Supplementary Data for

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

Wenlong Ma1,2,4, †, Siyuan Chen1,2,5, †, Yuhong Qi1,2, †, Minggui Song1, Jingjing Zhai1,2, Ting Zhang1,2, Shang Xie1, Guifeng Wang3, Chuang Ma1,2, \*

†These authors have contributed equally to this work.

\*Corresponding author: chuangma2006@gmail.com

**Supplemental Materials and Methods**

**1. Bioinformatics pipeline for generating maize gene expression matrix from raw reads**

easyMF presents a customized bioinformatics pipeline to generate gene expression matrix from raw RNA-Seq reads (**Supplementary Figure S1**). This bioinformatics pipeline has been applied to generate the maize gene expression matrix G1. In brief, 1,066 maize RNA-Seq datasets were firstly collected from NCBI’s Gene Expression Omnibus (GEO) and/or Sequence Read Archive (SRA) databases (as of 26 July 2019). Raw RNA-Seq data were preprocessed using fastp (version 0.20.0) [1] for quality control, including sequencing adapter trimming and low-quality read filtering. Subsequently, high-quality RNA-Seq reads from each sample were aligned to maize reference genome (APGv4, https://plants.ensembl.org/Zea\_mays/Info/Index) using HISAT2 (version 2.1.0) [2], generating a BAM (binary alignment map) file recording read-genome alignments. BAM files were then used as inputs of StringTie (version 1.3.6) [3] to estimate gene expression abundance in terms of FPKM (fragments per kilobase of transcript per million mapped reads).

****

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

**2. Generation of the maize gene expression matrix G1**

Using the customized bioinformatics pipeline, 1,066 RNA-Seq datasets from maize B73 samples manually collected from the NCBI GEO and SRA databases were firstly processed to generate a raw gene expression matrix. To obtain a high-quality gene expression matrix, a two-step quality control was implemented to filter genes expressed at low levels and remove outlier samples. For expression-level quality control, genes with FPKM ≥ 1 in at least 15 RNA-Seq samples were retained. For low-quality samples, we firstly averaged the statistical duplicated samples based on PCC with criteria: PCC > 0.999 [4]. Then, outlier samples whose correlation with the first principal component (sample based principal component analysis) less than 0.75 were removed. Finally, a gene expression matrix G1 with 28,874 protein-coding genes and 940 samples was obtained for the downstream application.

**3. Identification of signature genes**

easyMF decomposes a high-dimensional gene expression matrix (genes in rows and samples in columns) into a product of two low-dimensional metagene-based matrices: an amplitude matrix (AM; genes in rows and metagenes in columns) and a pattern matrix (PM; metagenes in rows and samples in columns). Using gene-level relationships in the AM and sample-level relationships in the PM, easyMF identifies genes exhibiting dominant patterns (defined as signature genes) for each metagene using the patternMarkers [5] and the Pearson’s correlation coefficient (PCC) statistics. patternMarkers calculates a Euclidean distance (*ED*) between normalized AM coefficients (i.e., coefficients in the AM) and a 0-1 pattern of metagenes (**Supplementary Figure S2A**). Suppose the number of metagenes is *M*, the *ED* score for gene *i* and metagene *k* (1≤ *k* ≤ *M*) is calculated using following formula:

,

where *Aij* represents the AM coefficient of gene *i* in metagene *j* (1≤ *j* ≤ *M*), *maxAi* denotes the maximum value of AM coefficient of gene *i* amongall *M* metagenes, is a numeric unit vector specifying the status of each component ( = 1 only when *k* = *j*, otherwise equals to 0). For each gene *i*, easyMF repeats this process to generate a vector of *ED* scores for all metagenes.

easyMF uses the PCC statistic to quantify the correlation between gene expression abundance and PM coefficients (**Supplementary Figure S2B**). For gene *i* and metagene *k*, the PCC score is calculated using the following formula:

Where represents the *i*-th gene’s expression values, represents the variance of *i*-th gene’s expression values, represents the *i*-th gene’s PM coefficients, represents the variance of *i*-th gene’s PM coefficients. The gene *i* is regarded as a signature gene of metagene *k*, if it satisfied with three conditions: i) equals to the minimal ED score; ii) PCC ≥ 0.6; iii) *P*-value ≤ 1.0E-03. Of note, the thresholds of PCC and *P*-value can be user-adjusted in web interface of easyMF.



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

**4. Metagene-based gene prioritization**

For a given set of genes (denoted as labeled genes), easyMF firstly examines the difference in the distribution of AM coefficients between labeled genes and unlabeled (all except labeled genes in the AM) genes by using Student’s t-test, following by a transformation of the significance level *P*-value to *z*-score using the standard normal quantile function ‘qnorm’ in R. A higher *z*-score indicates a larger difference in the AM coefficient between labeled and unlabeled genes, thus corresponding to stronger biological association between the metagene and the gene set. This results to a *z*-score vector with a length of metagene number for the given gene set. Subsequently, the association between *z*-scores and AM coefficients of corresponding genes is examined using the PCC statistic [4]. Finally, easyMF prioritizes candidate genes functionally associated with the given gene set based on the decreasing PCC values (**Supplementary Figure S3**).



**Supplementary Figure S3**. Graphic representation of easyMF for prioritizing candidate genes involved in a pre-specific function.

**4. Performance evaluation of gene prioritization approaches**

The leave-one-out cross-validation (LOOCV) experiment was used to evaluate the performance of easyMF and MaizeNet [6] in gene prioritization. In LOOCV experiment, each labeled gene and all unlabeled genes were used as testing samples and their normalized scores (min-max normalization) were then calculated. The performance of easyMF and MaizeNet were further evaluated using the area under the receiver operating characteristic (ROC) curve (AUC) and the area under the curve of self-ranked curve (AUSR). The ROC curve is a plot of true-positive rate (TPR) along the y axis versus false-positive rate (FPR) along the x axis. While the self-rank curve is a plot of ratio (*Ra*)along the *y* axis versus self-rank along the *x* axis [7], the *Ra* can be calculated by the following formula:

,

Where represents the ranks of all positive genes, represents the ratio of ranks lower than a pre-defined level of *l* (e.g., 1000). Both AUC and AUSR, ranging from 0 to 1, were finally calculated using the trapezoid rule [8], with greater value indicating better prediction performance.

**Supplementary References**

[1] S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics. 34 (2018) i884–i890.

[2] D. Kim, B. Langmead, S.L. Salzberg, HISAT: a fast spliced aligner with low memory requirements, Nat. Methods. 12 (2015) 357–360.

[3] M. Pertea, G.M. Pertea, C.M. Antonescu, T.-C. Chang, J.T. Mendell, S.L. Salzberg, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads, Nat. Biotechnol. 33 (2015) 290–295.

[4] R.S.N. Fehrmann, J.M. Karjalainen, M. Krajewska, H.-J. Westra, D. Maloney, A. Simeonov, T.H. Pers, J.N. Hirschhorn, R.C. Jansen, E.A. Schultes, H.H.H.B.M. van Haagen, E.G.E. de Vries, G.J. te Meerman, C. Wijmenga, M.A.T.M. van Vugt, L. Franke, Gene expression analysis identifies global gene dosage sensitivity in cancer, Nat. Genet. 47 (2015) 115–125.

[5] G.L. Stein-O’Brien, J.L. Carey, W.S. Lee, M. Considine, A.V. Favorov, E. Flam, T. Guo, S. Li, L. Marchionni, T. Sherman, S. Sivy, D.A. Gaykalova, R.D. McKay, M.F. Ochs, C. Colantuoni, E.J. Fertig, PatternMarkers & GWCoGAPS for novel data-driven biomarkers via whole transcriptome NMF, Bioinformatics. 33 (2017) 1892–1894.

[6] T. Lee, S. Lee, S. Yang, I. Lee, MaizeNet: a co-functional network for network-assisted systems genetics in *Zea mays*, Plant J. 99 (2019) 571–582.

[7] O. Tzfadia, D. Amar, L.M.T. Bradbury, E.T. Wurtzel, R. Shamir, The MORPH algorithm: ranking candidate genes for membership in *Arabidopsis* and tomato pathways, Plant Cell. 24 (2012) 4389–4406.

[8] P. Radivojac, W.T. Clark, T.R. Oron, A.M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, G. Pandey, J.M. Yunes, A.S. Talwalkar, S. Repo, M.L. Souza, D. Piovesan, R. Casadio, Z. Wang, J. Cheng, H. Fang, J. Gough, P. Koskinen, P. Törönen, J. Nokso-Koivisto, L. Holm, D. Cozzetto, D.W.A. Buchan, K. Bryson, D.T. Jones, B. Limaye, H. Inamdar, A. Datta, S.K. Manjari, R. Joshi, M. Chitale, D. Kihara, A.M. Lisewski, S. Erdin, E. Venner, O. Lichtarge, R. Rentzsch, H. Yang, A.E. Romero, P. Bhat, A. Paccanaro, T. Hamp, R. Kaßner, S. Seemayer, E. Vicedo, C. Schaefer, D. Achten, F. Auer, A. Boehm, T. Braun, M. Hecht, M. Heron, P. Hönigschmid, T.A. Hopf, S. Kaufmann, M. Kiening, D. Krompass, C. Landerer, Y. Mahlich, M. Roos, J. Björne, T. Salakoski, A. Wong, H. Shatkay, F. Gatzmann, I. Sommer, M.N. Wass, M.J.E. Sternberg, N. Škunca, F. Supek, M. Bošnjak, P. Panov, S. Džeroski, T. Šmuc, Y.A.I. Kourmpetis, A.D.J. van Dijk, C.J.F. ter Braak, Y. Zhou, Q. Gong, X. Dong, W. Tian, M. Falda, P. Fontana, E. Lavezzo, B. Di Camillo, S. Toppo, L. Lan, N. Djuric, Y. Guo, S. Vucetic, A. Bairoch, M. Linial, P.C. Babbitt, S.E. Brenner, C. Orengo, B. Rost, S.D. Mooney, I. Friedberg, A large-scale evaluation of computational protein function prediction, Nat. Methods. 10 (2013) 221–227.