Science Bulletin 67 (2022) 903–906

Contents lists available at ScienceDirect

Science Bulletin

journal homepage: www.elsevier.com/locate/scib

Correspondence

MODAS: exploring maize germplasm with multi-omics data association studies

Songyu Liu [a,b,1], Feng Xu [a,1], Yuetong Xu [a], Qian Wang [a], Jun Yan [a], Jinyu Wang [a], Xianbing Wang [b], Xiangfeng Wang [a,*]

[a] National Maize Improvement Center, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100094, China

[b] State Key Laboratory of Agro-Biotechnology, College of Biological Sciences, China Agricultural University, Beijing 100094, China

Exploiting crop germplasm with multi-omics data can greatly enhance the power of gene discovery and interpret the genetic relations of genes. Here, we present MODAS (Multi-Omics Data Association Studies) software, freely available at https://modas-bio.github.io/, to cope with high dimensional, noisy, and heterogeneous features typical of multi-omics data with advanced machine learning and statistical methods. MODAS features four analytical modules with extraordinary computing efficiency (Fig. S1 online). It first performs dimensionality reduction (DR) on genotypic data and generates a pseudo-genotype index file representing a simplified atlas of whole-genome variation of a studied population. The index file is then mainly used for an initial screening of biologically-meaningful molecular traits (mTraits), such as mRNA transcripts and metabolic compounds, showing significant association with genomic variation. With the second module, MODAS performs two steps of regional association (RA) analyses and one DR to reduce redundancy across mTraits. During the third module, MODAS performs expression GWAS (eGWAS) and metabolome GWAS (mGWAS) using the mTraits identified above, followed by visualization of GWAS signals and integration of annotation information for the genes within the candidate interval in a web-based user-friendly interface. In the fourth module, MODAS applies the Mendelian randomization (MR) algorithm on the summarized RA results to infer causal relations between transcription factors (TFs) and target genes, gene expression or metabolic compounds and phenotypic traits. The inferred relations then may facilitate biologists in formulating a hypothesis for molecular validation.

Millions of single nucleotide polymorphisms (SNPs) are too excessive to perform efficient association analysis with multi-omics data. We therefore first applied MODAS to reduce the complexity of the genotypic data and generate a simplified pseudo-genotype index that nevertheless captures the genomic variation across samples (Supplementary material online). MODAS first scanned the genome in 1-Mbp sliding windows with a 0.5-Mbp step and calculated Jaccard similarity coefficients [1] to measure the genotypic similarity between all pairs of SNPs within the window. Determination of the proper sizes of sliding window and step may be based on the distance of linkage disequilibrium (LD) decay of the studied population analyzed by the PopLDdecay software (Fig. S2, Supplementary materials online) [2]. It then applied the clustering algorithm DBSCAN (density-based spatial clustering of applications with noise) [3] to the resulting genotype similarity matrix to generate genomic blocks with clustered SNPs. Finally, principal component analysis (PCA) was performed on the genotypes of clustered SNPs within each block and selected the first principal component (PC1) for each block to represent the overall genomic variation across the population (Supplementary materials). As Fig. S3 (online) shows, PC1 contributions may explain 80% to 100% of the variances for the majority of the blocks. The genome is then partitioned into tens of thousands of genomic blocks used for the subsequent analysis, from which the pseudo-genotype index file was derived for screening of mTraits in the second module (Fig. 1a).

Multi-omics datasets are large, generally noisy, and possibly highly redundant. Without appropriate data filtration, GWAS results may include a high fraction of false-positive signals (Fig. S4 online). MODAS performs two steps of RA analysis for initial screening of mTraits independently of either differential or correlation analysis (Supplementary materials online). The first step implements a mixed linear model (MLM) to identify mTraits significantly associated ($P \leq 1 \times 10^{-6}$) with the genomic blocks, with consideration of the kinship matrix between samples [4]. In the second step, all SNPs within a genomic block showing association with a mTrait are extracted from the original genotype data and used as input to rerun MLM to determine the exact boundaries of the interval with significant association to that mTrait, followed by summarizing candidate quantitative trait loci (QTLs) for the subsequent analysis (Supplementary materials online).

MODAS adopts another DR step to remove data redundancy embedded in mTraits (Supplementary materials online). Redundancy is inherent not only to omics-type approaches but also to the nature of biological pathways. For example, metabolic gene clusters (MGCs) commonly exist in plant genomes, encoding enzymes that catalyze enzymatic reactions in the same metabolic pathway [5]. Thus, one genomic region might be identified repeatedly as associated with intermediate compounds and/or the final

[*] Corresponding author.
E-mail address: xwang@cau.edu.cn (X. Wang).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.scib.2022.01.021

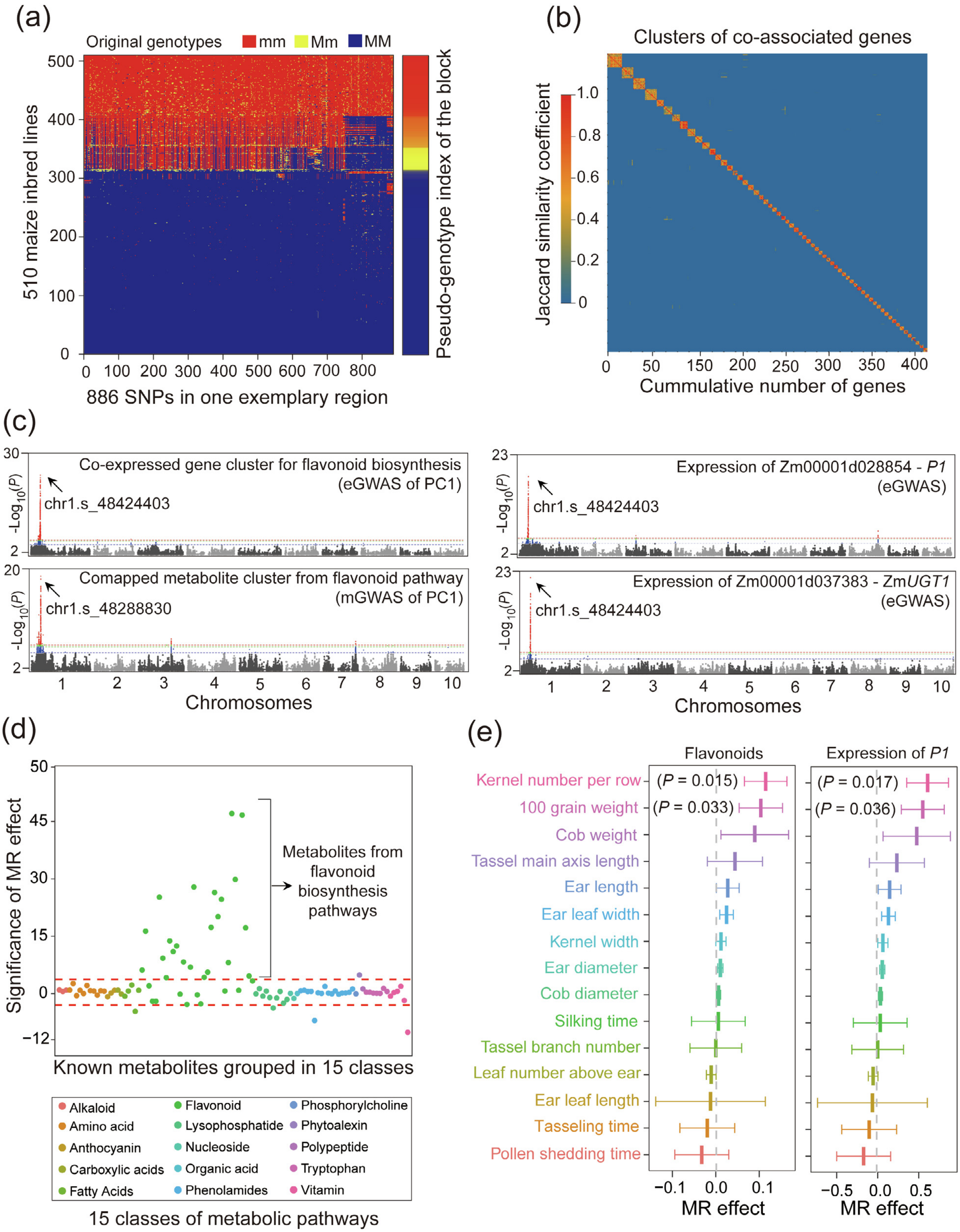

Fig. 1. Four major analytical modules of MODAS. (a) Module of dimensionality reduction. The pseudo-genotype index reflects the overall genotypic variation of an exemplary region (Chr1: 142.0 to143.5 Mb). "M" and "m" represent the major and minor alleles of the 886 SNPs, respectively. (b) Module of regional association (RA) analysis. The heatmap illustrates the 62 clusters of co-associated genes identified by RA analysis. (c) Module of expression GWAS (eGWAS) and metabolome GWAS (mGWAS). *P1* gene is located in the genomic region peaked at Chr1: s_48424403. Zm*UGT1* gene is located at Chr6: 123,804,940 to 123,806,574, regulated by *P1* in *trans*. (d) Module of Mendelian randomization (MR). Causal relations between the *P1* gene and 135 known metabolites were inferred. The *y* axis is $-\log_{10}(P)$, in which the *P* represents the statistical significance of the MR effect estimated by χ^2 test. The red dashed lines represent the significance threshold ($P = 1 \times 10^{-5}$) for positive and negative MR effects. (e) MR-based estimation of the contribution of flavonoid levels and *P1* expression levels to the 15 agronomic traits of maize.

product of the pathway. Additionally, if there is crosstalk between different pathways, one metabolic product might be associated with multiple genomic regions, together making mGWAS results difficult to interpret. MODAS applied a PCA-based DR step to all mTraits within each block separately, thus reducing the matrix of compounds × samples to one dimension. The resulting PC1 values are used for GWAS instead of the absolute mTrait values. We illustrate this functionality here with metabolome data. After RA analysis, six and eight metabolites showed significant associations with two genomic blocks—blocks 40,619 and 58,979, respectively (Fig. S5a online). The abundances of the compounds within each block were highly correlated, suggesting that their respective metabolism-related genes may lie in these blocks, most likely encoding enzymes acting early in the biosynthetic pathway (Fig. S5b online). We then applied a PCA-based DR step to all compounds within each block separately, thus reducing the matrix of compounds × samples to one dimension. Finally, we performed GWAS with the two resulting PC1 values and identified two genomic regions with a strong association with each PC1 value on chromosomes 6 and 10 (Fig. S5c online), corresponding to blocks 40,619 and 58,979, respectively. The RA and DR steps above therefore identified a subset of metabolic compounds that can be subsequently submitted to any GWAS software to detect associated regions harboring their metabolism-related genes and generate Manhattan plots.

One genetic variant may not only influence the expression of nearby genes in *cis*, but also that of distal genes in *trans*, and will be reflected by co-association of such genes to the same polymorphic genomic region. As co-associated genes may be functionally related, we adopted a clustering strategy to identify co-associated genes from transcriptome data (Supplementary materials online). After RA analysis, MODAS generated an association matrix of filtered genes × peak SNPs, followed by calculating Jaccard similarity coefficients between all possible gene pairs, and clustering genes significantly associated with the same peak SNPs by DBSCAN (Fig. 1b). We hypothesized that this approach would allow the discovery of genes with correlated expression patterns resulting from the same genomic variants acting both in *cis* and in *trans*, which also forms the biological basis to infer the causal relationships between genes, metabolites, and traits with the MR algorithm.

MR has been applied to infer causality among genetic variants, risk factors, and common human disease based on GWAS summary data [6–8]. Because metabolite contents reflect the expression of genes in the relevant biosynthesis, MR may also be applicable for inferring the causality between genes and metabolites and estimating their contributions to traits. We explored this possibility with regard to the well-studied flavonoid biosynthesis pathway of maize. GWAS analysis of co-associated genes (eGWAS) and clustered metabolites (mGWAS) identified the same genomic region on the short arm of chromosome 1 (Fig. 1c, left two panels). The abundance of apigenin, hesperetin, and quercetin, the products from the flavonoid biosynthesis pathway [9], also showed an association with this genomic region. Thus, we reasoned that genes involved in flavonoid biosynthesis may reside in this region. Indeed, this region contained the two previously reported *Pericarp color1* (*P1*), which encodes an R2R3-MYB domain TF) regulating pigmentation of maize kernel [10]. Additional flavonoid biosynthetic genes showed association with the same region, including *FNS1* (*Flavone synthase1*), *A1* (*Anthocyaninless1*), *C2* (*Colorless2*), *PR1* (*Purple aleurone1*), and *UGT1* (*UDP-glucosyl transferase1*) [11], but they were located on different chromosomes (Fig. 1c, right two panels). Thus, flavonoid biosynthesis provides a typical example of *cis*-acting variants affecting *P1* expression and *P1*-regulated genes in *trans*.

The causal relations of co-mapped *P1* and flavonoids may be quantitatively inferred with the MR method (Supplementary materials online). Out of the 983 compounds profiled in maize kernels, 135 are known metabolites derived from 15 metabolic pathways. We extracted an MR effect from our MR analysis representing the statistical significance, and thus the degree of causality, between each metabolite and the peak SNP (Chr1: s_48424403) associated with *P1* expression. Of the 135 metabolites, 27 passed the threshold; of those, 23 were products from the flavonoid biosynthetic pathway (Fig. 1d), in agreement with the reported role of *P1* [10].

We finally applied the MR method to estimate the contributions of flavonoids and their biosynthetic genes to agronomic traits. We first tested the causality of *P1* and flavonoids: they exhibited positive causal effects on yield-related traits with very similar patterns, such as kernel number per row (KNPR), cob weight, 100-grain weight (100 GW), and ear length, with flavonoid levels and *P1* expression contributing significantly to KNPR and 100GW (Fig. 1e). This implies that *P1*-regulated flavonoid biosynthesis may be involved in grain yield, perhaps through effects on KNPR.

We identified two genomic regions on chromosomes 5 and 8 with MODAS (Fig. S6a online), which contain six pairs of TF and autophagy genes (Table S1 online). These two regions fall into two QTLs previously mapped by traditional linkage analysis, but the causative genes remain uncharacterized due to the long list of candidate genes. From each region, we selected a pair of TF and gene with MR effects >30.0 and $P < 1 \times 10^{-7}$ to validate if interactions occur between the TF and targets with yeast one-hybrid (Y1H) assay (Supplementary materials, Table S2 online). The two pairs of TF and target are Zm00001d018258 (Zm*MYBR67*, *MYB-related transcription factor 67*) and Zm00001d018259 (Zm*ATG12*, *Autophagy-related 12*) in the QTL of *qtl_ch5-217.5mb*, and Zm00001d012015 (Zm*SBP18*, *SBP-transcription factor 18*) and Zm00001d011984 (Zm*ATG6b*, *Beclin-1-like protein*) in the QTL of *qtl_ch8-166.2mb*. The Y1H result showed that the two TFs Zm*MYBR67* and Zm*SBP18* can directly bind to the promoter regions of Zm*ATG12* and Zm*ATG6b*, respectively, thus worthy of further investigation (Fig. S6b online).

Previously published tools, such as Mergeomics, SMR, GSMR, and CoMM, mostly infer genetic regulation based on GWAS summary data [12–15]. In contrast, MODAS covers the whole procedure of analyzing population-scale multi-omics data, including dimensionality-reduction on genotypic data, screening of biologically-meaningful mTraits through a two-step RA analysis, and MR-based inference of causal relationship from summary data. MODAS also includes a visualization module to browse Manhattan plots and gene annotations of significantly associated regions identified by GWAS. After manual curation of identified regions, if necessary, MODAS combines all Manhattan plots into a single one to visually integrate GWAS signals. These analytical modules may be designed as a streamlined pipeline in MODAS, featuring extraordinary computing efficiency. To test the software performance, we run the MODAS pipeline on an RNA-Seq dataset containing expression data of 16,000 genes profiled from 510 maize samples. MODAS accomplished the six steps of analyses with only 5.7 h on a desktop server (Table S3 online).

GWAS with common agronomic traits has reached a bottleneck, due to innate limitations when dissecting complex, polygenic traits. Multi-omics analysis of a core crop germplasm may greatly enhance the resolution of gene mapping, and improve the chance of identifying causative genes for experimental validation. MODAS adopts novel strategies and algorithms to reduce data complexity and vastly accelerate computing efficiency for multi-omics data association analysis in maize. It also has the potential of extending to other plant species with necessary changes on some of the key software parameters based on the features of the genome and population of a studied plant. MODAS will expedite the discovery of agronomically important genes from plant germplasm in the era of omics.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the Key Research and Development Program of China (2018YFA0901000 and 2018YFA0901003), the National Natural Science Foundation of China (31871706), and the China Postdoctoral Science Foundation (2020TQ0355).

Appendix A. Supplementary materials

Supplementary materials to this article can be found online at https://doi.org/10.1016/j.scib.2022.01.021.

References

[1] Levandowsky M, Winter D. Distance between sets. Nature 1971;234:34–5.

[2] Zhang C, Dong SS, Xu JY, et al. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics 2019;35:1786–8.

[3] Hahsler M, Piekenbrock M, Doran D. dbscan: fast density-based clustering with R. J Stat Softw 2019;91:1–30.

[4] Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet 2012;44:821–4.

[5] Schläpfer P, Zhang P, Wang C, et al. Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. Plant Physiol 2017;173:2041–59.

[6] Zhu Z, Zheng Z, Zhang F, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. Nat Commun 2018;9:224.

[7] Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet 2016;48:481–7.

[8] Wu Y, Zeng J, Zhang F, et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. Nat Commun 2018;9:918.

[9] Yonekura-Sakakibara K, Higashi Y, Nakabayashi R. The origin and evolution of plant flavonoid metabolism. Front Plant Sci 2019;10:943.

[10] Grotewold E, Drummond BJ, Bowen B, et al. The myb-homologous P gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset. Cell 1994;76:543–53.

[11] Morohashi K, Casas MI, Ferreyra LF, et al. A genome-wide regulatory framework identifies maize *Pericarp Color1* controlled genes. Plant Cell 2012;24:2745–64.

[12] Shu Le, Zhao Y, Kurt Z, et al. Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems. BMC Genomics 2016;17:874.

[13] Ding J, Blencowe M, Nghiem T, et al. Mergeomics 2.0: a web server for multi-omics data integration to elucidate disease networks and predict therapeutics. Nucleic Acids Res 2021;49:W375–87.

[14] Ong JS, MacGregor S. Implementing MR-PRESSO and GCTA-GSMR for pleiotropy assessment in Mendelian randomization studies from a practitioner's perspective. Genet Epidemiol 2019;43:609–16.

[15] Yang Yi, Yeung KF, Liu J. CoMM-S^4: a collaborative mixed model using summary-level eQTL and GWAS datasets in transcriptome-wide association studies. Front Genet 2021;12:704538.

Songyu Liu received his B.S. degree in Life Sciences from Northwest A&F University in 2015. At present, he is a Ph.D. candidate of Bioinformatics in the College of Biological Sciences at China Agricultural University. His project is focused on developing software for association analysis with multi-omics data, and novel methods to infer gene regulation networks from omics big data in plants.

Feng Xu received his B.S. degree in Bioinformatics from Chongqing University of Posts and Telecommunications in 2015. At present, he is a Ph.D. candidate of Bioinformatics in the College of Agronomy and Biotechnology at China Agricultural University. His project is focused on developing analytical pipelines and software for the analysis of mRNA alternative splicing events in plants based on long reads generated from PacBio Iso-Seq and Oxford Nanopore platforms.

Xiangfeng Wang received his B.S. degree from China Agricultural University in 2002, and his Ph.D. degree in Bioinformatics from Peking University in 2007. He was a postdoc at Harvard University during 2007–2010. He then worked at the University of Arizona as an Assistant Professor during 2010–2014. He has been working in the National Maize Improvement Center at China Agricultural University since 2014. His team develops models for precision breeding in maize, and bioinformatics tools for mining multi-omics data in crops.