Analyzing bivariate cross-trait genetic architecture in GWAS summary statistics with the BIGA cloud computing platform

Running title: BIGA GWAS cloud platform

Yujue Li[1], Fei Xue[1], Bingxuan Li[2], Yilin Yang[3], Zirui Fan[4], Juan Shu[1], Xiyao Wang[2], Jinjie Lin[5], Carlos Copana[1], and Bingxin Zhao[4]*

[1]Department of Statistics, Purdue University, West Lafayette, IN 47907, USA.

[2]Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA.

[3]Department of Computer and Information Science and Electrical and Systems Engineering, School of Engineering & Applied Science, University of Pennsylvania, Philadelphia, PA 19104, USA.

[4]Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104, USA.

[5]Yale School of Management, Yale University, New Haven, CT 06511, USA.

**Corresponding to*:

Bingxin Zhao

413 Academic Research Building

265 South 37th Street, Philadelphia, PA 19104.

E-mail: bxzhao@upenn.edu Phone: (215) 898-8222

Abstract

As large-scale biobanks provide increasing access to deep phenotyping and genomic data, genome-wide association studies (GWAS) are rapidly uncovering the genetic architecture behind various complex traits and diseases. GWAS publications typically make their summary-level data (GWAS summary statistics) publicly available, enabling further exploration of genetic overlaps between phenotypes gathered from different studies and cohorts. However, systematically analyzing high-dimensional GWAS summary statistics for thousands of phenotypes can be both logistically challenging and computationally demanding. In this paper, we introduce BIGA (http://bigagwas.org/), a website that offers unified data analysis pipelines and centralized data resources for cross-trait genetic architecture analyses using GWAS summary statistics. We have developed a framework to implement statistical genetics tools on a cloud computing platform, combined with extensive curated GWAS data resources. Through BIGA, users can upload data, submit jobs, and share results, providing the research community with a convenient tool for consolidating GWAS data and generating new insights.

Keywords: GWAS; Cross-trait analysis; Cloud computing; Online platform.

The rapid development of biobank-scale biomedical databases, encompassing phenotyping and genomic data, has occurred globally[1]. Numerous genome-wide association studies (GWAS) have been conducted to determine the genetic architecture underlying a wide range of complex traits and clinical outcomes, with the aim of improving disease prevention and treatment[2]. Publicly available GWAS summary-level data (or GWAS summary statistics) encompass thousands of phenotypes[3-8]. These summary statistics, derived from large-scale studies, provide valuable opportunities for in-depth investigations into genetic overlaps and shared architectures between phenotypes across studies and cohorts. Various statistical genetic tools have been developed to analyze GWAS summary statistics and examine the shared genetic components between pairs of phenotypes, including LDSC[9], LAVA[10], Coloc[11], and Mendelian randomization[12]. These methods offer insights into genetic links from various perspectives and have been widely applied to clinical biomarkers and outcomes[13,14].

However, implementing and batch-running these tools often requires robust computing and data infrastructure, which may not always be available to all researchers. Consequently, systematic bivariate cross-trait analyses using massive GWAS summary statistics for thousands of phenotypes can be logistically and computationally challenging. As more complex and deep phenotyping data are obtained from biobanks[15], addressing these limitations becomes increasingly urgent. For example, the UK Biobank (UKB) imaging study[16] collected multimodal brain imaging data, generating over 5,000 imaging-derived phenotypes using different imaging modalities and processing pipelines[17-21]. Researchers interested in a specific disease and its genetic connections with imaging biomarkers have traditionally downloaded all the GWAS summary statistics for over 5,000 imaging biomarkers from the Oxford BIG40 Project (http://big.stats.ox.ac.uk) and the BIG-KP project (https://bigkp.org/), and run their statistical tools in local clusters, which can be inefficient. Several online research platforms based on cloud computing have been developed, most of which focus on one database (such as the UKB study, https://ukbiobank.dnanexus.com/), univariate trait GWAS analysis (such as FUMA[22]), or single data analysis method/function (such as LD Hub[23] and Locus Compare[24]). Developing an integrated platform for cross-trait analyses of GWAS summary data resources will make existing large-scale GWAS summary data more accessible to researchers.

To address these limitations, we developed BIGA (http://bigagwas.org/), an online cloud-based platform that offers unified data analysis pipelines and centralized data resources for cross-trait analyses using GWAS summary statistics. BIGA aims to provide various tools for quantifying cross-trait genetic architectures, such as genome-wide genetic correlation methods (e.g., LDSC[9], Popcorn[25], and SumHer[26]), polygenic overlap estimation between two traits (e.g., MiXeR[27]), genomic structural equation modeling (e.g., GenomicSEM[28]), local genetic correlation analysis (e.g., LAVA[10]), Mendelian randomization (e.g., MR-Base[29], MR-RAPS[30], DIWV[31], and GRAPPLE[32]), and colocalization (e.g., Coloc[11]). We have also aggregated and preprocessed GWAS summary statistics from various resources (e.g., the UKB[15], GWAS Atlas[4], FinnGen[6], Biobank Japan[8], BIG-KP[18,19,21], and Oxford BIG40[17,20]) and provided curated datasets. Our framework can easily be extended to incorporate additional methods and GWAS summary statistics.

Figure 1 provides an overview of the BIGA architecture. We offer users several options for inputting GWAS summary statistics data, including uploading their own data, querying harmonized data from public databases (such as the IEU OpenGWAS[7] and Pan-UK Biobank, https://pan.ukbb.broadinstitute.org/), and querying data from the BIGA in-house database. Users can specify the tools and job types they are interested in and submit their requests. After submission, the job request will be passed to the back-end and executed on our cloud computing platform using the specified tools and datasets. Once completed, the results will be presented to the user through the front-end interface. Users have the option to share their results and/or data with our public database.

Here we present an LDSC data analysis example for schizophrenia[33]. Specifically, we investigated the genetic correlation between schizophrenia and complex traits and diseases available in our current curated datasets. These datasets include 4,575 UKB phenotypes provided by the Neale Lab (http://www.nealelab.is/uk-biobank), 3,095 clinical outcomes from FinnGen[6], 199 brain-related phenotypes in GWAS Atlas[4], 3,905 brain imaging phenotypes from the Oxford BIG40[17,20], and 3,016 multi-organ imaging phenotypes from the BIG-KP[18,19,21]. **Figure S1** displays the job information for these LDSC data analyses, including Job ID, Job Name, Job Type, Status, Download Link, Result link,

and a link to share your results with the public. Once the job is complete, users can either download the full set of results or view them in our online table. For example, among the 199 brain-related phenotypes in GWAS Atlas[4], we discovered that the top-ranking phenotypes exhibiting strong genetic correlations with schizophrenia included neuroticism and its subclusters, daytime napping, and daytime sleepiness (**Fig. S2**). These results can be effortlessly shared on our Public Results Portal (**Fig. S3**). This data analysis demonstrates that our platform facilitates rapid analysis of extensive GWAS summary statistics and enables seamless result sharing with the community.

In summary, our platform enables researchers to easily perform multiple cross-trait analyses without needing access to a local research computing cluster, implementing methods locally, or downloading large datasets. BIGA will help reduce the imbalance in the research community caused by unequal computing resources. The source code to build the BIGA platform will be made publicly available on GitHub, which can serve as an example for creating more cloud computing-based genomic data analysis websites. The BIGA website welcomes user feedback and requests, which aids in improving the project and implementing new tools and functions to better meet the needs of the research community.

ADDITIONAL INFORMATION

One supplementary information pdf file is available, where the figures of an example data analysis using BIGA are displayed.

ACKNOWLEDGEMENTS

The study has been partially supported by funding from the Wharton Dean's Research Fund and Analytics at Wharton, as well as start-up funds from Purdue Statistics Department. We would like to thank the research computing team at the Wharton School of the University of Pennsylvania and Purdue University for providing computational resources and support that have contributed to this project. We would like to thank all the developers of the tools and methods implemented in our project. We gratefully acknowledge all the studies and databases that made GWAS summary data available and thank the individuals who represented these studies for their participation and the

research teams for their work in collecting, processing, and disseminating these datasets for analysis.

AUTHOR CONTRIBUTIONS

Y.L. and B.Z. designed the study, developed the BIGA website, and wrote the manuscript with feedback from all authors. B.L. helped with the implementation of statistical genetic methods and website functions. Y.Y., Z.F., J.S., X.W., B.L., and C.C. processed the GWAS summary statistics, developed the curated datasets, and contributed to the development of the website. F.X. and J.L. provided feedback on the study design and website.

CORRESPONDENCE AND REQUESTS FOR MATERIALS should be addressed to B.Z.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

REFERENCES

1. Zhou, W. *et al.* Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genomics* **2**, 100192 (2022).
2. Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 1-21 (2021).
3. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research* (2022).
4. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nature genetics* **51**, 1339-1348 (2019).
5. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nature genetics* **50**, 1593-1599 (2018).
6. Kurki, M.I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508-518 (2023).
7. Elsworth, B. *et al.* The MRC IEU OpenGWAS data infrastructure. *BioRxiv* (2020).
8. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nature genetics* **53**, 1415-1424 (2021).

9. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47**, 1236-1241 (2015).
10. Werme, J., van der Sluis, S., Posthuma, D. & de Leeuw, C.A. An integrated framework for local genetic correlation analysis. *Nature genetics* **54**, 274-282 (2022).
11. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics* **10**, e1004383 (2014).
12. Burgess, S., Scott, R.A., Timpson, N.J., Davey Smith, G. & Thompson, S.G. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *European journal of epidemiology* **30**, 543-552 (2015).
13. Romero, C. *et al.* Exploring the genetic overlap between twelve psychiatric disorders. *Nature Genetics*, 1-8 (2022).
14. Lee, P.H. *et al.* Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell* **179**, 1469-1482. e11 (2019).
15. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
16. Littlejohns, T.J. *et al.* The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nature communications* **11**, 1-12 (2020).
17. Elliott, L.T. *et al.* Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562**, 210-216 (2018).
18. Zhao, B. *et al.* Common genetic variation influencing human white matter microstructure. *Science* **372**, eabf3736 (2021).
19. Zhao, B. *et al.* Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nature genetics* **51**, 1637-1644 (2019).
20. Smith, S.M. *et al.* An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature neuroscience* **24**, 737-745 (2021).

21. Zhao, B. *et al.* Common variants contribute to intrinsic human brain functional networks. *Nature Genetics* **54**, 508-517 (2022).
22. Watanabe, K., Taskesen, E., Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications* **8**, 1826 (2017).
23. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272-279 (2017).
24. Liu, B., Gloudemans, M.J., Rao, A.S., Ingelsson, E. & Montgomery, S.B. Abundant associations with gene expression complicate GWAS follow-up. *Nature genetics* **51**, 768-769 (2019).
25. Brown, B.C., Ye, C.J., Price, A.L., Zaitlen, N. & Consortium, A.G.E.N.T.D. Transethnic genetic-correlation estimates from summary statistics. *The American Journal of Human Genetics* **99**, 76-88 (2016).
26. Speed, D. & Balding, D.J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nature genetics* **51**, 277-284 (2019).
27. Frei, O. *et al.* Bivariate causal mixture model quantifies polygenic overlap between complex traits beyond genetic correlation. *Nature communications* **10**, 1-11 (2019).
28. Grotzinger, A.D. *et al.* Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature human behaviour* **3**, 513-525 (2019).
29. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *elife* **7**(2018).
30. Zhao, Q., Wang, J., Hemani, G., Bowden, J. & Small, D.S. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *The Annals of Statistics* **48**, 1742-1769 (2020).
31. Ye, T., Shao, J. & Kang, H. Debiased inverse-variance weighted estimator in two-sample summary-data Mendelian randomization. *The Annals of statistics* **49**, 2079-2100 (2021).

32. Wang, J. *et al.* Causal inference for heritable phenotypic risk factors using heterogeneous genetic instruments. *PLoS genetics* **17**, e1009575 (2021).
33. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502-508 (2022).

Code availability

All source code to develop the BIGA platform will be made publicly available on the BIGA GitHub repository. The statistical tools and methods implemented in the BIGA platform are also open source, and their source code has already been made available to the public by their authors.

Data availability

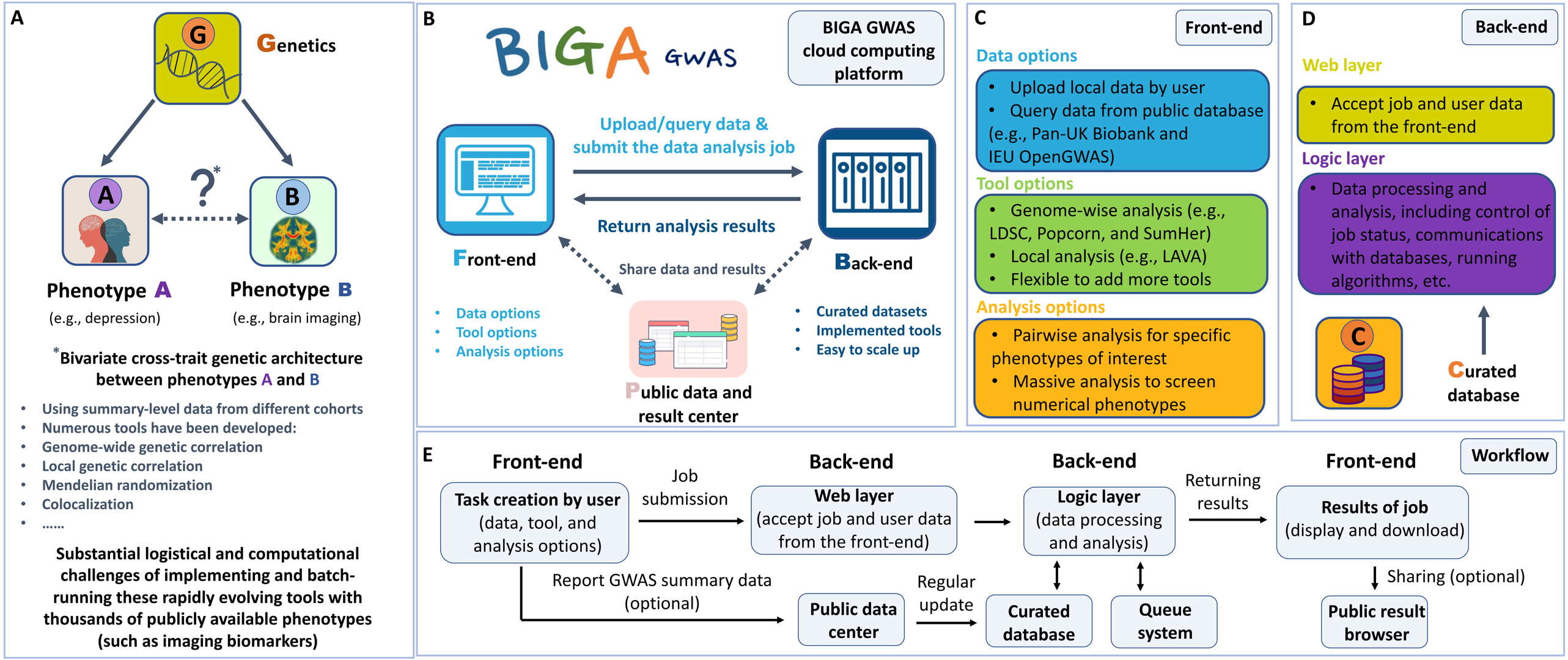
GWAS summary statistics used in the BIGA platform are publicly available and can be found in several public databases, such as the
Pan-UK Biobank (https://pan.ukbb.broadinstitute.org/),
IEU OpenGWAS (https://gwas.mrcieu.ac.uk/),
GWAS Atlas (https://atlas.ctglab.nl/),
FinnGen (https://www.finngen.fi/en/access_results),
Biobank Japan (https://pheweb.jp/),
BIG-KP (https://bigkp.org/), and
Oxford BIG40 (https://open.win.ox.ac.uk/ukbiobank/big40/).

Figure Legend

Fig. 1 Overview of BIGA GWAS cloud computing platform.

(A) The motivation of this project is to address the substantial logistical and computational challenges associated with implementing and batch-running the constantly evolving tools for cross-trait genetic architecture analysis. Our aim is to offer a cloud computing-based solution that can effectively overcome these challenges. **(B)** Overview of the BIGA GWAS platform. Users can easily upload their GWAS summary level data and submit data analysis jobs through the front-end interface. These jobs are then processed on the back-end, and the results are subsequently returned to the users. In

addition, we have established public data and result centers where users can share their data and outcomes with the public. **(C)** The front-end interface of the BIGA GWAS platform offers users a comprehensive set of options to manage their data resources, choose the appropriate tools, and select the desired mode of data analysis. **(D)** Details of the back-end of the BIGA GWAS platform. **(E)** Overview of the analysis workflow.


A
Genetics
?*
Phenotype A
(e.g., depression)
Phenotype B
(e.g., brain imaging)
*Bivariate cross-trait genetic architecture between phenotypes A and B
Using summary-level data from different cohorts
Numerous tools have been developed:
Genome-wide genetic correlation
Local genetic correlation
Mendelian randomization
Colocalization
.....
Substantial logistical and computational challenges of implementing and batch-running these rapidly evolving tools with thousands of publicly available phenotypes (such as imaging biomarkers)
B
BIGA GWAS
BIGA GWAS cloud computing platform
Upload/query data & submit the data analysis job
Return analysis results
Share data and results
Front-end
Back-end
Data options
Tool options
Analysis options
Public data and result center
Curated datasets
Implemented tools
Easy to scale up
C
Front-end
Data options
Upload local data by user
Query data from public database (e.g., Pan-UK Biobank and IEU OpenGWAS)
Tool options
Genome-wise analysis (e.g., LDSC, Popcorn, and SumHer)
Local analysis (e.g., LAVA)
Flexible to add more tools
Analysis options
Pairwise analysis for specific phenotypes of interest
Massive analysis to screen numerical phenotypes
D
Back-end
Web layer
Accept job and user data from the front-end
Logic layer
Data processing and analysis, including control of job status, communications with databases, running algorithms, etc.
Curated database
E
Workflow
Front-end
Task creation by user (data, tool, and analysis options)
Job submission
Back-end
Web layer (accept job and user data from the front-end)
Back-end
Logic layer (data processing and analysis)
Returning results
Front-end
Results of job (display and download)
Report GWAS summary data (optional)
Public data center
Regular update
Curated database
Queue system
Sharing (optional)
Public result browser