MPGK: A user-friendly tool for Mendelian randomization, polygenic risk score, Geno Ontology, and the Kyoto Encyclopedia of Genes and Genomes analysis

Jungang Yang[1,2], Xuejiao Song[3], Jingkai Xu[4], Yue Yang[5], Xianbo Zuo[3], Yong Cui[1,2]

[1]Department of Dermatology, China-Japan Friendship Hospital, Beijing 100029, China;
[2]Department of Dermatology, Peking University China-Japan Friendship School of Clinical Medicine, Beijing 100029, China;
[3]Big Data Center, China-Japan Friendship Hospital, Beijing 100029, China;
[4]NHC Key Laboratory of Clinical Big Data Standardization & Integration, Beijing 100029, China;
[5]Derpartment of Dermatology, China-Japan Friendship Institute of Clinical Medical Sciences, Beijing 100029, China.

Abstract

Background: Mendelian randomization (MR), polygenic risk score (PRS), Geno Ontology (GO), and the Kyoto Encyclopedia of Genes and Genomes (KEGG) are powerful bioinformatic analysis tools. However, the analysis of MR, PRS, GO, and KEGG may pose a challenge for novices. This article intends to introduce a program that adeptly guides beginners in implementing these analysis functions, ensuring that even those new to the field can confidently use them.

Methods: The MPGK program was developed to run on the command line. It conveniently implements the MR, PRS, GO, and KEGG analysis functions by calling well-written R programs. The results of our analyses were validated using genome-wide association study (GWAS) summary data for diabetes and psoriasis, as well as gene sequencing data for diabetes.

Results: Three demo analyses using the MPGK program demonstrate the comprehensive capabilities of the MPGK program in conducting advanced bioinformatics analysis. First, the MPGK program revealed a causal relationship between diabetes and psoriasis. Additionally, the PRS analysis generated polygenic risk scores for diabetes, demonstrating the implementation of PRS analysis within the MPGK framework. Furthermore, the GO and KEGG analyses indicated that psoriasis is associated with infection and T helper 17 cells. These findings are consistent with the previous literature.

Conclusion: MPGK can be easily used to perform comprehensive analysis, including MR, PRS, GO, and KEGG analyses, by both beginners and researchers.

Keywords: MPGK; Mendelian randomization; Polygenic risk score; Gene Ontology; Kyoto Encyclopedia of Genes and Genomes

Introduction

Advancements in next-generation sequencing (NGS) technology have significantly propelled the identification of genetic variations across extensive sample populations. These genetic variations are systematically recorded in variant call format (VCF) files, adhering to established VCF specifications.[1] VCF files not only encapsulate genotype information but also serve as a fundamental resource for subsequent bioinformatics analyses. Such analyses include genome-wide association studies (GWAS), Mendelian randomization (MR) analysis, polygenic risk score (PRS) analysis, and differential gene function enrichment analysis. Consequently, the initial step in these bioinformatics workflows involves accurately parsing the VCF files. The PLINK software, widely recognized in the field, offers robust functionalities to facilitate the processing of VCF files.[2]

Access this article online

Quick Response Code:

Website: www.cmj.org

DOI: 10.1097/CM9.0000000000003980

MR is a robust bioinformatics methodology employed to assess causal relationships between exposure and outcome by leveraging genetic information as an instrumental variable.[3,4] The TwoSampleMR package stands out as the predominant R package designed for conducting MR analysis.[5]

In GWAS, genetic variations—predominantly single nucleotide polymorphisms (SNPs)—have been identified for a wide array of traits. PRS involves summing risk alleles carried by an individual, each weighted by their respective effect sizes, providing predictive estimates at the individual level.[6,7] A variety of software tools are available for calculating PRS, including PRSice,[8,9] Ldpred,[10] PRS-CS,[11] JAMPred,[12] lassosum,[13] and MultiPRS.[14]

Correspondence to: Xianbo Zuo, Big Data Center, China-Japan Friendship Hospital, No. 2, Yinghua East Street, Beijing 100029, China
E-Mail: zuoxianbo@qq.com;
Yong Cui, Department of Dermatology, China-Japan Friendship Hospital, No. 2, Yinghua East Street, Beijing 100029, China
E-Mail: wuhucuiyong@vip.163.com

Chinese Medical Journal 2026;139(10)

Received: 10-03-2025; **Online:** 09-03-2026 **Edited by:** Sihan Zhou and Xiuyuan Hao

Functional enrichment analysis is a prevalent bioinformatics technique used to decipher differential gene function.[15] Annotated bioinformatics databases such as Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) offer extensive data on the biological functions of genes and proteins. The clusterProfiler R package is particularly notable for supporting both GO and KEGG enrichment analyses, coupled with excellent visualization capabilities.[16,17]

As genomics and genetics research continues to advance, the generation of large volumes of multi-omics data has become a common practice, necessitating the extraction of valuable insights. Bioinformatics analyses—including MR, PRS, and functional enrichment analysis—remain at the forefront of this endeavor.

The aforementioned analysis system is both extensive and complex, often presenting conceptual challenges to novices in the field. The diverse array of R packages and software tools, each implementing specific functions, is dispersed across various online platforms, making the learning curve steep and time-consuming. To address this issue and to streamline the learning process, we developed the MPGK program.

Psoriasis is a chronic and often refractory inflammatory skin disease, with clinical studies recognizing its comorbidities such as hypertension, coronary artery disease, diabetes mellitus, and hyperlipidemia.[18,19] In addition, numerous bioinformatics studies have been conducted to explore the genetic underpinnings of psoriasis. This paper elucidates the operation of the MPGK program and exemplifies its utility through a demonstration using psoriasis-related demo data.

Methods

Software development

The MPGK program was designed as an integrated, command line-based bioinformatics pipeline to standardize and automate post-GWAS analyses, thereby reducing manual intervention and improving reproducibility across platforms. MPGK integrates MR, PRS, GO, and KEGG analysis modules into a unified workflow that can be executed across different operating systems via a single shell script [Figure 1].

The MR analysis module reads GWAS summary statistics for exposure and outcome traits and performs two-sample MR analyses based on the TwoSampleMR package (version 0.0.6; MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK). Multiple MR diagnostic and sensitivity analyses are supported, and visualization is implemented using ggplot2 (version 3.5.1; R Foundation for Statistical Computing, Vienna, Austria), generating scatter plots, forest plots, funnel plots, and leave-one-out plots.

The PRS analysis module processes genotype data using PLINK (version 1.90b7; Broad Institute of MIT and Harvard, Cambridge, MA, USA) for quality control and data transformation, followed by PRS calculation using PRSice (version 2.3.3; MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK). The module outputs standardized PRS results and automatically generates probability distribution plots to facilitate phenotype-genotype interpretation.

For functional annotation, the GO and KEGG analysis module selects genome-wide significant variants ($P < 5 \times 10^{-8}$) and maps them to corresponding genes for downstream enrichment analysis. Functional enrichment and visualization are performed using the clusterProfiler package (version 4.8.3; Bioconductor Project, Fred Hutchinson Cancer Research Center, Seattle, WA, USA), enabling the generation of dot plots, bar plots, directed acyclic graphs (DAGs), and network plots.

Each analytical module can be executed independently or combined within a single automated workflow, allowing flexible configuration according to specific research objectives.

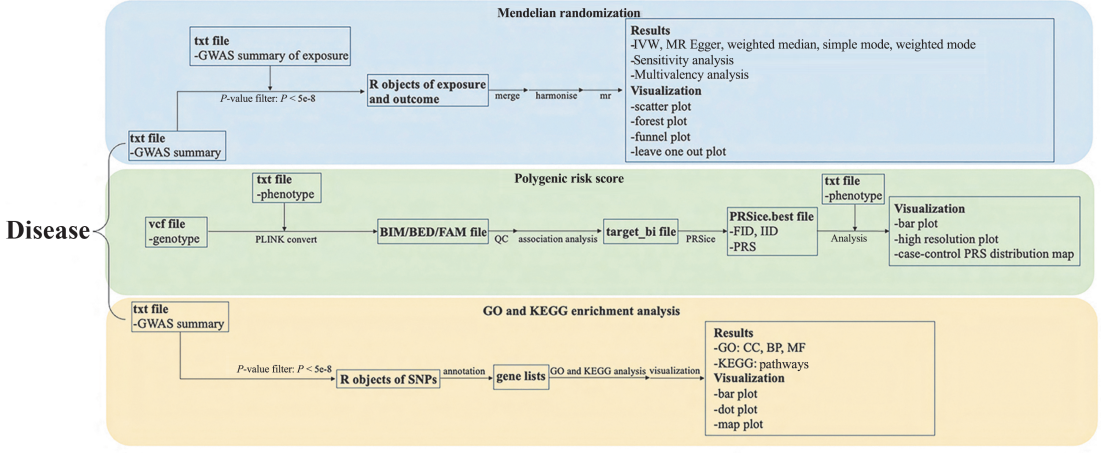

Figure 1: An overview of MPGK program workflows. The MPGK program comprises three main modules: Mendelian randomization (MR), polygenic risk score (PRS), and Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses. BP: Biological process; CC: Cellular component; FID: Family identifier; GWAS: Genome-wide association study; IID: Individual identifier; IVW: Inverse variance weighted; MF: Molecular function; QC: Quality control; SNP: Single nucleotide polymorphism.

Software usage

MPGK employs a concise, flag-based command line interface to improve usability and facilitate reproducible analysis. The program includes the following parameters: M for MR, P for PRS, G for GO, and K for KEGG.

Required input files include an exposure GWAS summary file for MR analysis (specified with e), an outcome GWAS summary file for MR analysis (specified with o), a genotype VCF file for PRS analysis (specified with v), a phenotype file containing sample identifiers (specified with f), and GWAS summary data for GO and KEGG analysis (specified with d). In addition, a *P* value threshold must be specified with the p parameter. Detailed parameter settings and usage examples are provided in Supplementary Material 1, http://links.lww.com/CM9/C771.

Software availability

The MPGK program, together with test datasets and example scripts, is publicly available at the GitHub repository: https://github.com/BioinfoEasy/MPGK.

Test data

All datasets used in this study were obtained from publicly available GWAS repositories or institutional sequencing resources. The psoriasis GWAS datasets (demo1.txt and demo4.txt) were downloaded from the IEU OpenGWAS database (ukb-b-10537) and converted from .vcf to .txt format, retaining SNPs with $P < 5 \times 10^{-8}$. The diabetes GWAS dataset (demo2.txt) was obtained from the IEU OpenGWAS database (ukb-b-12948) and processed using the same criteria. The genotype VCF file (demo3.vcf) and corresponding phenotype file (demo3_phe.txt) for diabetes were derived from unpublished institutional sequencing data and used solely for demonstrating the fuctionality of the MPGK program.

Study overview

To demonstrate the practical application of MPGK, the program was applied to a representative post-GWAS analysis workflow integrating MR, PRS, GO, and KEGG analyses. MR, GO, and KEGG modules directly processed GWAS summary statistics in text format, whereas the PRS module utilized genotype data in VCF format together with phenotype information provided in text files.

For MR analysis, SNPs significantly associated with the exposure (diabetes) at $P < 5 \times 10^{-8}$ (customizable) were selected as instrumental variables, and corresponding SNP effects were extracted from the outcome (psoriasis) GWAS dataset. MPGK automatically performed heterogeneity and horizontal pleiotropy assessments and generated corresponding visualizations, including scatter plots, forest plots, funnel plots, and leave-one-out plots.

For PRS analysis, genotype data were merged with phenotype files using PLINK and converted into binary formats. Following stringent quality control, association analyses were conducted to generate the target_bi binary file. MPGK subsequently invoked PRSice to calculate polygenic risk scores and generate the PRSice.best file, which contains sample identifiers, PRS values, and phenotype information. The program then automatically plotted PRS probability distributions by jointly analyzing the PRSice.best file and phenotype data.

For functional enrichment analyses, genome-wide significant SNPs ($P < 5 \times 10^{-8}$) were annotated to their corresponding genes, and GO and KEGG enrichment analyses were performed on the resulting gene sets. MPGK generated multiple types of enrichment visualizations to facilitate biological interpretation of the results.

Results

MR analysis using MPGK.

Using MPGK, the MR analysis revealed a significant causal association between diabetes and psoriasis. Table 1 presents the findings from the five MR analyses. The MR conducted using the MPGK program indicated a causal relationship between diabetes and psoriasis. The *P* value for the inverse variance weighted (IVW) method was 3.75×10^{-25} ($P < 0.05$). The funnel plot [Figure 2A] enables the observation of SNP heterogeneity, with the points on either side of the IVW line appearing roughly symmetrical, indicating low heterogeneity. Various MR tests consistently suggested a positive causal relationship between diabetes and psoriasis [Figure 2B].

PRS analysis of MPGK.

PRS analyses were conducted using gene sequencing results of the diabetes and phenotypic data. The optimal *P* value threshold was found to be 0.03 [Figure 3A and 3B]. However, the SNPs used in the PRS analysis did not fully distinguish between case and control phenotypes [Figure 3C].

GO and KEGG pathway analysis using MPGK.

SNPs in the psoriasis GWAS summary file were screened using a threshold of *P* value $< 5 \times 10^{-8}$. Gene annotation, as well as GO and KEGG functional enrichment analyses, were subsequently performed. In the GO analysis, the cellular component (CC) was visualiszed. Dot plots

Table 1: MR results for the causal effect of diabetes on psoriasis using multiple MR methods.

Methods	nSNP	*P* value	OR
MR Egger	1282	0.16	1.08
Weighted median	1282	0	1.39
IVW	1282	3.75×10^{-25}	1.21
Simple mode	1282	0	1.59
Weighted mode	1282	0	1.51

IVW: Inverse variance weighted; MR: Mendelian randomization; nSNP: Number of SNPs used as instrumental variables; OR: Odds ratio; SNP: Single nucleotide polymorphism.

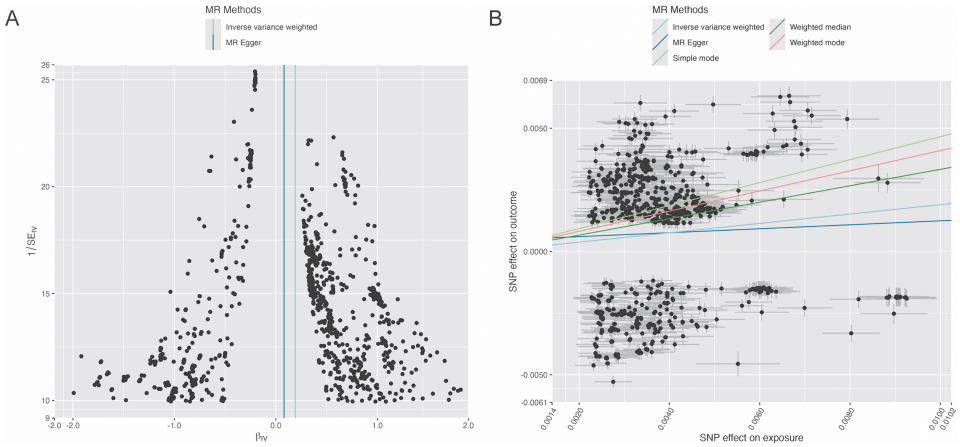

Figure 2: Visualization of MR analysis. (A) The funnel plot; (B) The scatter plot. The effect of SNPs on exposure and outcome is shown. MR: Mendelian randomization; SNPs: Single nucleotide polymorphisms.

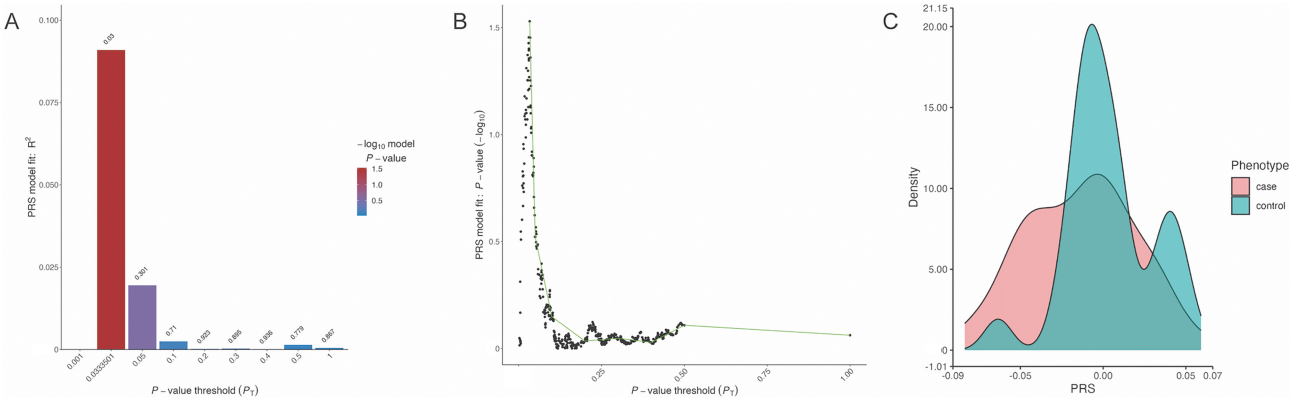

Figure 3: Visualization of polygenic risk score analysis. (A) The bar plot showing PRS model fit (R^2) across different GWAS SNP inclusion thresholds (*P* value threshold, *Pt*); (B) The high-resolution plot illustrating the statistical significance (-$\log_{10}$ model *P* value) of PRS models across different Pt values; (C) Probability distribution of PRS scores in case and control groups. PRS: Polygenic risk score; GWAS: genome-wide association study; SNP: single nucleotide polymorphism.

and bar plots were generated and enhanced using the ggplot2 package. The analysis revealed that genes related to psoriasis were enriched in various locations, including the external side of the plasma membrane, the MHC protein complex, and transport vesicles [Figure 4A and 4B]. Specific cellular components (CCs) in psoriasis patients were notably related to structures such as cell membranes, vesicles, and the Golgi apparatus [Figure 4C and 4D]. Additionally, the KEGG analysis showed that these genes were enriched in Epstein–Barr virus infection, antigen processing and presentation, and type 1 diabetes mellitus pathways [Figure 5A and 5B].

Disscussion

This paper presents the function and usage of the MPGK program with four demo datasets. We obtained the results of MR analysis, heterogeneity analysis, and pleiotropy analysis on the datasets of psoriasis (demo1.txt) and diabetes (demo2.txt) using the MPGK program. Additionally, we performed PRS analysis on our in-house datasets (demo3.vcf and demo3_phe.txt) using the MPGK program, and showed PRS probability distribution for case and control groups. We implemented the MPGK program to perform SNP screening, gene annotation, and GO and KEGG pathway analyses on GWAS dataset of psoriasis (demo4.txt).

These examples demonstrated the comprehensive capabilities of the MPGK program in conducting advanced bioinformatics analysis, and the results analyzed by MPGK were consistent with previous research findings. In the first demo analysis, the MPGK program has identified a causal relationship between diabetes and psoriasis, which is consistent with the findings of Zhao *et al*.[20] They investigated the association between psoriasis and lifestyle and comorbidities using cross-sectional analysis and MR, and found that poor blood sugar control can increase the risk of psoriasis. In the second demo analysis, MPGK calculated PRS scores for each individual diabetic patient and plotted PRS probability distributions for case and control. In the third demo analysis, the MPGK program conducted functional enrichment analyses on significant genes associated with psoriasis, revealing enrichment in pathways related to diabetes, asthma, viral and bacterial infections, T helper (Th)1 and Th2 cell differentiation, and Th17 cell differentiation. These insights are consistent with the existing literature, including a review by Junko *et al*[21], which discusses psoriasis-related comorbidities such as metabolic syndrome, cardiovas-

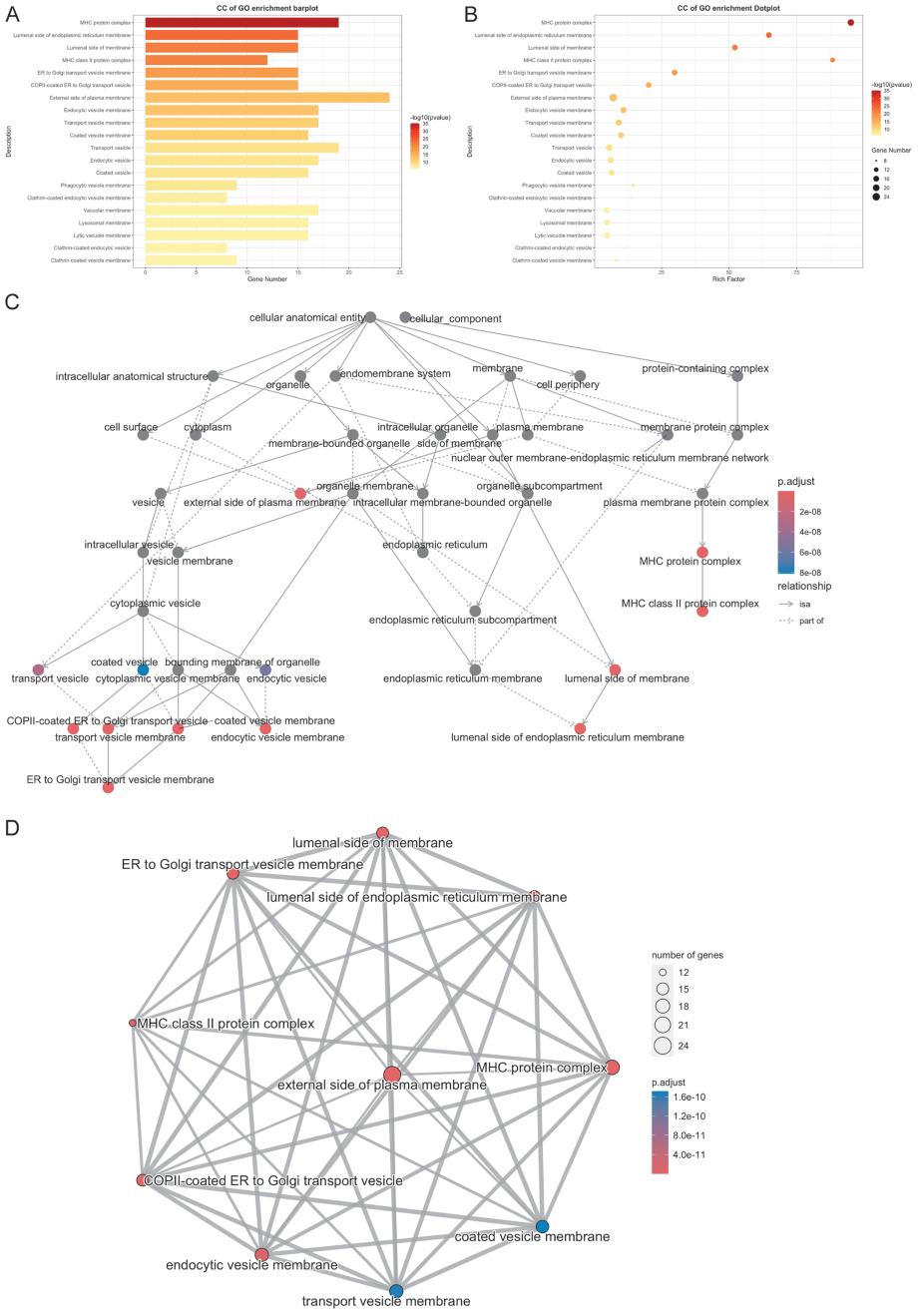

Figure 4: Visualization of GO analysis (CC as an example). (A) The bar plot; (B) The dot plot; (C) The DAG; (D) The net plot. The bar plot and dot plot display the top 20 CC terms. DAG: Directed acyclic graph; CC: Cellular component; GO: Gene ontology; COPII-coated: Coat protein II; ER: Endoplasmic reticulum; MHC: Major histocompatibility complex; p.adjust: Adjusted *P* values.

cular disease, depression, and infections. Zhou and Yao *et al*[22] also summarized the impact of bacterial, viral, and fungal infectious factors on psoriasis, noting that the pathogenesis involves interferon-gamma (IFN-γ), IL-12/IL-23, Th1, Th17, and Th22 pathways.[22,23]

The MPGK program boasts several key features. Firstly, it offers a user-friendly interface that facilitates quick access to a comprehensive range of bioinformatics analysis results. Secondly, it significantly reduces the learning curve for researchers [Supplementary Material 2, http://links.lww.com/CM9/C864], enabling them to allocate more time to data analysis. Additionally, the program leverages the ggplot2 package to generate high-quality, visually appealing results. Finally, the MPGK program is implemented in the R language and incorporates a few command line codes, making it cross-platform compatible and capable of being running on both Linux and MacOS systems without the need for high computer specifications. In conclusion, the MPGK program is a valuable tool for facilitating MR, PRS, GO, and KEGG analyses for beginners and researchers alike.

This MPGK program currently has certain limitations. With the growth of multi-omics research, the study of diseases has expanded beyond genomic variants. However,

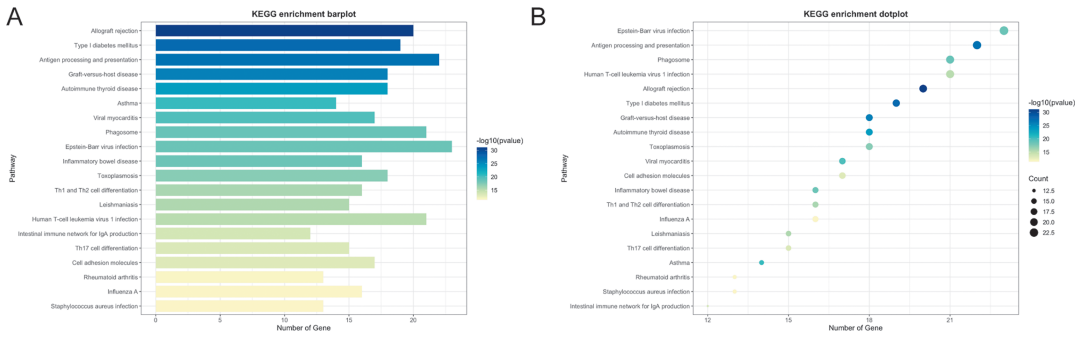

Figure 5: Visualization of KEGG analysis. (A) The bar plot; (B) The dot plot. The top 20 KEGG pathways of psoriasis are presented. IgA: Immunoglobulin A; KEGG: Kyoto Encyclopedia of Genes and Genomes; Th: T helper cell.

the MPGK program currently employs only basic research methods in genomics. We plan to incorporate additional omic analysis methods in the future to address this limitation. Additionally, with the advancement of artificial intelligence, we intend to expand the Python language in our framework to facilitate a the more comprehensive analysis capabilities of MPGK. This expansion also enables researchers and clinicians to easily utilize the machine learning libraries in Python.

Funding

This study was supported by grants from the National Key Research & Development Program of China (No. 2022YFC3602002), and National Natural Science Foundation Project (Nos. 82330097 and 82273523).

Conflicts of interest

None.

References

1. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, *et al*. The variant call format and VCFtools. Bioinformatics 2011;27:2156–2158. doi: 10.1093/bioinformatics/btr330.
2. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, *et al*. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–575. doi: 10.1086/519795.
3. Smith GD, Ebrahim S. Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol 2003;32:1–22. doi: 10.1093/ije/dyg070.
4. VanderWeele TJ, Tchetgen EJ, Cornelis M, Kraft P. Methodological challenges in Mendelian randomization. Epidemiology 2014;25:427–435. doi: 10.1097/EDE.0000000000000081.
5. Zheng J, Baird D, Borges MC, Bowden J, Hemani G, Haycock P, *et al*. Recent developments in mendelian randomization studies. Curr Epidemiol Rep 2017;4:330–345. doi: 10.1007/s40471-017-0128-6.
6. Lewis CM, Vassos E. Polygenic risk scores: From research tools to clinical instruments. Genome Med 2020;12:44. doi: 10.1186/s13073-020-00742-5.
7. Choi SW, Mak TS, O'Reilly PF. Tutorial: A guide to performing polygenic risk score analyses. Nat Protoc 2020;15:2759–2772. doi: 10.1038/s41596-020-0353-1.
8. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic risk score software. Bioinformatics 2015;31:1466–1468. doi: 10.1093/bioinformatics/btu848.
9. Choi SW, O'Reilly PF. PRSice-2: Polygenic risk score software for biobank-scale data. Gigascience 2019;8:giz082. doi: 10.1093/gigascience/giz082.
10. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, *et al*. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am J Hum Genet 2015;97:576–592. doi: 10.1016/j.ajhg.2015.09.001.
11. Ge T, Chen CY, Ni Y, Feng YA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun 2019;10:1776. doi: 10.1038/s41467-019-09718-5.
12. Newcombe PJ, Nelson CP, Samani NJ, Dudbridge F. A flexible and parallelizable approach to genome-wide polygenic risk scores. Genet Epidemiol 2019;43:730–741. doi: 10.1002/gepi.22245.
13. Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. Genet Epidemiol 2017;41:469–480. doi: 10.1002/gepi.22050.
14. Márquez-Luna C, Loh PR, South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium, Price AL. Multiethnic polygenic risk scores improve risk prediction in diverse populations. Genet Epidemiol 2017;41:811–823. doi: 10.1002/gepi.22083.
15. Chicco D, Jurman G. A brief survey of tools for genomic regions enrichment analysis. Front Bioinform 2022;2:968327. doi: 10.3389/fbinf.2022.968327.
16. Yu G, Wang LG, Han Y, He QY. clusterProfiler: An R package for comparing biological themes among gene clusters. OMICS 2012;16:284–287. doi: 10.1089/omi.2011.0118.
17. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, *et al*. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb) 2021;2:100141. doi: 10.1016/j.xinn.2021.100141.
18. Amin M, Lee EB, Tsai TF, Wu JJ. Psoriasis and co-morbidity. Acta Derm Venereol 2020;100:adv00033. doi: 10.2340/00015555-3387.
19. Yamazaki F. Psoriasis: Comorbidities. J Dermatol 2021;48:732–740. doi: 10.1111/1346-8138.15840.
20. Zhao SS, Bellou E, Verstappen SMM, Cook MJ, Sergeant JC, Warren RB, *et al*. Association between psoriatic disease and lifestyle factors and comorbidities: Cross-sectional analysis and Mendelian randomization. Rheumatology (Oxford) 2023;62:1272–1285. doi: 10.1093/rheumatology/keac403.
21. Takeshita J, Grewal S, Langan SM, Mehta NN, Ogdie A, Van Voorhees AS, *et al*. Psoriasis and comorbid diseases: Epidemiology. J Am Acad Dermatol 2017;76:377–390. doi: 10.1016/j.jaad.2016.07.064.
22. Zhou S, Yao Z. Roles of infection in psoriasis. Int J Mol Sci 2022;23:6955. doi: 10.3390/ijms23136955.
23. Brembilla NC, Senra L, Boehncke WH. The IL-17 family of cytokines in psoriasis: IL-17A and beyond. Front Immunol 2018;9:1682. doi: 10.3389/fimmu.2018.01682.

How to cite this article: Yang JG, Song XJ, Xu JK, Yang Y, Zuo XB, Cui Y. MPGK: A user-friendly tool for Mendelian randomization, polygenic risk score, Geno Ontology, and the Kyoto Encyclopedia of Genes and Genomes analysis. Chin Med J 2026;139:1532–1537. doi: 10.1097/CM9.0000000000003980