Supplementary materials for

# Ancient sex compatibility genes underlie tissue differentiation in mushrooms

Torda Varga[1], Hongli Wu[2]*, Anna Szekeres[2]*, Árpád Csernetics[2]*, Csenge Földi[2]*, Máté Virágh[2]*, Nikolett Miklovics[2]*, Xiao-Bin Liu[2]*, Zhihao Hou[2]*[$], Zsolt Merényi[2], Gábor Berend[3], Blythe Angus[4], Stefan Manolache[4], Victor Roces[4], András Vágner[2], Ji-Peng Li[2], Zsuzsanna Darula[2,5], Dorottya Szafián[2], Edit Ábrahám[2], Péter Horváth[2], Zsolt Kristóffy[2], Jinwei Zhang[2], Balázs Bálint[2], Botond Hegedüs[2], Nikolett Zsibrita[2], Hajk Drost[4,6], Zoltán Lipinszki[2], László G. Nagy[2,7,8&]

[1] Royal Botanic Gardens Kew, Richmond, TW9 3DS UK
[2] HUN-REN Biological Research Centre, Szeged 6726, Hungary
[3] Institute of Informatics, University of Szeged, Szeged 6726, Hungary
[4] Digital Biology Group, Faculty of Life Sciences, University of Dundee, Dundee, United Kingdom
[5] Single Cell Omics Advanced Core Facility, Hungarian Centre of Excellence for Molecular Medicine,Szeged 6728, Hungary
[6] Computational Biology Group, Max Planck Institute for Biology Tübingen, Tübingen, Germany
[7] Korea University, Seoul 02841, Republic of Korea
[8] Shandong Agricultural University, Tai'an, China

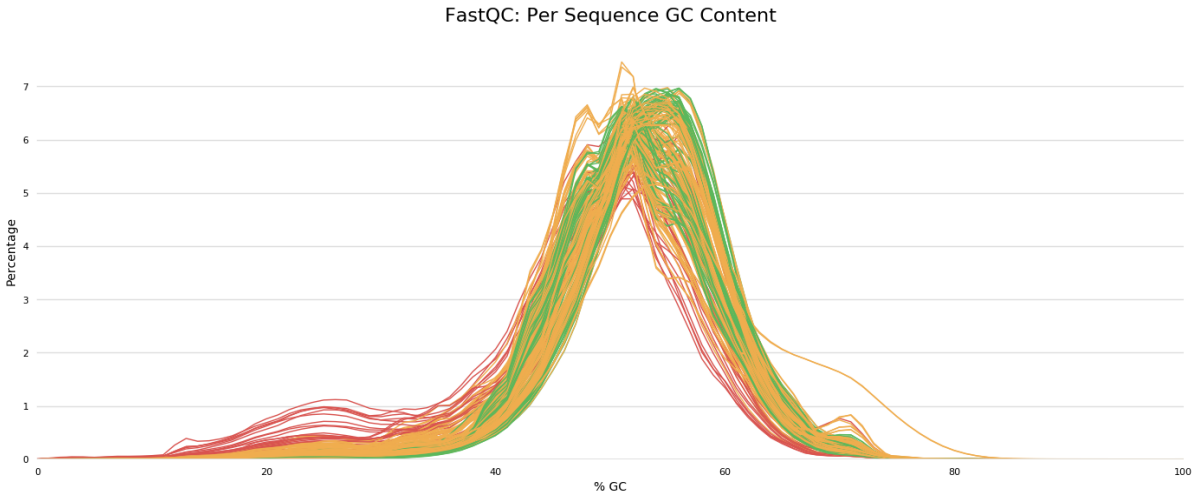
**Contents**:

Supplementary Note 1.
Supplementary Figures 1-20.

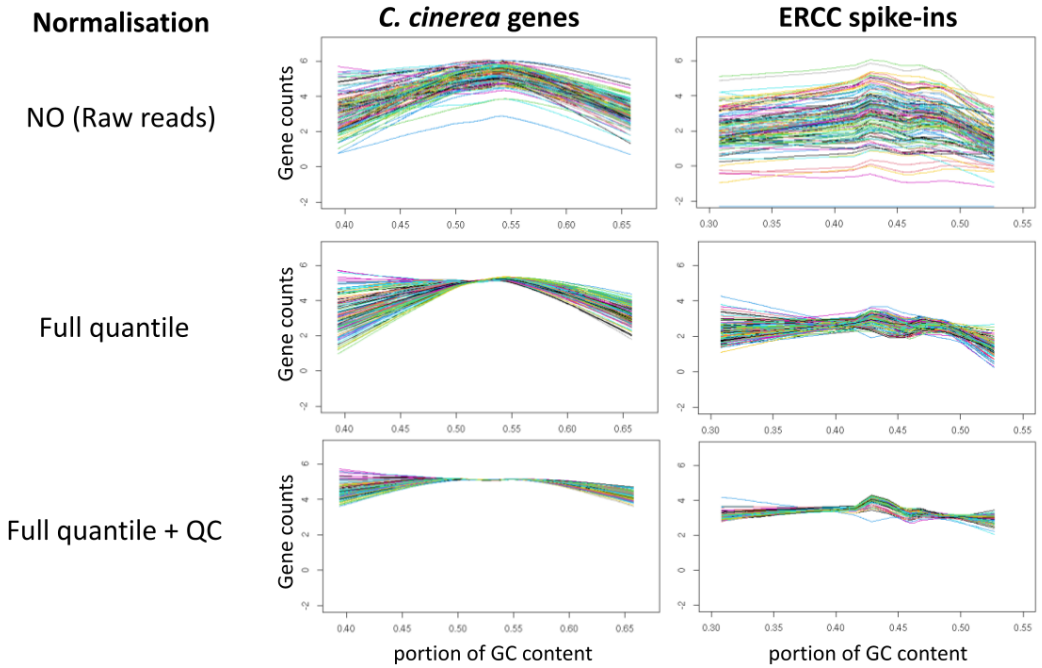
**Supplementary Note 1.** Detailed results of RNA-seq normalisation and outlier detection.

## GC bias

We combined fastqc reports by using MultiQC program and found that out of the 236 paired-end read libraries, 162 showed a certain level of bias in GC content (**Supplementary Note 1, Figure 1.**). Then, we checked the gene-specific GC effects on the read counts using the biasPlot function in the EDAseq R package. This analysis further supported the GC bias because full quantile normalisation did not completely eliminate differences between libraries. (**Supplementary Note 1, Figure 2.**). This difference could be mitigated by normalising read counts using a GC content scaling factor.




**Supplementary Note 1, Figure 1.** The GC distribution of 236 libraries (118 samples). 138 libraries were mildly biased (yellow: indicated as warnings in the MultiQC report) and 24 libraries were severely biased (red: indicated as failed in the MultiQC report).

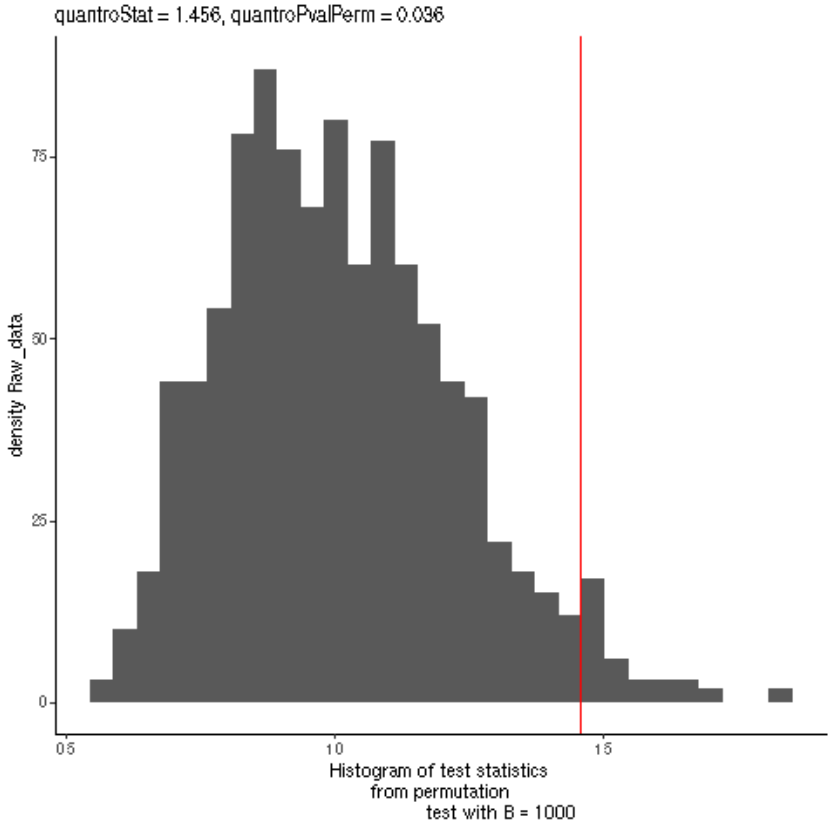



**Supplementary Note 1, Figure 2. The effect of normalisation on read counts.** We expect that the average read count (Y-axis) is similar between libraries and independent of GC content (x-axis). We analysed 118 libraries of *C. cinerea* genes (left column) and ERCC spike-in external RNA controls (right column), showing

that full quantile library normalisation did not eliminate all differences between libraries, but after applying GC-normalisation, differences were minimised.

## Testing the assumption of the global normalisation

We tested if there is a global trend in our dataset that could be attributed to biological variance and would be masked by global normalisation methods such as full quantile normalisation. Performing 1,000 permutations with the quantro function of the quantro R package, we found that the between-group (class) variation is significantly higher than the within-group variation (P = 0.036), suggesting that no global normalisation method should be applied to our data (**Supplementary Note 1, Figure 3.**).

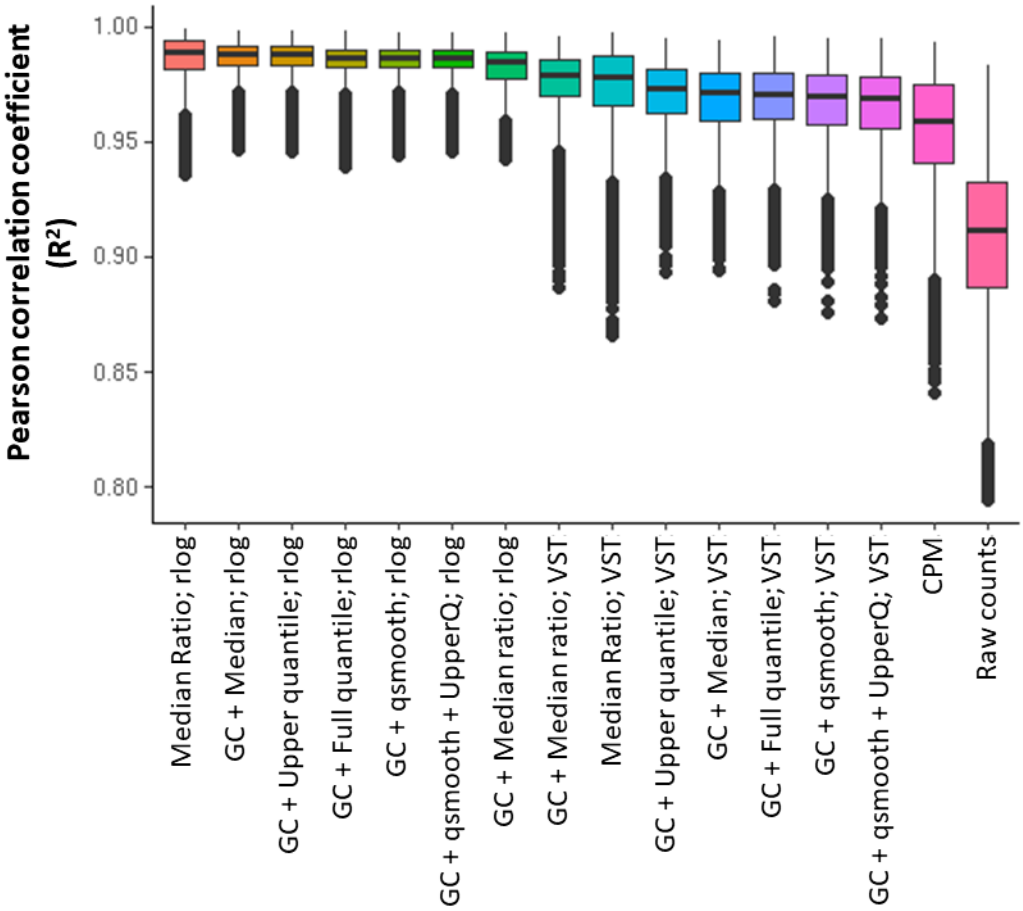



**Supplementary Note 1, Figure 3.** The result of the permutation test performed by the quantro function. The histogram shows the generated null test statistics. The red line depicts the observed statistic.

## Comparing rlog and VST transformations

To use log-transformed normalised counts in subsequent analyses, two options are provided in the DESeq2 package: variance stabilising transformation (VST) and regulized logarithm (rlog) transformation methods. To determine which transformation best suits our data, we performed eight different transformations and compared the input concentration of the ERCC molecules with their normalised transformed read counts. The eight normalisation methods were Median ratio, GC normalisation coupled median, GC normalisation coupled median ratio, GC normalisation coupled full quantile, GC normalisation coupled qsmooth, GC normalisation coupled qsmooth and upper quantile and finally count per million (CPM) normalisation. GC and full quantile normalisation were performed by using the EDAseq package. We combined normalisations (e.g., QC + qsmooth) by first calculating a normalised count of the first method, then subsequent normalisation was applied to this data. In each case, we generated a size factor that was used in DESeq2 calculations. The size factor was calculated by the exp(-1 x dataoffset) formula in the case of EDAseq-based normalisations ("GC" and "full quantile" normalisations). For size factors generated in DEseq2 (e.g., median ration) the *estimateSizeFactors* function was used. In the case of qsmooth normalisation, we used the strategy of the EDAseq package by calculating an offset manually, and then using this offset, we generated a size factor as follows. First, we calculated a normalised counts

with the qsmooth function of the qsmooth R package, then we subtracted the log2 counts + 0.1 from the log2 normalised counts + 0.1 to calculate an offset. Then the size factor was calculated with the formula of exp(-1 x dataoffset).
By calculating the Pearson correlation coefficient ($R^2$) between the input concentration of the ERCC molecules and their normalised and transformed read counts in each library we found that rlog transformation outperformed VST, CPM and raw data. Among the rlog transformed data, the different normalisation methods gave similar Pearson correlation coefficients, including the GC coupled qsmooth normalisation that was our final chosen method. (**Supplementary Note 1, Figure 4.**).

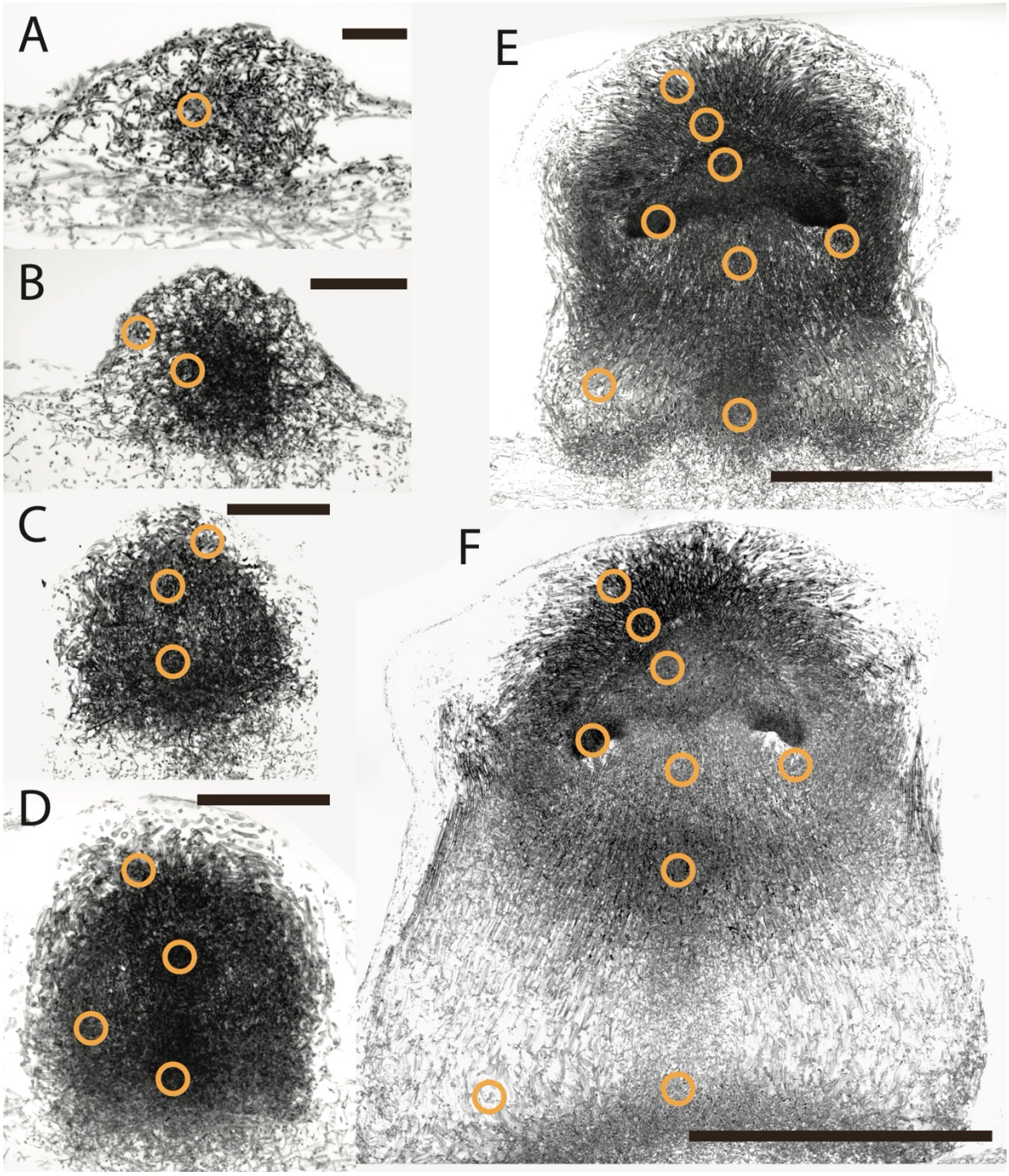



**Supplementary Note 1, Figure 4.** The comparison of different normalisation and transformation methods based on the correlation of the input ERCC molecule concentration and their normalised and transformed counts. The data is organised in decreasing order along the mean Pearson correlation coefficient. Note that there is a small drop and then a gradual decrease in the $R^2$ when starting the VST transformation.

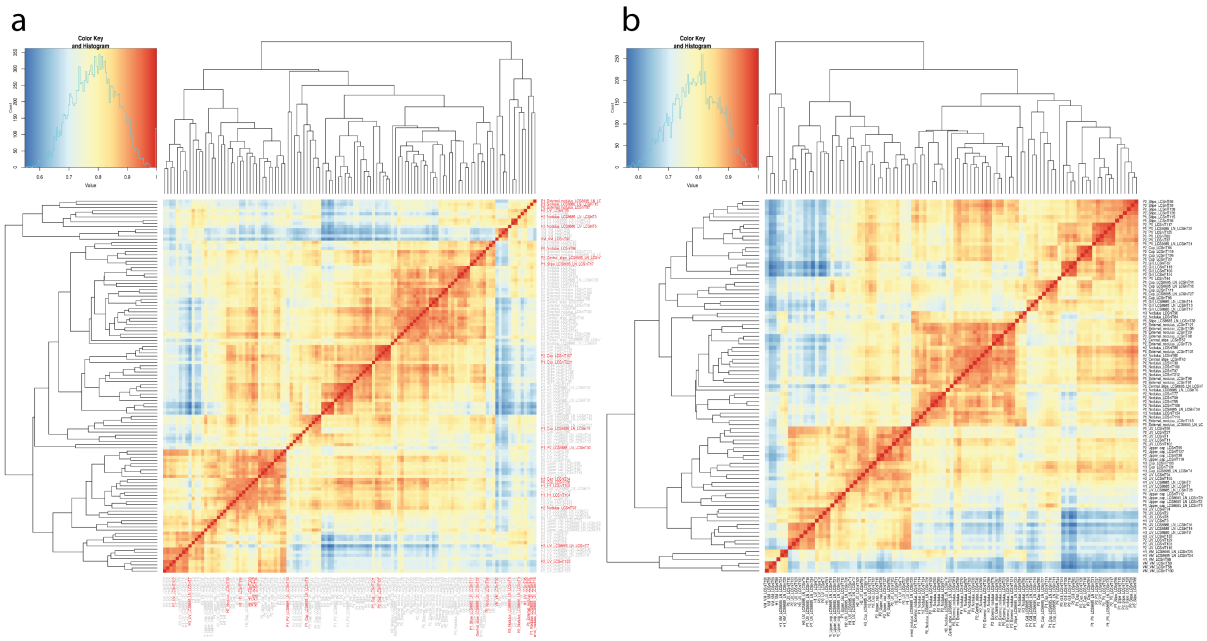
## Outlier detection

We filtered out outlier libraries based on visually inspecting PCA and correlation plots as well as evaluating ERCC statistics (the percentage of the detected spike-in molecules; the Pearson correlation coefficient between the input ERCC spike-in concentration and the normalised read counts; the lower level detection of molecules), number of zero count genes and outliers based on Cook's distance.
We excluded 21 libraries (Supplementary Table 1), out of which 11 samples experienced technical bias based on ERCC statistics (n=9), a high number (>3000) of zero-count genes (n=6), and a high number of outlier genes based on Cook's distance statistics (n=2). Another 15 samples were excluded because they did not cluster well with other replicates and tissues based on PCA (n = 15) and hierarchical clustering (n = 10). Interestingly, seven samples showed a level of technical bias but resulted in good clustering with other replicates and ten samples were excluded solely based on poor grouping on PCA plots and

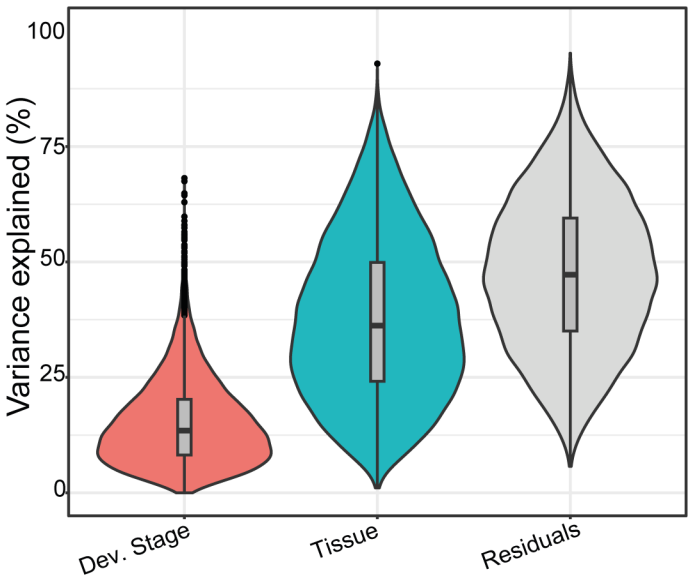
hierarchical clustering. After removing the 21 libraries from the dataset, the hierarchical clustering was improved (Supplementary Fig. 2), and some primordia samples that showed high overlap between different developmental stages based on size dimensions were also eliminated. The latter indicates that some of the differences at the sequencing library level could have originated from picking primordia at an unsuitable stage.



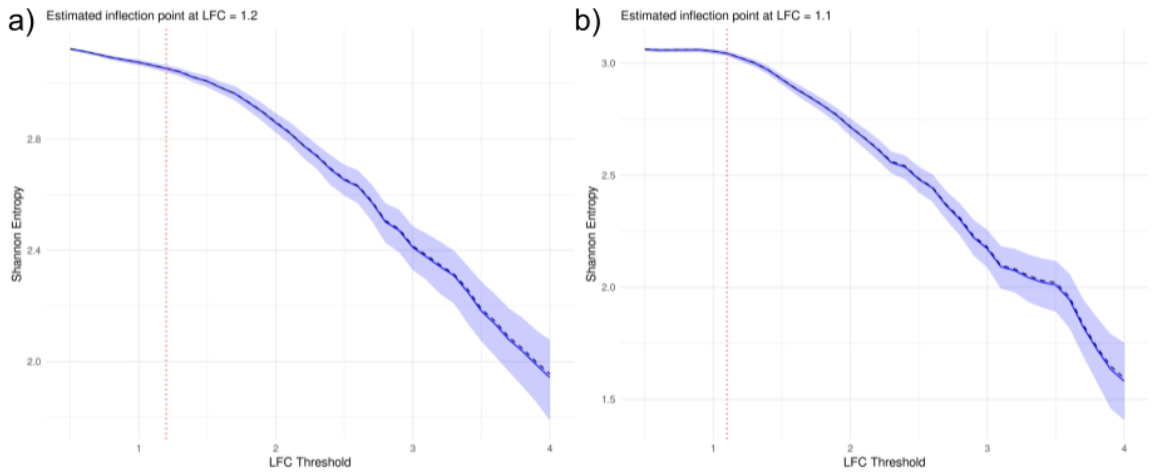

**Supplementary Figure 1.** Light microscopic images of developmental stages and tissue types (yellow circles) in each stage sample for ultralow input RNA-seq in this study.



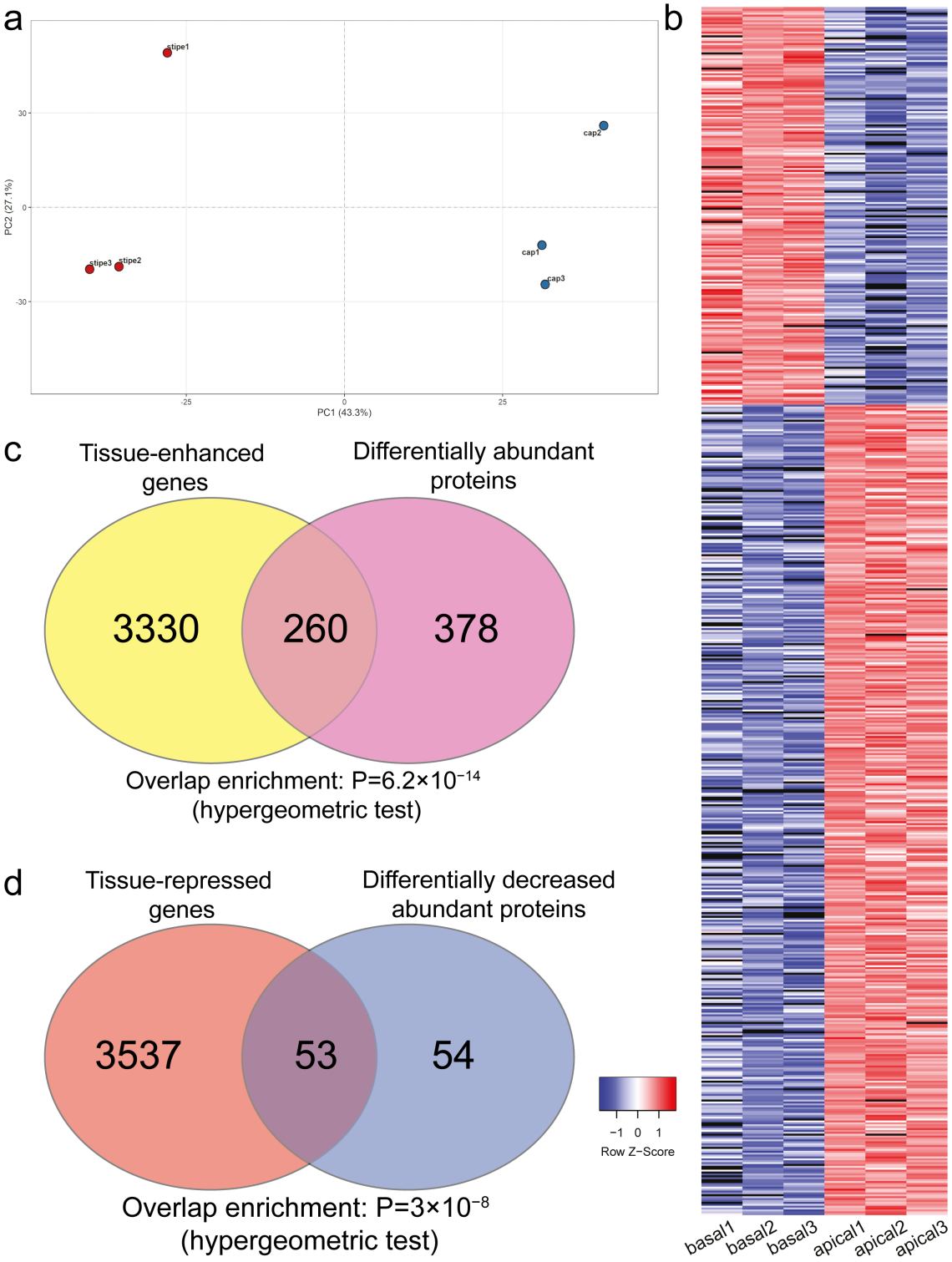

**Supplementary Figure 2.** Hierarchical clustering of all RNAseq libraries based on the Pearson correlation coefficient calculated using the rlog-transformed GC and qsmooth normalised reads. **a.** All RNAseq libraries. Samples highlighted in red were excluded during the outlier detection process. **b.** RNA-seq libraries after outlier samples were excluded.



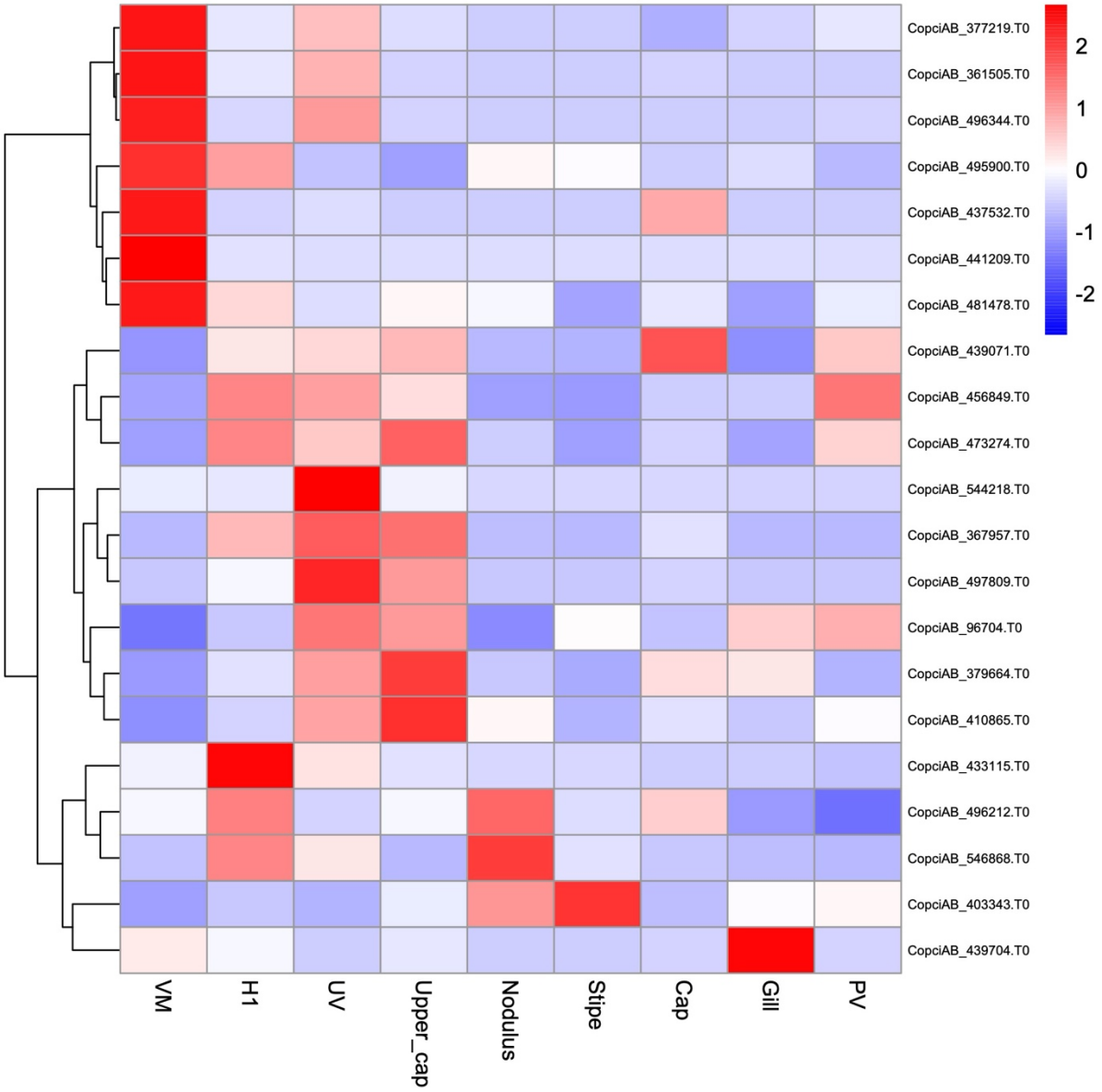

**Supplementary Figure 3.** The distribution of variance across tissues and stages



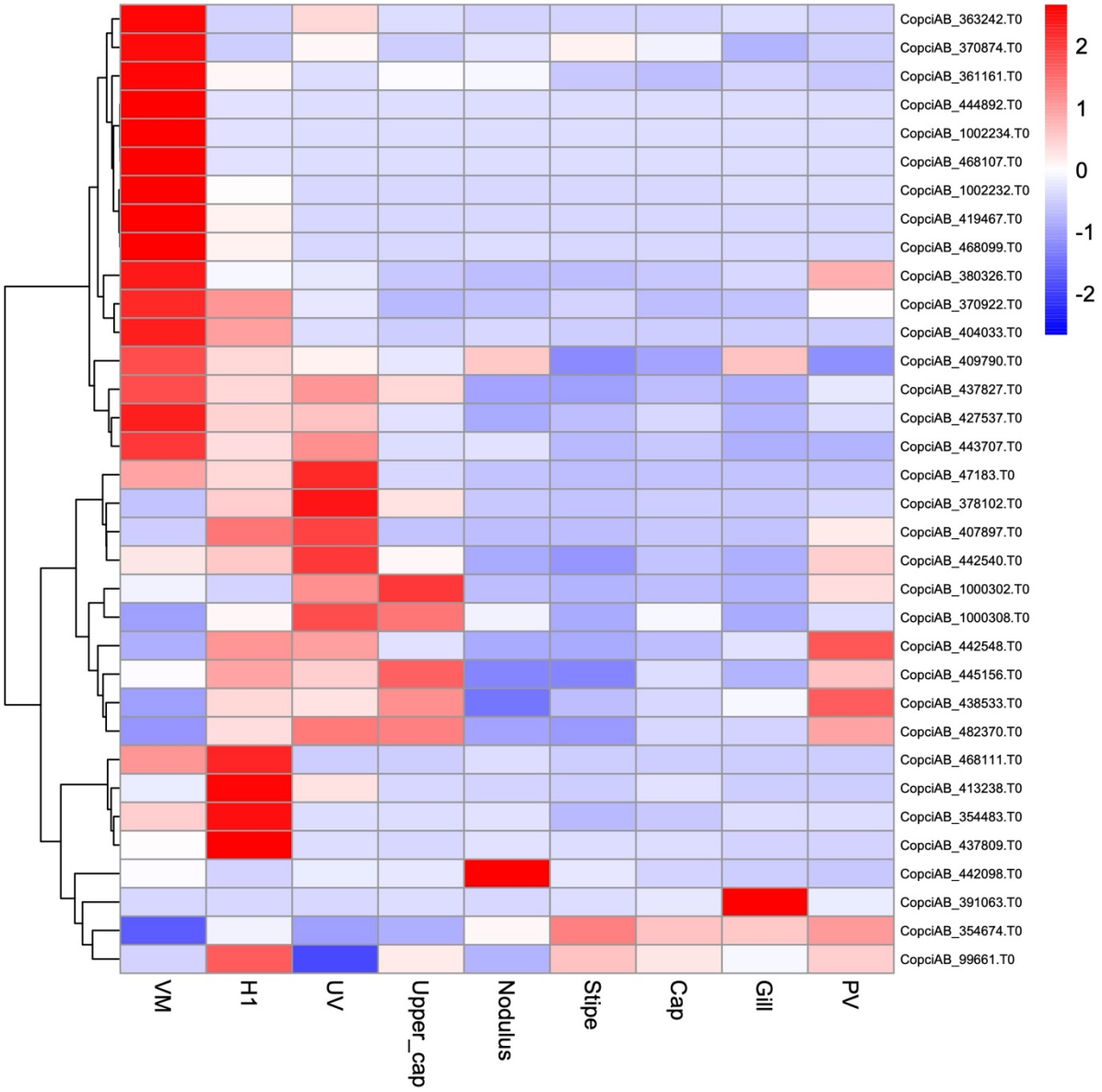

**Supplementary Figure 4.** The effect of the log-fold threshold on the transcriptome diversity measured by Shannon entropy. An inflection point (red dashed line) was calculated to determine where the entropy drops substantially, indicating that a higher threshold would cause significant loss of transcriptome diversity. a) Enhanced genes. b) Repressed genes.



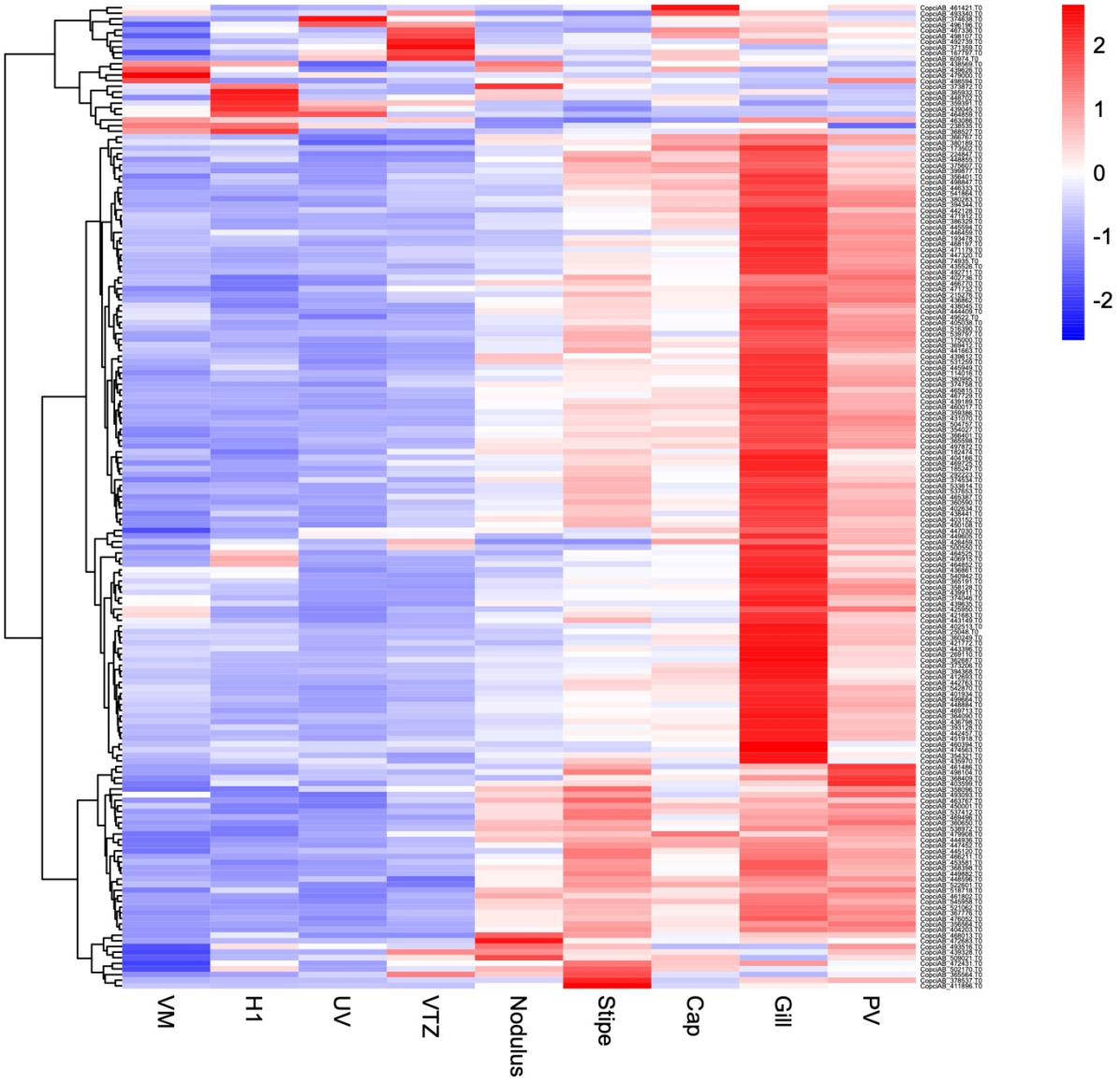

**Supplementary Figure 5. a.** Principal component analysis of biological replicates of apical and basal regions of stage 2 primordia. **b**. Protein abundance heatmap of 638 differentially abundant (p=<0.05) proteins between apical and basal regions of stage 2 primordia of *C. cinerea*. **c**. Venn diagram illustrating the overlap between tissue-enhanced genes and proteins with differential abundance between apical and basal regions. **d**. Venn diagram illustrating the overlap between tissue-repressed genes and proteins detected only in either apical or basal regions.



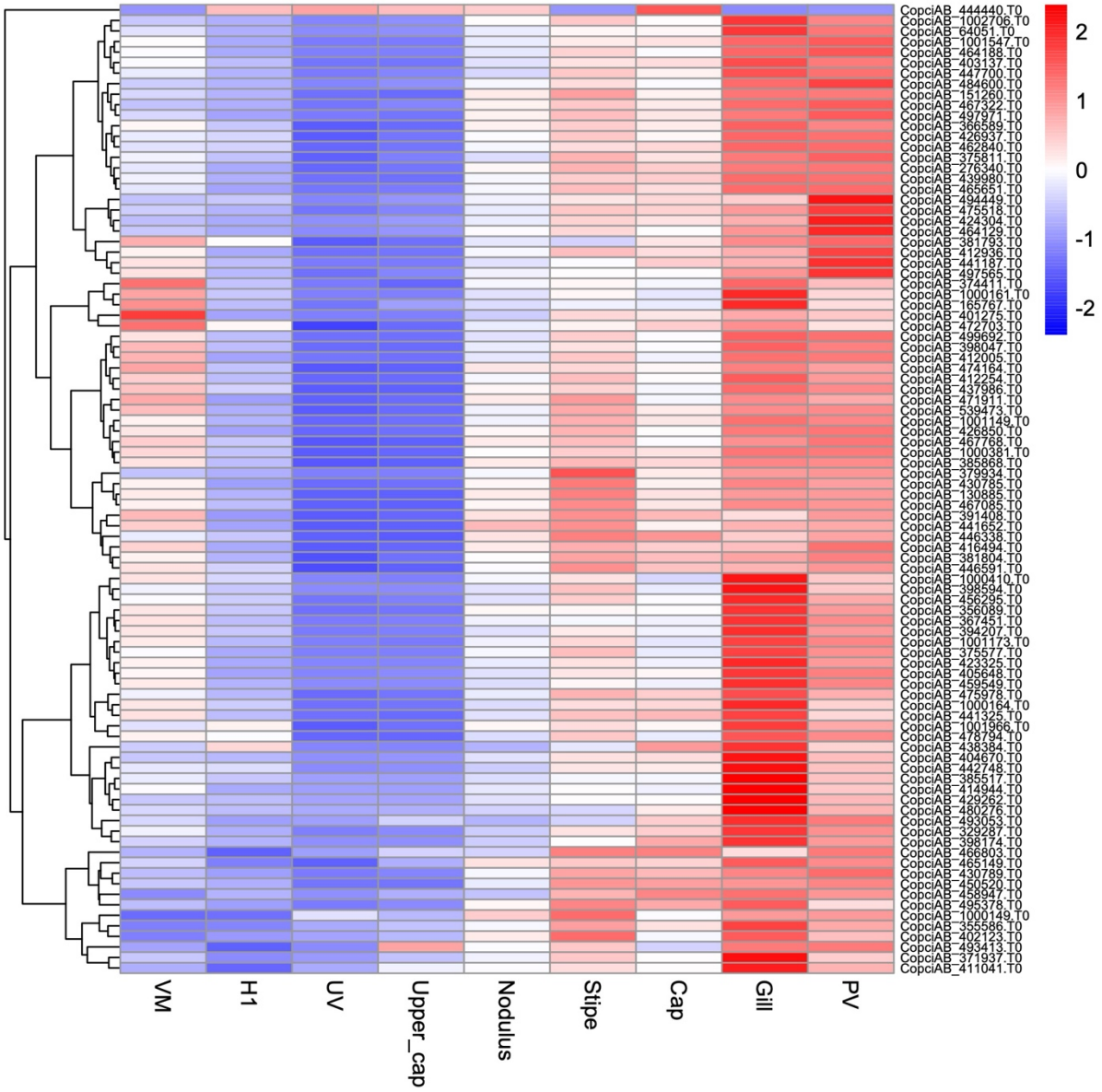

**Supplementary Figure 6.** Expression heatmap for putative defense-related genes across the development and major tissues of *C. cinerea*.



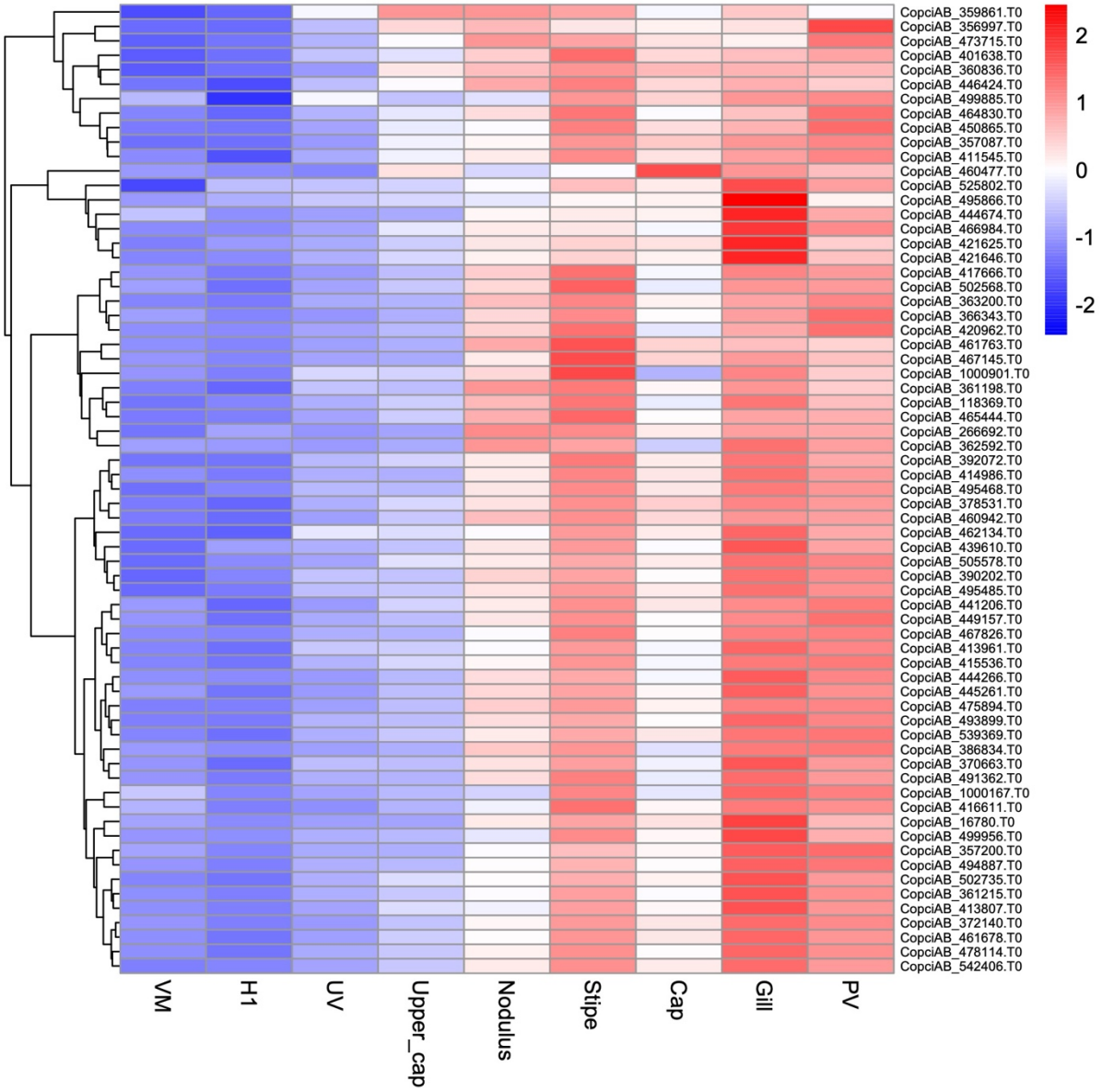

**Supplementary Figure 7.** Expression heatmap for hydrophobin-encoding genes across the development and major tissues of *C. cinerea*.



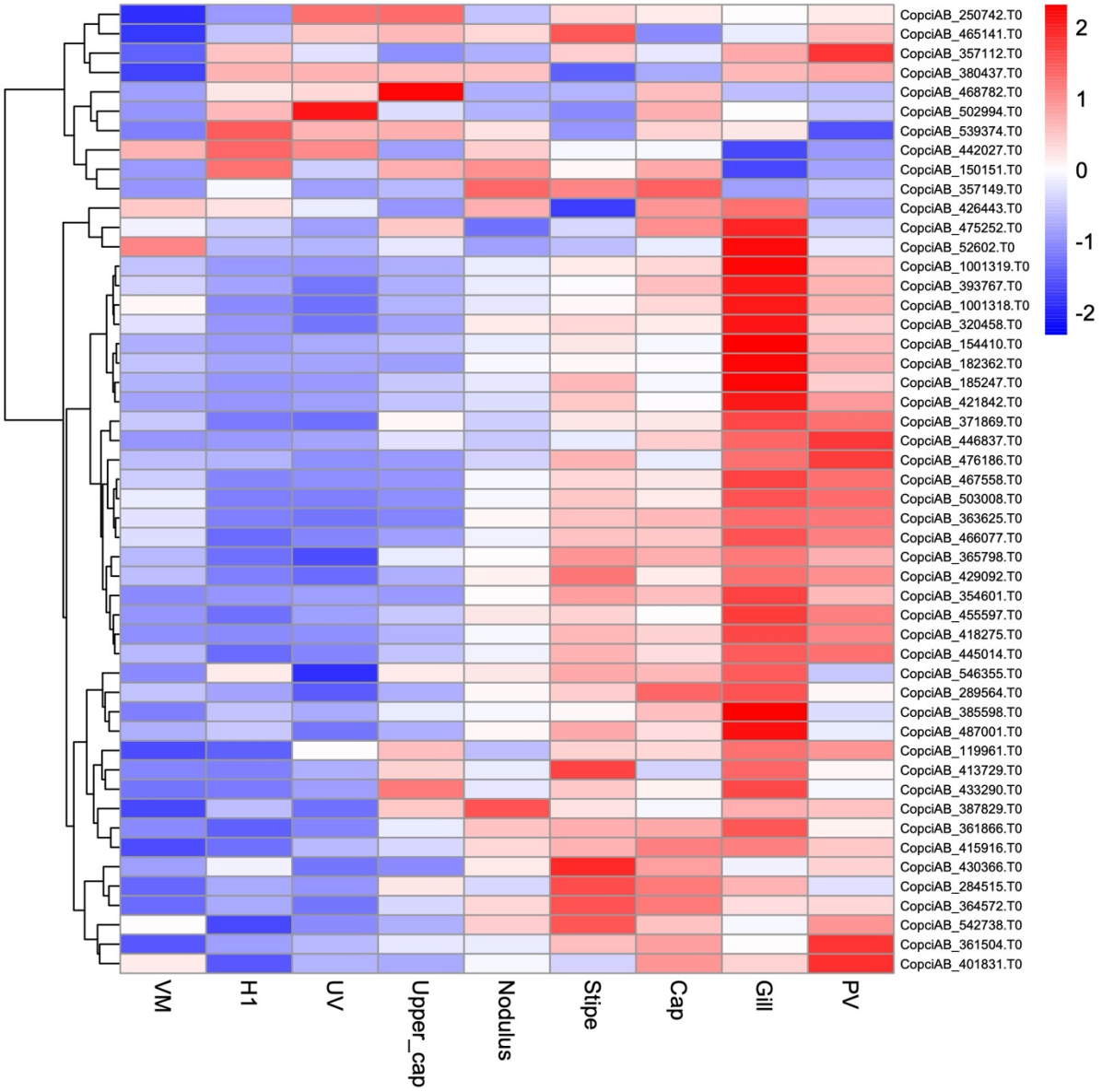

**Supplementary Figure 8.** Expression heatmap for putative cell division (mitosis and meiosis)-related genes across the development and major tissues of *C. cinerea*.



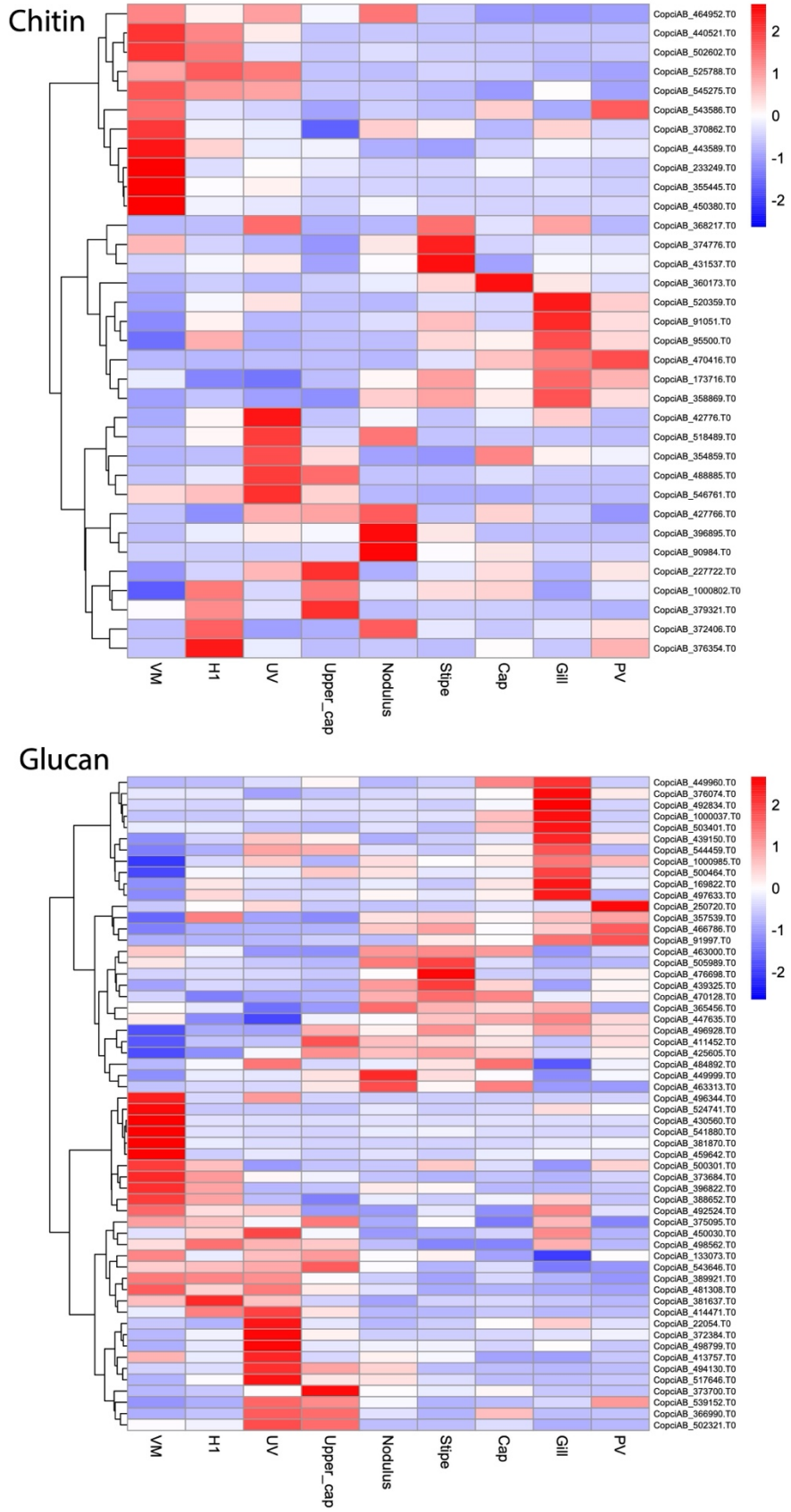

**Supplementary Figure 9.** Expression heatmap for ribosomal protein-encoding genes across the development and major tissues of *C. cinerea*.



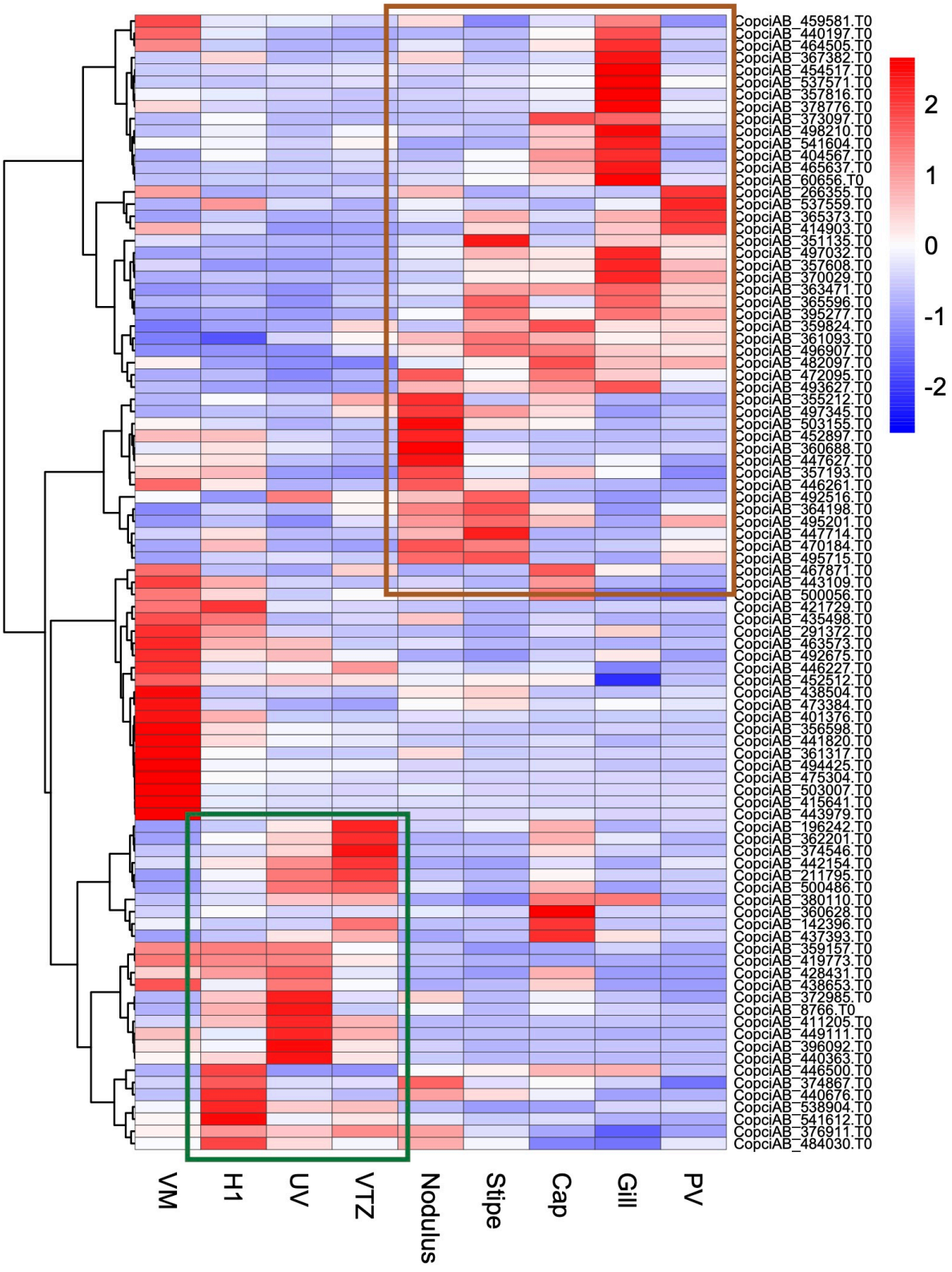

**Supplementary Figure 10.** Expression heatmap for mitochondrial protein-encoding genes across the development and major tissues of *C. cinerea*.



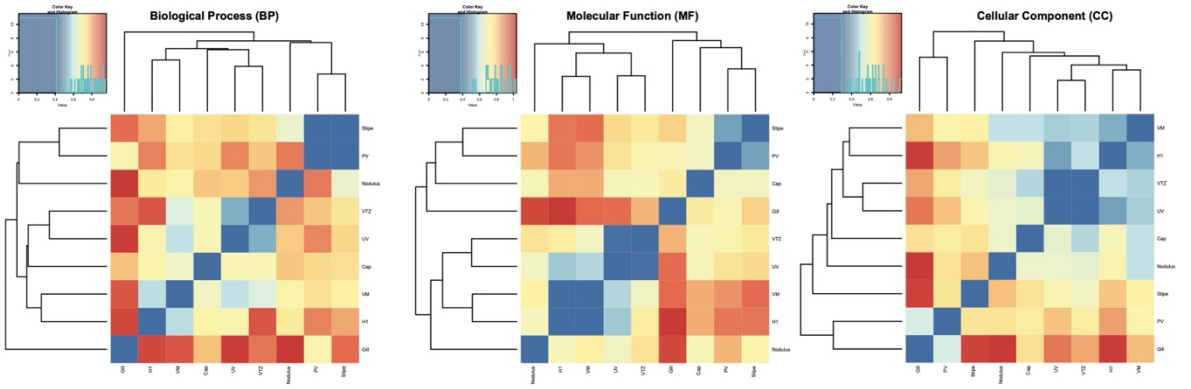

**Supplementary Figure 11.** Expression heatmap for chromatin remodelling-related genes across the development and major tissues of *C. cinerea*.



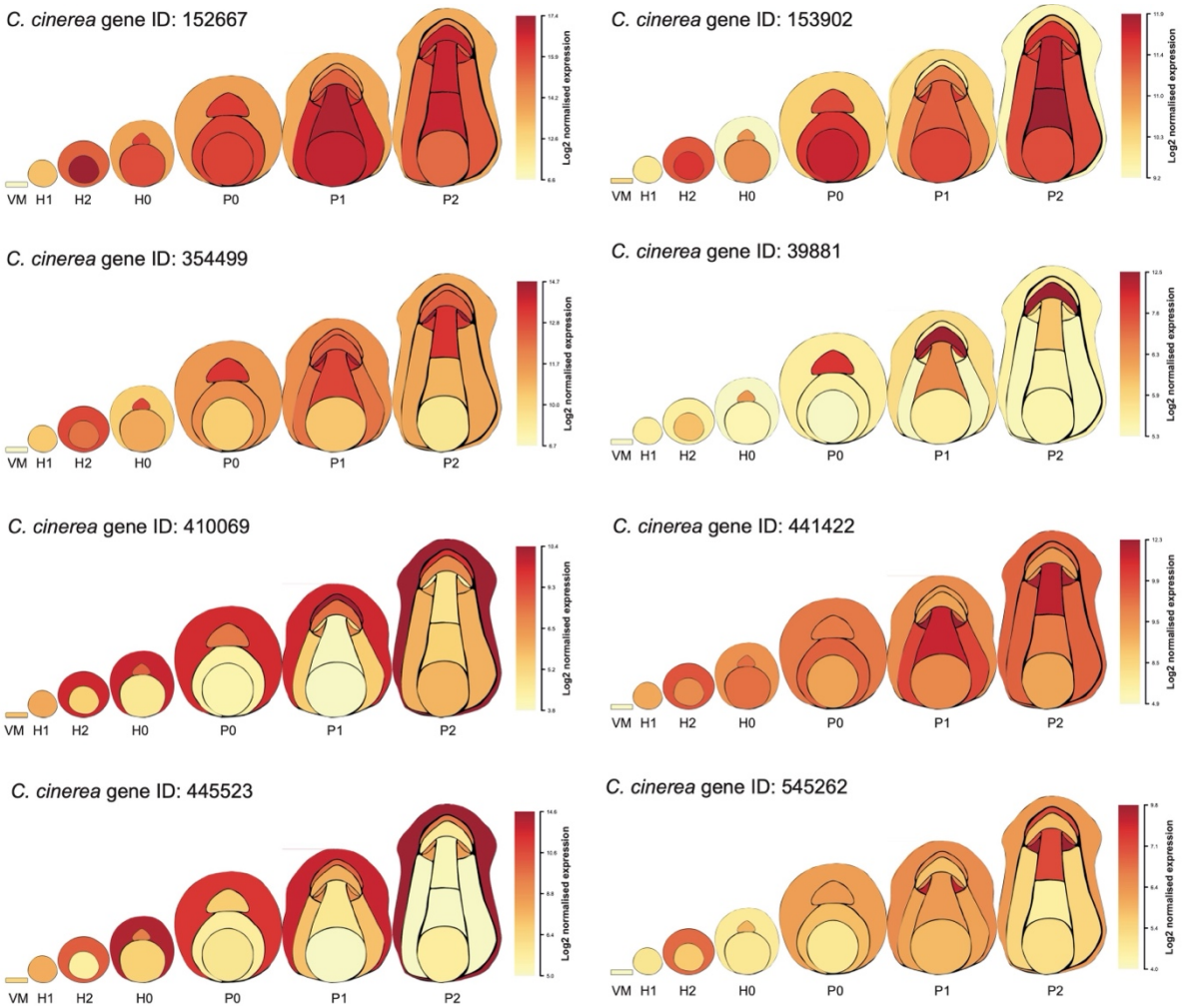

**Supplementary Figure 12.** Expression heatmap for cell wall remodeling genes across the development and major tissues of *C. cinerea*.



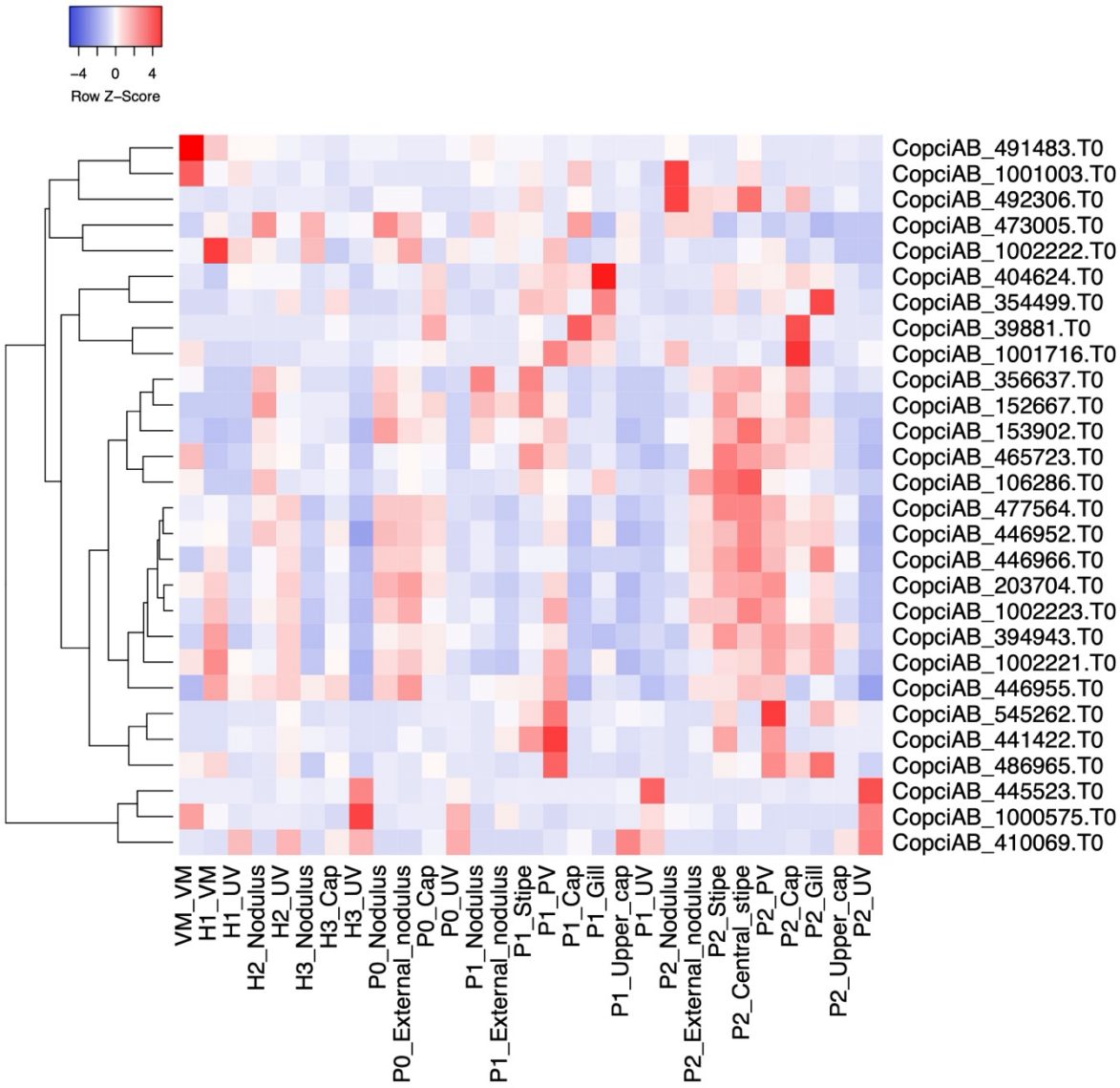

**Supplementary Figure 13.** Expression heatmap for 93 tissue-enhanced or -repressed TFs across the development and major tissues of *C. cinerea*.



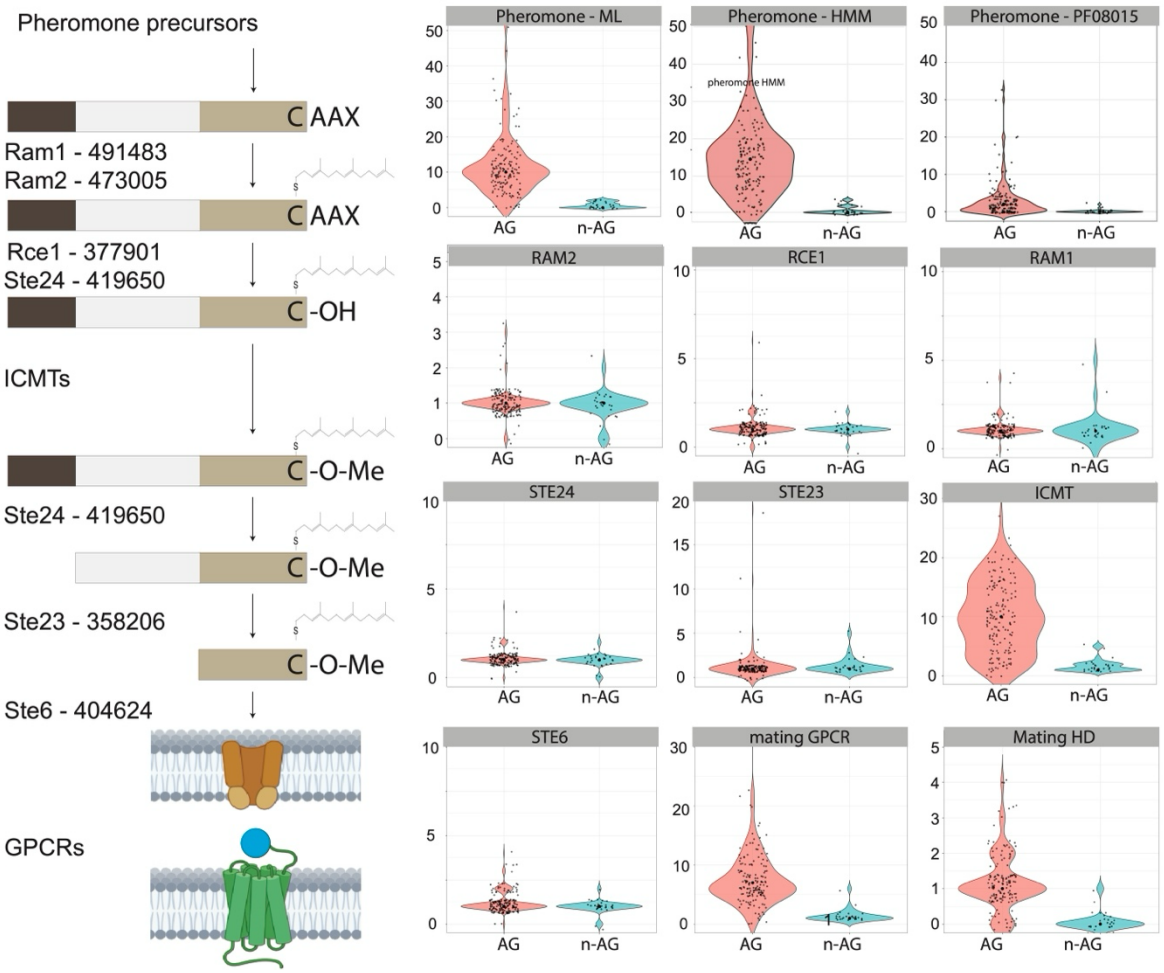

**Supplementary Figure 14.** Functional divergence of main tissues based on GO enrichment among tissue-enhanced genes. Distances were calculated by one minus the Pearson correlation coefficient between all pairs of tissues. Pearson correlation was based on the GO fold enrichment scores.



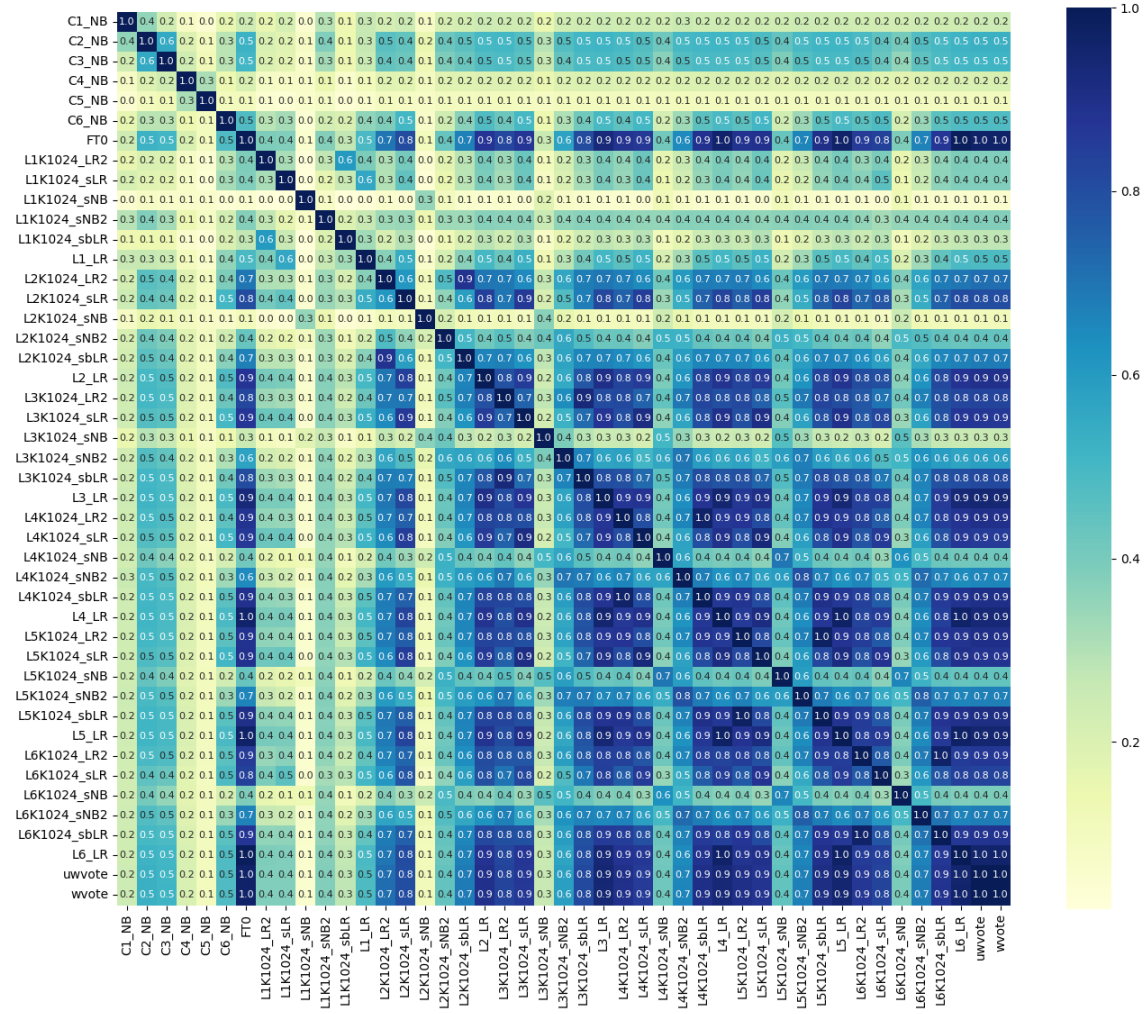

**Supplementary Figure 15.** The spatial expression of the eight Isoprenylcysteine carboxyl methyltransferases (ICMT) paralogues of *C. cinerea*.



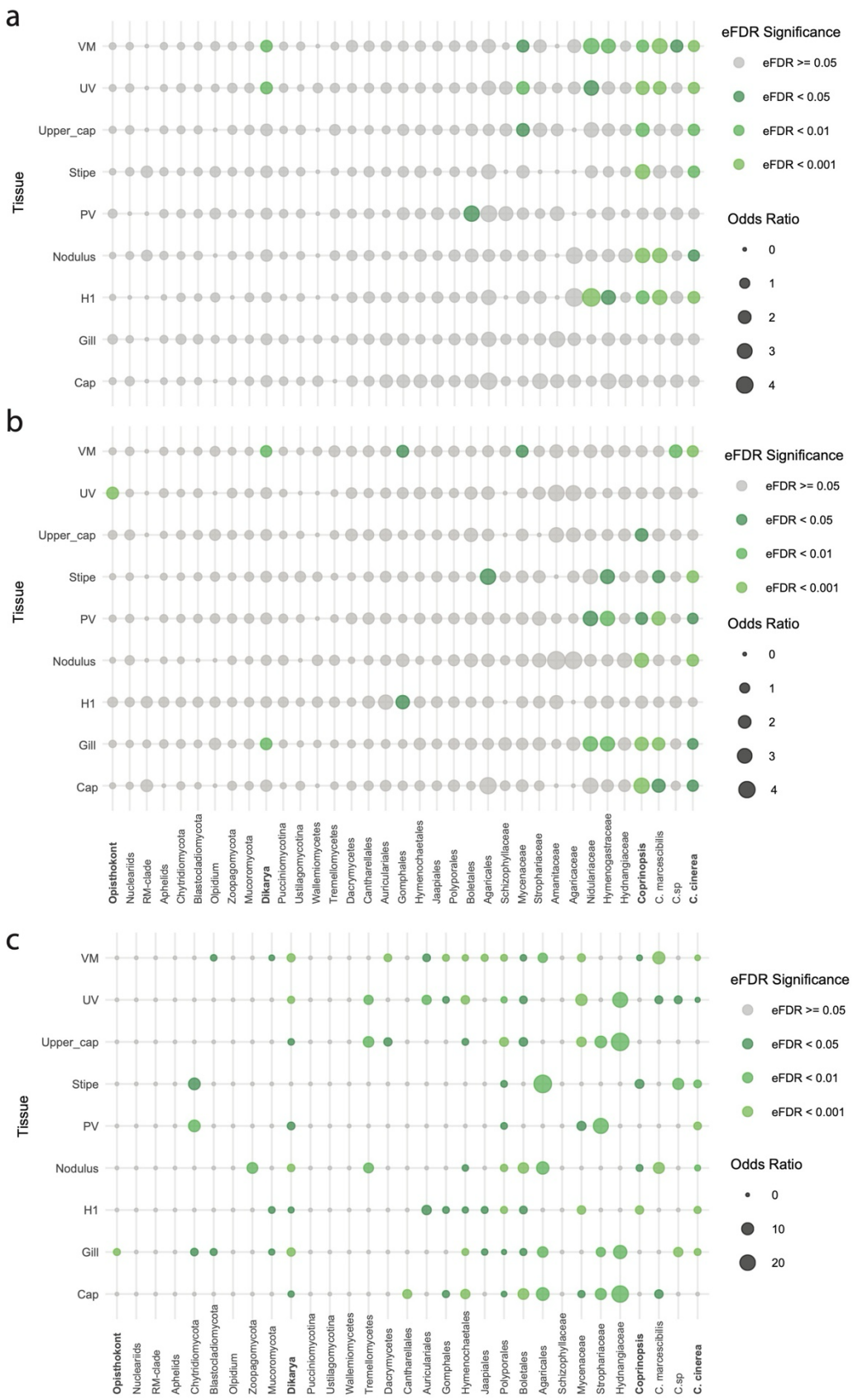

**Supplementary Figure 16.** Expression heatmap of pheromone precursor and GPCR-encoding genes across developmental stages and tissue types.




**Supplementary Figure 17.** Violin plots of ICMTs, pheromones, pheromone-sensing GPCRs, mating HD transcription factors and elements of the pheromone maturation pathway, based on homology searches across 187 fungal species.

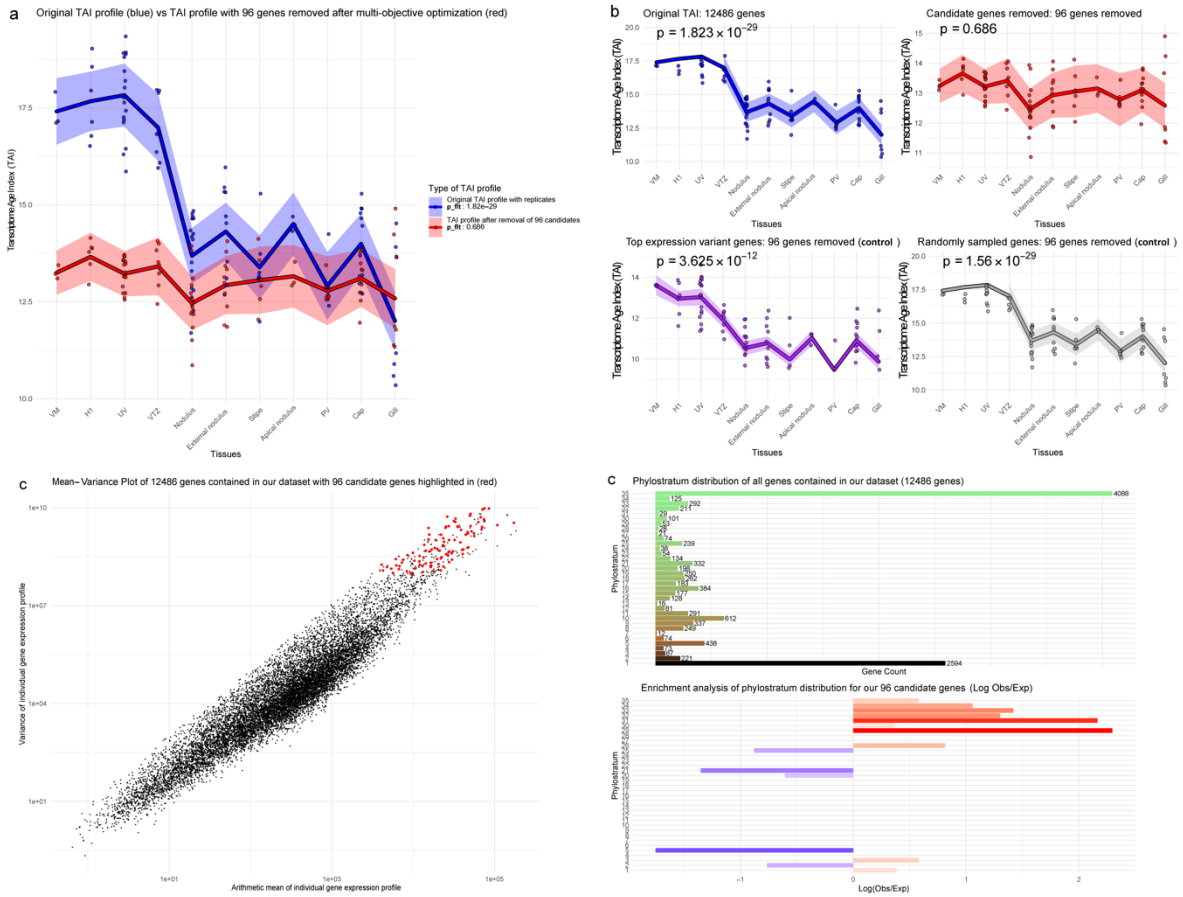



**Supplementary Figure 18.** Heatmap of Jaccard similarities across machine learning methods for predicting mating-type pheromone genes in genome assemblies, calculated by taking the mean of pairwise similarities for each individual species.




**Supplementary Figure 19.** Overrepresentation analysis (ORA, mulea package) of tissue-enhanced and repressed genes across phylostratigraphic layers representing the fungal kingdom. Enrichment of **a.** all tissue-enhanced genes, **b.** all tissue-repressed genes and **c.**

tissue-enhanced transcription factor (TF) genes tested against gene groups defined by their phylostratum (most recent common ancestor, MRCA). Taxon names in bold represent the name of the clade at the respective MRCA, while the other taxonomic names refer to a representative clade that split off from that MRCA. Bubble plot: each dot shows, for a given tissue (rows) and MRCA node (columns), the odds ratio (dot size) of TF enrichment among genes originating at that phylostratum, colored by eFDR significance level.




**Supplementary Figure 20.** Extraction of genes driving the observed Transcriptome Age Index Profile. We performed a multi-objective optimization procedure to extract the minimal set of genes that drives the observed TAI patterns in **Fig. 5** (Methods). Through this inference procedure, we identified 96 genes that can generate a non-significant TAI pattern (flat-line test p-value >0.75) when removed from the overall transcriptome dataset (which contains 12,486 genes overall). **a.** In (blue), TAI profile without any genes removed (full dataset: 12,486 genes). In (red), re-computed TAI profile after removal of our 96 candidate genes from the full dataset (12,486 – 96 = 12,390 genes). **b.** As a control, we also removed 96 genes with the highest variance of their gene expression profile (purple) and 96 randomly sampled genes (grey) to show that, in contrast to our 96 candidate genes, they cannot "destroy" the TAI pattern. **c.** Gene expression mean-variance distribution of all 12,486 genes with 96 candidate genes highlighted in red. **d.** Phylostratum distribution of all 12,486 genes vs phylostratum enrichment of 96 candidate genes. The findings suggest that the 96 candidate genes are largely enriched in evolutionarily young genes compared to the overall dataset background.