***Supplementary Materials***

**Matrix factorization-based biological discovery from large-scale transcriptome data using easyMF**

Wenlong Ma\*, Siyuan Chen\*, Jingjing Zhai, Yuhong Qi, Shang Xie, Minggui Song, Chuang Ma

Corresponding Author: Chuang Ma, State Key Laboratory of Crop Stress Biology for Arid Areas, Center of Bioinformatics, College of Life Sciences, Northwest A&F University, Shaanxi, Yangling 712100, China. Tel.: +86-29-87091109; Email: chuangma2006@gmail.com; cma@nwafu.edu.cn

\*These authors contributed equally to this work.

**Supplementary Figures and Tables are available online at the website of the easyMF project (https://github.com/cma2015/easyMF).**

**Supplementary Table Legends**

**Supplementary Table S1.** Summary of MF-based software tools

**Supplementary Table S2.** Description of functional modules in easyMF

**Supplementary Table S3.** Summary of 940 maize RNA-Seq datasets used in this study

**Supplementary Table S4.** Summary of 75 GO terms used to evaluate the performance of gene prioritization methods

**Supplementary Table S5.** List of 70 experimentally validated genes functionalized in maize seed development

**Supplementary Table S6.** Genome-wide prioritization of candidate seed-related genes using easyMF and MaizeNet

**Supplementary Table S7.** List of signature genes identified from gene expression matrix G1, Gt, and Gs

**Supplementary Table S8.** GO enrichments of signature genes identified from gene expression matrix Gt

**Supplementary Table S9.** GO enrichments of signature genes identified from gene expression matrix Gs

**Supplementary Figure Legends**

**Figure S1.** The bioinformatics pipeline for the generation of a gene expression matrix from RNA-Seq data

**Figure S2.** Identification of signature genes using patternMarkers (A) and Pearson’s correlation coefficient (PCC) algorithm (B)

**Figure S3.** Prioritization of candidate genes involved in a pre-specific function

**Figure S4.** PCA statistics of optimal metagenes. The blue and black dots represent the Cronbach's α value and the explained variance of each metagene, respectively. The red dots represent the cumulative explained variance. At the threshold of Cronbach’s *α* of 0.7, easyMF generated 161 optimal metagenes, capturing 96.4% of the variation in gene expression

**Figure S5.** Identification of cell types of unknown cells from single-cell RNA-Seq data

**Figure S6.** t-Distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction of 4,043 single *Arabidopsis* root cells, which are represented by individual points. All captured cells were clustered into 13 populations corresponding to six cell types

**Supplementary Documents**

**1. Metagene-based pathway activity analysis**

easyMF defined a pathway-level statistic to examine a pathway activity for any gene set of interest. The analysis generally starts with an AM decomposed from a gene expression matrix, and two gene-level statistics are firstly calculated for background and case through Student’s t-test, respectively. To be specific, for background, difference in the distribution of AM coefficients between annotated genes in the pathway and remaining genes in the AM is estimated for each metagene, while a similar procedure was applied to user-specific gene set. The significance of pathway activity is then estimated based on the two gene-level statistics using a paired Student’s t-test, and further adjusted as false discovery rate (FDR) through 1000 permutations.

**2. Metagene-based single-cell transcriptomic data analysis**

Single-cell RNA-Seq, which measures gene expressions at the level of a single cell, has been developed as a powerful technique to investigate the function of individual cells [1]. One of key steps in the process of single-cell transcriptome analysis is to identify cell types of unknown cells. To do so, easyMF builds a bioinformatics pipeline starting from the decomposition of single-cell gene expression matrix to the identification of cell types (Figure S5). In brief, easyMF firstly decomposes the single-cell gene expression matrix (genes in rows and unknown cells in columns) into two matrices (*i.e.*, AM and PM), and then exports into a Seurat object using the R package Seurat v2.3.4 [2]. Then, easyMF performs cell clustering based on the PM using a shared nearest neighbor modularity optimization-based clustering algorithm, as implemented with the ‘FindClusters’ function in Seurat package. Subsequently easyMF characterizes the cell identity of each unknown cell using the Index of Cell Identity (ICI) algorithm, which takes account of prior information of marker genes driven from known tissue expression matrix [3−5]. Finally, easyMF labels unknown cells with different cell types based on cell identities in the corresponding clusters.

easyMF was applied to high-throughput single-cell RNA sequencing data of >4000 individual cells (GEO accession number: GSE116614) from *Arabidopsis* root [3]. easyMF grouped 4043 unknown cells into 13 clusters. According to the gene expression matrix generated from 69 samples (collected manually from NCBI GEO database; corresponding 11 known tissues), easyMF generated a Spec score matrix through ‘getAllSpec’ function, in which each gene has 11 Spec scores for corresponding known tissues. Subsequently, easyMF identified marker genes for each known tissue based on the Spec score matrix through ‘getMarkerList’ function. Finally, easyMF generated 11 ICI scores for each unknown cell with the ‘getIdentity’ function using the 4043 unknown cells’ expression matrix, the Spec score matrix and these identified marker genes. These functions were predefined by [5], and have been incorporated in easyMF. The ICI score from a known tissue *t* is calculated using the following formula:

,

where is the number of marker genes for tissue *t*, represents the unknown single cell expression level of marker gene *g*, represents the unknown single cell expression status of marker gene *g* ( = 1, if > 0; otherwise = 0), and represents the Spec score of marker gene *g* at tissue *t*. is calculated by a law of total probability:

,

where *P*(*x|t*) represents conditional probability of gene expression level *x* at given tissue *t*, and the *P*(*t|x*) represents conditional probability of the tissue *t* given at gene expression level *x*.

**Supplemental References**

[1] Tang F, Barbacioru C, Nordman E, Li B, Xu N, Bashkirov VI, *et al*. RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat Protoc* 2010;**5**:516−35.

[2] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;**36**:411−20.

[3] Shulse CN, Cole BJ, Ciobanu D, Lin J, Yoshinaga Y, Gouran M, *et al*. High-throughput single-cell transcriptome profiling of plant cell types. *Cell Rep* 2019;**27**:2241−7 e4.

[4] Birnbaum KD, Kussell E. Measuring cell identity in noisy biological systems. *Nucleic Acids Res* 2011;**39**:9093−107.

[5] Efroni I, Ip PL, Nawy T, Mello A, Birnbaum KD. Quantification of cell identity from single-cell gene expression profiles. *Genome Biol* 2015;**16**:9.