.0.4) (Kim et al., 2019). Gene expression values were calculated on uniquely mapped reads using HTSeq (version 0.11.2) (Anders et al., 2015), and DEseq2 (version 1.4.5) was used to calculate differential expression (fold change >2 and $P < 0.01$) (Love et al., 2014).

Toluidine Blue staining

Five-day-old root tips (5 mm) of primary roots were excised from *L. japonicus* plants and then fixed in a formaldehyde-acetic acid solution (50% ethanol, 5% acetic acid, and 3.7% formaldehyde) (Sangon, China). The fixed tissue was dehydrated through an ethanol series and HistoClear. The samples were embedded in paraffin (Sigma, USA), cut into 10-μm sections with a thickness of 10 μm on Rotary microtome (Leica RM2255, Germany), and mounted on

microscope slides (Fisher Scientific). Sections of primary roots from *L. japonicus* were stained in a solution of 0.01% Toluidine Blue (Sangon, China). Paraffin was first removed from sections by dissolving in xylene and the sections were then rehydrated in distilled water. After staining in Toluidine Blue solution for 90 s, sections were washed in distilled water and dehydrated quickly through an ethanol series, and then covered with resinous mounting medium (Sigma, USA). After staining and mounting, the samples were sealed with cover slips and observed under a light microscope (Leica DM6B).

RNA *in situ* hybridization

Specific regions of marker genes were cloned into the pGEM-TEasy (Promega) vector, then Digoxigenin RNA Labeling Kit (Roche) was used for *in vitro* transcription and labeling. Hybridization and immunological detection were performed as described previously (Zhao et al., 2009; Liang et al., 2015). Bright-field microscopy was carried out using a Leica DM6B. The primers used in these experiments are listed in Table S25.

Pseudotime analysis

The pseudotime analysis of cell differentiation and determination of cell fate was performed using the Monocle 2 (v.2.10.0). To explore the developmental trajectory of specific cell types, a subset of raw data with target clusters was analyzed as follows. Firstly, the variance in expression of each gene across cells was calculated using the dispersionTable function. Genes with variable expression were chosen based on average expression level to define a developmental trajectory. Second, we reduced the dimensionality of the data to two (set max_components=2, method=DDRTree). With the expression data in a lower dimensional space, the state transition of single cells was then described using orderCells. The trajectory was plotted with plot_cell_trajectory in Monocle 2. To specify *a priori* the beginning of the trajectory, we ran orderCells again to set the root_state argument. Genes dynamically expressed

along the pseudotime trajectory were clustered and visualized using the `plot_pseudotime_heatmap` function. To determine genes that contribute to the branching of developmental trajectory, branch points were chosen. Then, the `differentialGeneTest` function was used to analyze and identify the pseudotime-dependent or branch-dependent genes. The genes that were significantly branch-dependent were visualized using the `plot_genes_branched_heatmap` function. To explore the differences in root hair cells between branches, the Hypergeometric Distribution test in R was used to test the significance of enrichment.

Comparative analysis of root scRNA-seq data from *L. japonicus* and *Arabidopsis*

Arabidopsis root scRNA-seq data were obtained from Zhang et al. (2019b). OrthoFinder was used to calculate pairwise similarity scores between genes from *Arabidopsis* and *L. japonicus* for constructing orthology groups. Finally, a total of 4743 one-to-one orthologous gene pairs were identified and retained for subsequent analysis. An intersection set of the top 1000 genes with the most highly variable expression from each dataset that are expressed in all samples was identified using the `FindIntegrationAnchors` function with other default parameters (`dims=1:20`) to discover integration anchors across all samples. The `IntegrateData` function was run on the anchor set with default additional arguments. `ScaleData` and `RunPCA` were then performed on the integrated assay to compute 20 principal components. UMAP dimensionality reduction was carried out and an SNN graph constructed using dimensions 1:17 as input features and default PCA reduction. Clustering was performed on the integrated assay at a resolution of 0.5. Reference-based integration across samples was performed using the Seurat package. In the Reference-based method, a subset of the scRNA-seq dataset (or a single dataset) is listed as a reference. This method provides advantages in both efficiency and in finding anchors. *Arabidopsis* scRNA-Seq results from an independent study (Zhang et

al., 2019) were used as the reference. The functions FindIntegrationAnchors and IntegrateData were used to integrate and combine the datasets before scaling, PCA, and clustering. Average expression values of each cluster were log2-transformed and Pearson's correlations were calculated to compare *L. japonicus* and *Arabidopsis*. Seurat was used to identify differentially expressed genes across homologous cell types between *L. japonicus* and *Arabidopsis*. Fold-differences of > 5-fold change were the threshold for identifying divergent differences in gene expression. Functional gene family enrichment (ftp://ftp.arabidopsis.org/home/tair/Genes/Gene_families/) was analyzed using the set of diverged genes.

Functional analysis

Cluster-enriched genes with statistical significance were applied to clusterProfiler (Yu et al., 2012). The R package was used for GO term enrichment analysis. *L. japonicus* phytohormone-related genes were manually chosen for visualization of expression profiles, the expression levels of the chosen genes were summarized, and the top 20% of the cells with high expression levels were visualized on UMAP.

Genome sequencing

The ~20 kb library was constructed and carried out based on the PacBio sequence platform (Pacific Biosciences, CA, USA), a total of 10 cells were sequenced and 54.85 Gb long fragment data were obtained. In order to construct Hi-C library, the genomic DNA in natural state in cells was fixed with formaldehyde and digested with restriction enzyme MboI. Then the constructed library was sequenced based on Hiseq X Ten platform, the raw data was filtered by SOAPnuke v.1.5.6 (Chen et al., 2018), and obtained 75.01 Gb clean data. The 10x Genomics long fragment library was constructed and sequenced by Hiseq X ten platform (Illumina, CA, USA), a total of 107.85 GB of filtered data were obtained. The 350 bp small fragment library was

constructed and sequenced using Hiseq 4000 platform (Illumina, CA, USA), and 51.61 GB clean data was obtained.

RNA was extracted from the roots, leaves, stems, flowers, seed and seed pods, respectively. The mixed RNA sequencing and full-length transcriptome sequencing were performed based on DBNseq (BGI, Shenzhen, Guangdong, China) and PacBio Sequel, respectively.

Genome assembly

In order to use the short reads to estimate the genome size and heterozygosity, we used jellyfish v2.2.6 software to obtain the frequency table based on 17-mer, and then used GenomeScope v1.0 (Vurture et al., 2017) software to evaluate the genome size and heterozygosity.

We assembled the PacBio subreads sequence into preliminary contigs by using MECAT (Xiao et al., 2017). The results were corrected by Polish (Arrow) and pilon (Walker et al., 2014). Compared 10x Genetics linked read to the above corrected contigs using *BWA* software, classified the barcode information at both ends of the original scaffold using ARCS (Assembly Roundup by Chromium Scaffolding) (Yeo et al., 2018), and connected the scaffolds that met the set conditions with 10 N to obtain the scaffolds using LINKS (Warren et al., 2015). Effective Hi-C connected reads were obtained using Judge v.1.5 (Durand et al., 2016), and above scaffolds were anchored to pseudochromosomes using 3 D-DNA pipeline v.180922 (Dudchenko et al., 2017).

Genome annotation

De novo prediction of repeat annotation was based on self sequence alignment, full long terminal repeats (LTRs) retrotransposons were predicted through LTR_Finder v1.06 (Xu & Wang, 2007). Firstly *de novo* (Price et al., 2005) repeat sequence library was established using softwares such as Piler (Edgar & Myers, 2005), RepeatScout, RepeatModeler, etc., and then *de novo*

TEs were predicted through Repeatmaker software. In addition, Tandem Repeats Finder v4.09 (Benson, 1999) was used to find tandem repeats in the genome.

MAKER pipeline (Cantarel et al., 2008) was used for two rounds of genome annotation based on homology, transcripts and de novo method. The mixed libraries ISOforms and unigenes are used as transcriptome evidence, and the protein sequences of *Arabidopsis* and other 7 published leguminous species are used as homologous evidence. In the first round, transcriptome evidence and homologous evidence were used to obtain annotation results and train the parameters of Augustus (Stanke et al., 2006) and SNAP (Johnson et al., 2008). In the second round, we combined Augustus and SNAP trained models with homologous evidence and transcriptional evidence to predict the gene set. For StringTie results, ORF was predicted based on Transedcode, and the complete structural genes without overlap with MAKER results were added to the gene set. GeMoMa V1.7 was used for homology annotation to adjust some genes in the above gene set.

Gene function annotation was based on the consistency of protein sequence and domain. Protein sequences were blastp with NCBI non-redundancy protein sequence (NR) database, Kyoto Encyclopedia of genes and genes (KEGG v89.0; (Kanehisa et al., 2017)), Swissprot and TrEMBL (UniProt release 2020-06) (Boeckmann et al., 2003). Domains of sequences were scanned in the databases ProDom (Servant et al., 2002), PRINTS (Attwood et al., 2000), Pfam (Bateman et al., 2004), PANTHER (Thomas et al., 2003), PROSITE profiles (Sigrist et al., 2010), and SMART (Letunic et al., 2012) using InterProScan v5.11-55.0 (Zdobnov & Apweiler, 2001). Gene Ontology (GO) terms (Ashburner et al., 2000) annotation of genes was based on the blast with NR library.

For the identification of miRNA, small RNA was enriched and purified from total RNA, single strand circle DNA (ssCir DNA) added unique molecular identifiers (UMI) labeled primer was formatted as the final library, which was

sequenced by DNBSEQ platform. Raw reads were filtered by removed contaminant tags and the length distribution of clean tags is then summarized, Bowtie2 (Langmead et al., 2009) was used to map clean reads to the reference genome and to other sRNA databases. To make sure every unique small RNA is mapped to only one category, we followed this priority rule: MiRbase>pirnabank>snoRNA(human/plant)>Rfam>othersRNA. We used miRA (Nawrocki & Eddy, 2013; Evers et al., 2015) to predict novel miRNA by exploring the characteristic hairpin structure of miRNA precursor. Finally, target genes were predicted using TargetFinder (Fahlgren & Carrington, 2010) and kept only if their score was greater than 3.

For the identification of LncRNA, After the rRNA was removed from total RNA, cDNA libraries were constructed and sequenced on BGISEQ-500. Clean reads were aligned and assembled using HISAT (Kim et al., 2015) and String Tie (Pertea et al., 2015), then compared with known mRNA and lncRNA by cuffcompare (Trapnell et al., 2010) to receive the newly assembly transcripts. We predicted the coding ability of above transcript to distinguish between mRNA and lncRNA using CPC v0.9-r2 (Kong et al., 2007), txCdsPredict, CNCI (Sun et al., 2013) and pfam (Bateman et al., 2004).

Supplementary figure

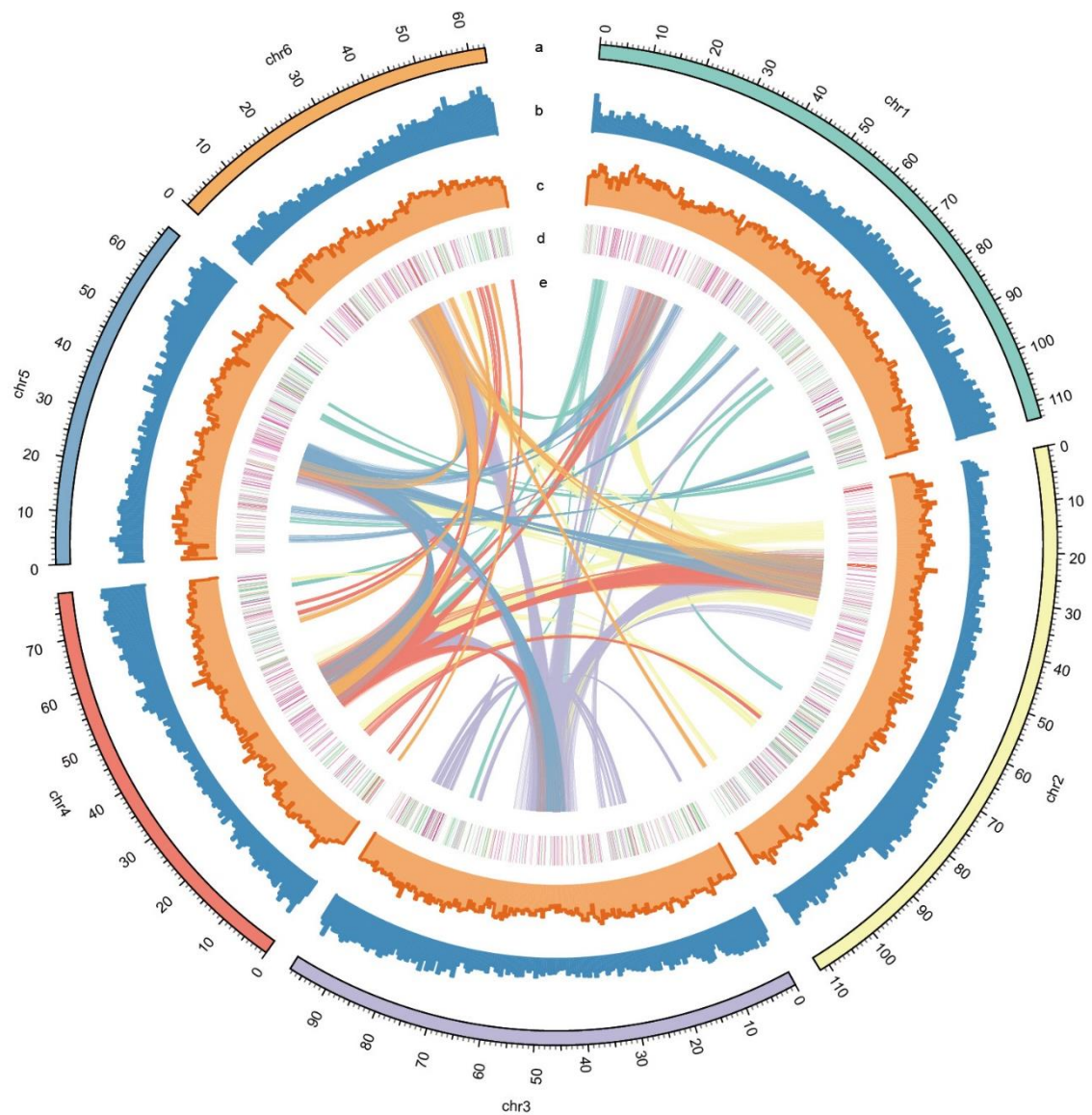


Figure S1. *L. japonicus* genome assembly and annotation

Circos diagram illustrating important features of the assembled *L. japonicus* genome. **a** six assembled chromosomes. **b** Gene density. **c** Repetitive elements density, **d** Non-coding RNAs distribution, tRNAs: green; rRNAs: red; snRNAs: rose; miRNAs: blue. **e** Blocks of synteny of at least five gene pairs between the genomes are connected by linked lines at the center of the Circos plot. Different colors represent different pseudochromosomes or syntenic blocks.

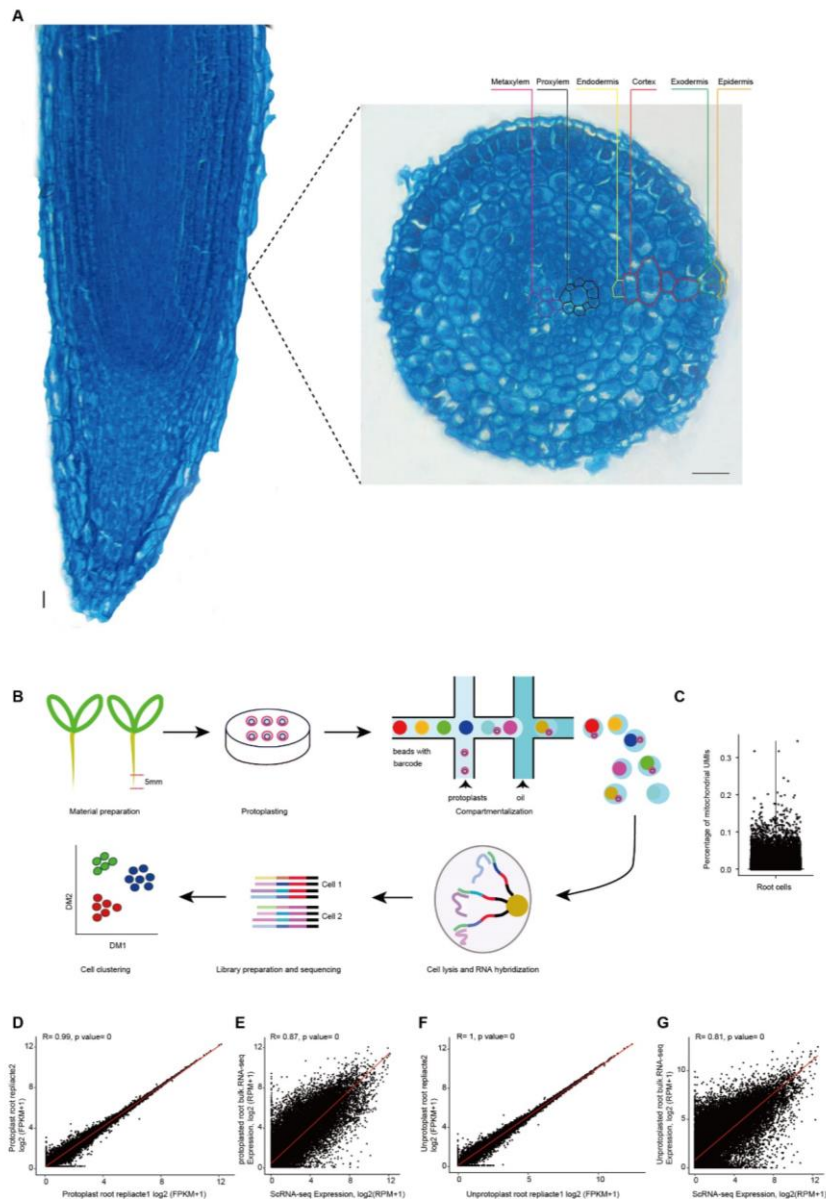


Figure S2. Morphology, scRNA-seq, and bulk RNA-seq of *L. japonicus* root tip

(A) Longitudinal section (left) and cross section (right) showing root tip morphology of *L. japonicus*. Scale bars, 20 μm for longitudinal section and 50 μm for cross section. **(B)** Overview of *L. japonicus* scRNA-seq workflow. **(C)** Percentage of mitochondrial UMIs in *L. japonicus* root tip cells. **(D)** Scatter plots showing the correlation between replicates of root protoplast bulk RNA-seq of *L. japonicus*. **(E)** Scatter plot showing the correlation between root protoplast bulk RNA-seq and combined scRNA-seq of *L. japonicus*. **(F)** Scatter plots showing the correlations between replicates of root cell bulk RNA-seq of *L. japonicus*. **(G)** Scatter plots showing the correlations between root cell bulk RNA-Seq and combined scRNA-seq of *L. japonicus*.

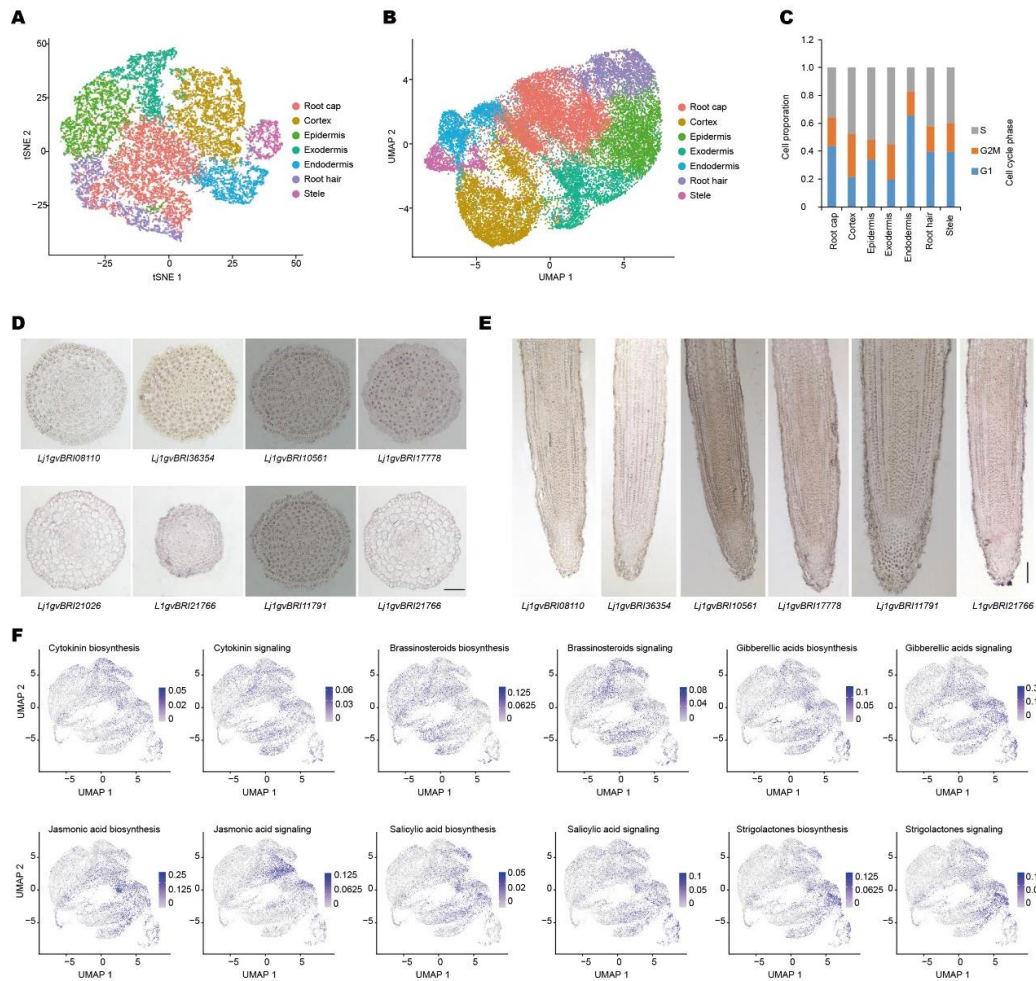


Figure S3. scRNA-seq cluster and annotation

(A) *t*-SNE visualization of seven clusters with cell-type annotation for *L. japonicus*. **(B)** After filtering out genes with differential expression introduced by preparation of protoplasts, UMAP visualization of seven clusters with cell-type annotation for *L. japonicus*. **(C)** Proportion of inferred cell cycle phases (G1, S and G2/M) in each cluster. **(D,E)** Cross section **(D)** and longitudinal section **(E)** for representative RNA *in situ* hybridization of cell-type marker genes with sense probes as negative control. Scale bars, 50 μ m. **(F)** UMAP visualization of expression patterns of the genes related to phytohormone biosynthesis and signaling, including those related to cytokinin, brassinosteroid, gibberellic acid, jasmonic acid (JA), salicylic acid, and strigolactone functions. Color bars indicate scaled expression levels. Related genes are listed in Table S23.

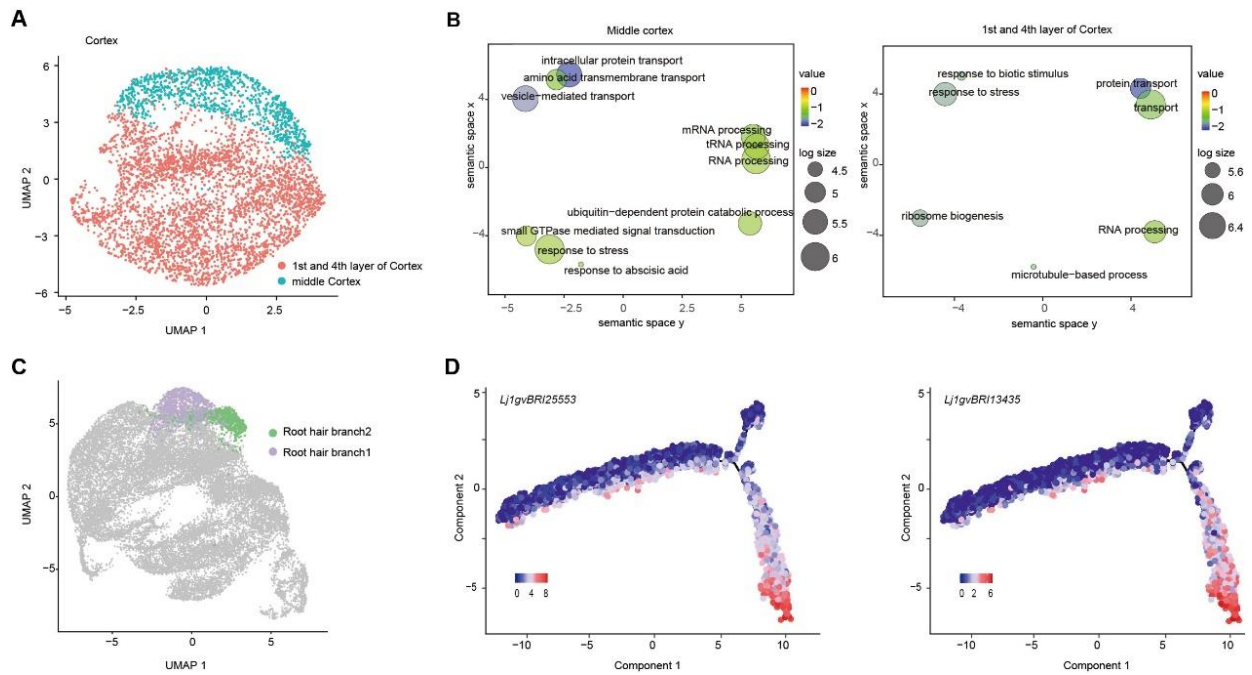


Figure S5. Analysis of the subclusters of root cortex and root hair cells

(A) UMAP visualization of the cluster of cells of the middle cortex (second and third layer of cortex) and the cluster of cells of the inner and outermost layer of the root cortex. **(B)** Scatter plots of GO term enrichment analysis of differentially expressed genes in middle cortex compared with the first and fourth layers of the root cortex. **(C)** UMAP visualization of cells of Branch 1 and Branch 2 (related to Fig. 2B) from pseudotime analysis of root hair cells. **(D)** Expression of cell-type marker genes on Branch 2 of the pseudotime trajectory. Color bars indicate scaled expression levels.

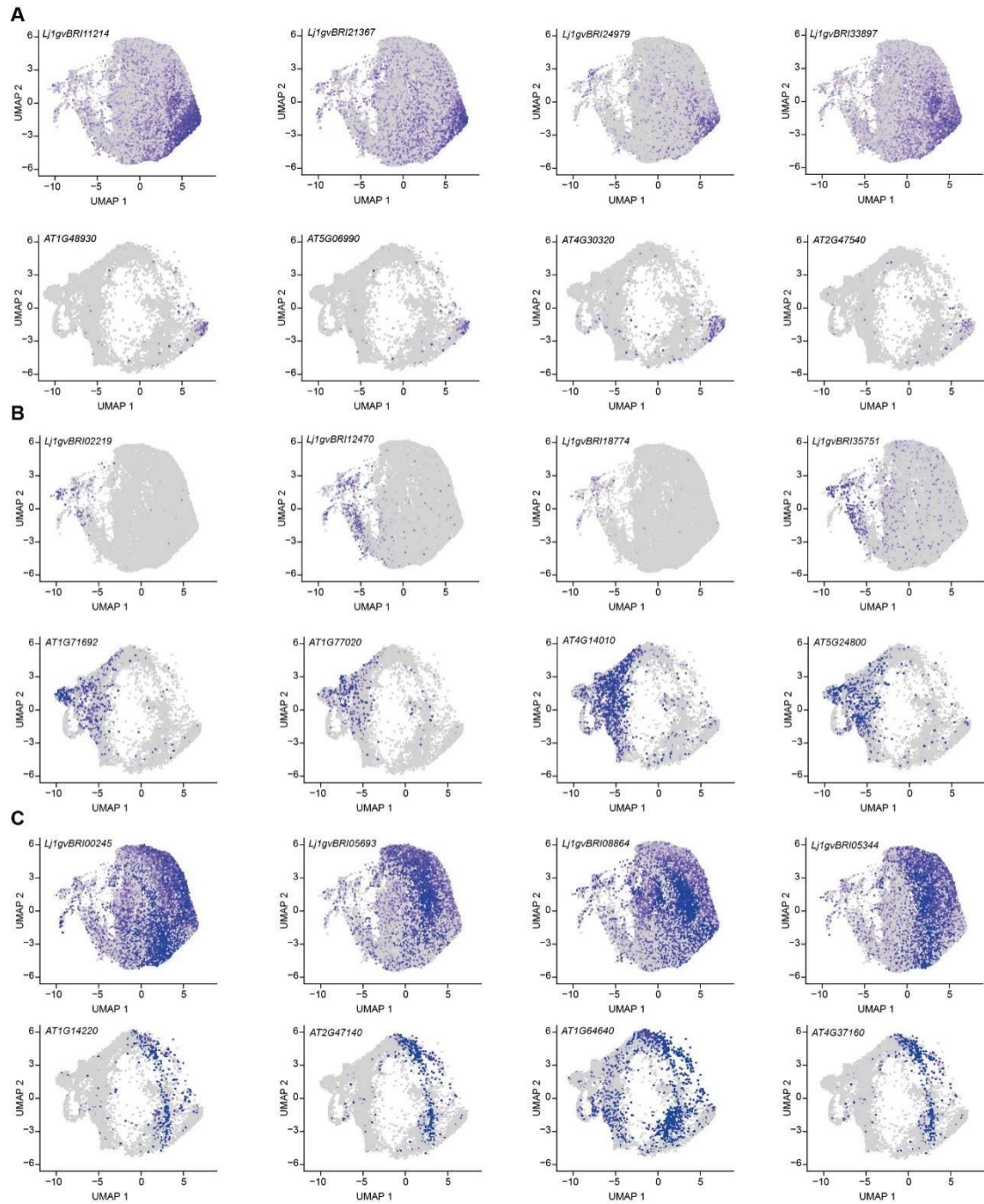


Figure S6. Analysis of conserved cell types and marker genes between *L. japonicus* and *Arabidopsis*

(A–C) UMAP visualization of expression of conserved marker genes for root hair (A), stele (B) and root cap (C) cells between *L. japonicus* and *Arabidopsis*. Color bars indicate scaled expression levels.

Table S1. Statistics of Illumina RNA-seq raw reads.

Table S2. Statistics of Illumina RNA-seq clean reads.

Table S3. Statistics of full-length isoform sequencing (Iso-seq) of *L. japonicus*.

Table S4. Statistics of 10x Genomics sequencing of *L. japonicus*.

Table S5. Statistics of PacBio sequencing of *L. japonicus*.

Table S6. Statistics of Hi-C sequencing of *L. japonicus*.

Table S7. Statistics of the assembled contigs and scaffolds of *L. japonicus* (10x Genomics).

Table S8. The mapping statistics of the high quality Hi-C data.

Table S9. Completeness of the genome based on BUSCO analysis.

Table S10. The summary of genome assembly results of *L. japonicus*.

Table S11. Summary of annotated genes.

Table S12. Summary of functional annotation of predicted genes.

Table S13. Summary of annotated non-coding RNA.

Table S14. List of annotated lncRNA in *L. japonicus*.

Table S15. List of annotated snRNA in *L. japonicus*.

Table S16. List of annotated rRNA in *L. japonicus*.

Table S17. List of annotated tRNA in *L. japonicus*.

Table S18. Statistics of repeat sequences in *L. japonicus* genome.

Table S1-S18 See excel file.

Table S19. Statistics of scRNA-seq in *L. japonicus*.

Cultivar	Number of cell	Mean reads per cell	Median genes per cell	Total genes detected	Median UMI per cell
MG-20	22688	48,898	3,183	28,464	7,947

Table S20. Number of cells captured by scRNA-seq for each cell type.

Cell type	Number
Root Cap	6760
Root hair	2115
Cortex	4639
Epidermis	3279
Exodermis	2624
Endodermis	2139
Metaxylem	1126

Table S21. DEGs between protoplasted bulk RNA-seq and un-protoplasted (root tissues) bulk RNA-seq.

See excel file.

Table S22. List of enriched genes for each cluster of *L. japonicus*.

See excel file.

Table S23. List of phytohormone related genes.

See excel file.

Table S24. List of nodulation related genes.

See excel file.

Table S25. List of primers used in this study.

Antisense probe for in situ hybridization

Primer ID	Primers (5'-3')
Lj1gvBRI08110	AAGACATCCGACCAGCGTAA AGAGCAAGACCGAGCCTAA
Lj1gvBRI36354	GCCAAAGCCATCCTTCCTC GCCGTCGTCGTAGTGTTTCG
Lj1gvBRI10561	CCGCCAAGGGTCACAAACAC CCTCAAGGAGGGAGCAGCAT
Lj1gvBRI21026	ACGAGACGCTGGCACCTGTA CGGGCTATCCCAACCTTATT
Lj1gvBRI17778	CAACCAGCAAATGTAGGAATAA GAGAACAACGAGGTCAGCAC
Lj1gvBRI11791	TCTGCCATCCTTGCTCATTC TGTTGTTGGTGCTTGCTGTG
Lj1gvBRI21766	ACTGCTGTGACGCCTGATTC CGCCCTTGAAAGTTGTTGTG

REFERENCES

- Anders, S., Pyl, P.T., and Huber, W. (2015).** Htseq-a python framework to work with high-throughput sequencing data. *Bioinform.* **31**: 166–169.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. (2000).** Gene ontology: Tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* **25**: 25–29.
- Attwood, T.K., Croning, M.D., Flower, D.R., Lewis, A.P., Mabey, J.E., Scordis, P., Selley, J.N., and Wright, W. (2000).** PRINTS-S: The database formerly known as PRINTS. *Nucleic Acids Res.* **28**: 225–227.
- Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C., and Eddy, S.R. (2004).** The Pfam protein families database. *Nucleic Acids Res.* **32**: D138–141.
- Becht, E., Mcinnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018).** Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**: 38–44.
- Benson, G. (1999).** Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27**: 573–580.
- Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003).** The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**: 365–370.
- Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018).** Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**: 411–420.
- Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., and Yandell, M. (2008).** MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**: 188–196.
- Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C., Li, Z., Zhang, X., Wang, J., Yang, H., Fang, L., and Chen, Q. (2018).** SOAPnuke: A MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* **7**: 1–6.
- Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., and Aiden, E.L. (2017).** *De novo* assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**: 92–95.
- Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S., Huntley, M.H., Lander,**

- E.S., and Aiden, E.L.** (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**: 95–98.
- Edgar, R.C., and Myers, E.W.** (2005). PILER: Identification and classification of genomic repeats. *Bioinform.* **21**: i152–i158.
- Evers, M., Huttner, M., Dueck, A., Meister, G., and Engelmann, J.C.** (2015). miRA: Adaptable novel miRNA identification in plants using small RNA sequencing data. *BMC Bioinformatics* **16**: 370.
- Fahlgren, N., and Carrington, J.C.** (2010). miRNA target prediction in plants. *Methods Mol. Biol.* **592**: 51–57.
- Johnson, A.D., Handsaker, R.E., Pulit, S.L., Nizzari, M.M., O'donnell, C.J., and De Bakker, P.I.** (2008). SNAP: A web-based tool for identification and annotation of proxy SNPs using hapmap. *Bioinform.* **24**: 2938–2939.
- Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K.** (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**: D353–D361.
- Kim, D., Langmead, B., and Salzberg, S.L.** (2015). HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **12**: 357–360.
- Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L.** (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**: 907–915.
- Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q., Zhao, S.Q., Wei, L., and Gao, G.** (2007). Cpc: Assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* **35**: W345–W349.
- Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L.** (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**: R25.
- Letunic, I., Doerks, T., and Bork, P.** (2012). SMART 7: Recent updates to the protein domain annotation resource. *Nucleic Acids Res.* **40**: D302–D305.
- Liang, Z., Brown, R.C., Fletcher, J.C., and Opsahl-Sorteberg, H.G.** (2015). Calpain-mediated positional information directs cell wall orientation to sustain plant stem cell activity, growth and development. *Plant Cell Physiol.* **56**: 1855–1866.
- Liang, Z., Shen, L., Cui, X., Bao, S., Geng, Y., Yu, G., Liang, F., Xie, S., Lu, T., Gu, X., and Yu, H.** (2018). DNA N⁶-adenine methylation in *Arabidopsis thaliana*. *Dev. Cell* **45**: 406–416.
- Love, M.I., Huber, W., and Anders, S.** (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**: 550.
- Nawrocki, E.P., and Eddy, S.R.** (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinform.* **29**: 2933–2935.
- Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L.** (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**: 290–295.

- Price, A.L., Jones, N.C., and Pevzner, P.A.** (2005). *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**: i351-i358.
- Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D., and Kahn, D.** (2002). ProDom: Automated clustering of homologous domains. *Brief Bioinform.* **3**: 246–251.
- Sigrist, C.J., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A., and Hulo, N.** (2010). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.* **38**: D161–D166.
- Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B.** (2006). AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**: W435–W439.
- Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., Liu, Y., Chen, R., and Zhao, Y.** (2013). Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* **41**: e166.
- Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A.** (2003). PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.* **13**: 2129–2141.
- Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L.** (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**: 381–386.
- Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L.** (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**: 511–515.
- Van Der Maaten, L., and Hinton, G.** (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**: 2579–2605.
- Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., and Schatz, M.C.** (2017). Genomescope: Fast reference-free genome profiling from short reads. *Bioinform.* **33**: 2202–2204.
- Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M.** (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**: e112963.
- Warren, R.L., Yang, C., Vandervalk, B.P., Behsaz, B., Lagman, A., Jones, S.J., and Birol, I.** (2015). LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *Gigascience* **4**: 35.
- Xiao, C.L., Chen, Y., Xie, S.Q., Chen, K.N., Wang, Y., Han, Y., Luo, F., and Xie, Z.** (2017). MECAT: Fast mapping, error correction, and de novo

- assembly for single-molecule sequencing reads. *Nat. Methods* **14**: 1072–1074.
- Xu, Z., and Wang, H.** (2007). LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**: W265–268.
- Yeo, S., Coombe, L., Warren, R.L., Chu, J., and Birol, I.** (2018). ARCS: Scaffolding genome drafts with linked reads. *Bioinform.* **34**: 725–731.
- Yu, G., Wang, L.G., Han, Y., and He, Q.Y.** (2012). ClusterProfiler: An R package for comparing biological themes among gene clusters. *Omics* **16**: 284–287.
- Zdobnov, E.M., and Apweiler, R.** (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinform.* **17**: 847–848.
- Zhang, T.Q., Xu, Z.G., Shang, G.D., and Wang, J.W.** (2019). A single-cell RNA sequencing profiles the developmental landscape of *Arabidopsis* root. *Mol. Plant* **12**: 648–660.
- Zhao, Y., Hu, Y., Dai, M., Huang, L., and Zhou, D.X.** (2009). The WUSCHEL-related homeobox gene *wox11* is required to activate shoot-borne crown root development in rice. *Plant Cell* **21**: 736–748.