Descriptor

Patterns

PYPE: A pipeline for phenome-wide association and Mendelian randomization in investigator-driven biobank scale analysis

Highlights

- PYPE is a Python pipeline for running phenome-wide association studies (PheWASs)
- PYPE automates the analysis and visualization of genotype-phenotype relationships
- PYPE uncovers genetically derived causal relationships between phenotypes

Authors

Taykhoom Dalal, Chirag J. Patel

Correspondence

tid4007@med.cornell.edu (T.D.), chirag_patel@hms.harvard.edu (C.J.P.)

In brief

Dalal and Patel have created a pipeline for running high-throughput phenome-wide association studies on large-scale biobank data sources. The pipeline, written in Python, provides a user-friendly end-to-end tool for analyzing genetic influences across a broad range of phenotypes. Dubbed PYPE, this tool simplifies and accelerates the process of discovering complex genotype-phenotype relationships and potential causal links between traits, making it invaluable for researchers seeking to explore the genetic underpinnings of health and disease.

Dalal & Patel, 2024, Patterns *5*, 100982
June 14, 2024
https://doi.org/10.1016/j.patter.2024.100982

CellPress

Patterns

CellPress
OPEN ACCESS

Descriptor

PYPE: A pipeline for phenome-wide association and Mendelian randomization in investigator-driven biobank scale analysis

Taykhoom Dalal[1,*] and Chirag J. Patel[1,2,*]
[1]Harvard Medical School Department of Biomedical Informatics, Boston, MA 02115, USA
[2]Lead contact
*Correspondence: tid4007@med.cornell.edu (T.D.), chirag_patel@hms.harvard.edu (C.J.P.)
https://doi.org/10.1016/j.patter.2024.100982

THE BIGGER PICTURE As large biobanks increasingly expand their measurements of health data from genetic to behavioral information, there is a growing need for tools to help analysts annotate and enhance the discovery of the links between these various data types. While there are databases that automatically deliver findings to investigators, they are "one size fits all" and do so without input or hypotheses from the user. To address these limitations, we developed PYPE, an open-source, one-stop shop optimized for identifying and interpreting genotypic and phenotypic relationships from large-scale biomedical biobanks. PYPE also only requires basic familiarity with a command line interface for use, ensuring that it is accessible to a wide range of researchers and hypotheses.

SUMMARY

Phenome-wide association studies (PheWASs) serve as a way of documenting the relationship between genotypes and multiple phenotypes, helping to uncover unexplored genotype-phenotype associations (known as pleiotropy). Secondly, Mendelian randomization (MR) can be harnessed to make causal statements about a pair of phenotypes by comparing their genetic architecture. Thus, approaches that automate both PheWASs and MR can enhance biobank-scale analyses, circumventing the need for multiple tools by providing a comprehensive, end-to-end tool to drive scientific discovery. To this end, we present PYPE, a Python pipeline for running, visualizing, and interpreting PheWASs. PYPE utilizes input genotype or phenotype files to automatically estimate associations between the chosen independent variables and phenotypes. PYPE can also produce a variety of visualizations and can be used to identify nearby genes and functional consequences of significant associations. Finally, PYPE can identify possible causal relationships between phenotypes using MR under a variety of causal effect modeling scenarios.

INTRODUCTION

While genome-wide association studies (GWASs) have been critical for determining the relationship between genetic variants along the genome and a phenotype, phenome-wide association studies (PheWASs) allow investigators to explore the relationship between phenotypes along the "phenome" and a genetic variant.[1] PheWASs have been shown to replicate known GWAS results[2] and to improve upon GWASs by helping to identify shared biological mechanisms across phenotypes (known as pleiotropy), reveal previously unknown associations between phenotypes, and identify new associations.[3,4] The development of tools that can take advantage of biobank-scale data sources for performing PheWASs is thus a prerequisite for interpreting and contextualizing genotype-phenotype associations.

Prior PheWAS software includes R PheWAS,[5] PyPheWAS,[6] and DeepPheWAS,[7] the latter of which uses Plink2, a popular open-source association analysis toolset,[8] to accelerate their analyses. Packages also exist for visualizing PheWAS results, such as PheWAS-View[9] and PheWeb.[10] While these tools are useful for quick lookups, they are not suitable for "bespoke" analysis or in a wide variety of use cases where analysts are re-analyzing a subset of the population, such as subsets of the population with image or multi-omics data (e.g., https://www.ukbiobank.ac.uk/enable-your-research/approved-research). In these research studies, no PheWAS results may exist.

In summary, none of these tools by themselves provide an end-to-end pipeline for running a PheWAS, visualizing the results, and interpreting potential causal effects. To this end, we developed a Python-based implementation of a PheWAS tool, including

downstream functionality that annotates the significant results with nearby genes and functional relevance according to external variant information databases and can natively run Mendelian randomization (MR), a method that uses genetic variants to approximate a randomized control trial to make causal inferences about the relationship between an exposure and an outcome.[11] While packages for MR exist, such as Mendelian Randomization[12] and TwoSampleMR,[13] none of them are incorporated into any previous PheWAS workflow, even though it is a very typical analysis to run along with PheWASs. To this end, we present PYPE, a computational tool optimized for UK Biobank (UKBB) data that is designed to accelerate PheWASs at every step of the process.

RESULTS AND DISCUSSION

Overview of PYPE and features

PYPE facilitates the three main components of a typical PheWAS analysis. First, the user inputs a file containing the independent variables (which may be a set of genotype or phenotype files), a file containing the dependent variables (the set of phenotypes provided by the UKBB), and any other optional arguments (e.g., covariates) and runs the PheWAS analysis. Next, the user can choose to visualize the PheWAS results in a variety of plot types, highlighting significant associations that were found in the PheWAS. Finally, the user can explore the functional consequences of the significantly associated variants and the genes these variants are physically close to, as well as run MR analyses to uncover possible causal relationships for the phenotypes of interest.

Automating PheWAS analysis

The UKBB is a large-scale medical database that provides in-depth genetic, health, and lifestyle data for around half a million UK residents, allowing for a wide variety of exploratory analyses that facilitate biological discovery.[14] With this in mind, we decided to focus PYPE's compatibility with this widely used resource to provide the most useful tool for most researchers. With the data provided by the UKBB, users can input either genotype or phenotype data as the independent variables and phenotype data as the dependent variables in the PheWAS. Users can also specify covariates such as age, sex, genetic principal components, or other phenotypes provided by the UKBB for use in the PheWAS. Furthermore, PYPE scrapes the UKBB website to annotate the specified phenotypes with description and categorization information, allowing users to specify broad classes of phenotypes directly using a link from the UKBB showcase. Although PYPE is best suited for use directly with the UKBB, it is important to note that data from other sources can also be used to run PheWASs in PYPE. As long as the genomic data are stored in the widely used BED/BIM/FAM file formats and the phenotypes are located in one file with an ID column linking the two, PYPE can be run on any dataset. Further information on how to run the pipeline with alternate data sources can be found at the provided GitHub repository (https://github.com/TaykhoomDalal/pype/tree/main?tab=readme-ov-file#non_ukbb).

When working with data from the UKBB, there are more than 7,000 phenotypes available to run the PheWAS with, encompassing a wide array of health-related outcomes including clinical diagnoses, physical measurements, and biochemical markers.[15] The data are present in a variety of formats, including binary, continuous, and unstructured formats, which provide the end user with ample resources for investigating the relationships between the target genetic variants and phenotypes and revealing potential shared biological mechanisms. This breadth of data is well suited for running PheWASs and is the reason why PYPE provides several UKBB-specific features to accelerate and detect associations that may not have been previously explored. Furthermore, unlike several existing tools, PYPE also does not exclusively focus on running PheWASs using ICD diagnosis codes (which are used as proxies for disease status) but rather provides the user the ability to utilize any type of trait, which could even include creating new phenotypes from existing ones to use for the PheWAS analysis.

PYPE currently uses the default linear regression model from statsmodels in Python to run the mass multivariate regressions (with added support for parallel processing), with phenotypes as the dependent variables and genotypes (or phenotypes) as the independent variables, outputting p values, beta coefficients, and standard errors. Here, the beta coefficients that are produced for each variant-phenotype association indicate whether having the variant increases or decreases the phenotype's value, a fact that can be used to interpret the biological impact of the variant on the target phenotype. PYPE also allows the user to specify the maximum allowed missingness rate or minimum sample requirements for the phenotypes included in the PheWAS to ensure that the analyses are based on enough data. This helps mitigate the effects of data incompleteness, which could otherwise affect the accuracy of the regression results and the subsequent interpretation of the associations. To correct for the occurrence of false positives, users can also specify multiple testing correction methods including Bonferroni correction, false discovery rate correction, Sidak correction, or simple p value correction methods. For each of these multiple testing correction methods, the total number of association tests run is used as an estimation for the space of tests. To verify the accuracy of our code, we ran our PheWAS method and R PheWAS on the data used in the exploratory analysis section below and got the same results. The output can be found in Tables S1 and S2.

Enabling customizable PheWAS visualizations

PYPE also provides a script to generate a variety of visualizations once the PheWAS has been run. Users can choose between generating traditional Manhattan plots, category enrichment plots, and volcano plots. When generating Manhattan plots, users can specify custom groupings for the phenotypes or use the categories provided by the UKBB, include annotations for the significant associations, alter plot features such as color maps and transparency, and plot aggregate-level Manhattan plots (Figure 1) or individual categories with significant associations (Figure 3). Users can also generate category enrichment plots as in Figure 2, plotting the enrichment of significant associations per category in a barplot. Finally, users can generate volcano plots for each independent variable, displaying the significant associations between the variable and the dependent phenotypes, with the ability to annotate the most significant associations as well, as shown in Figure S1.

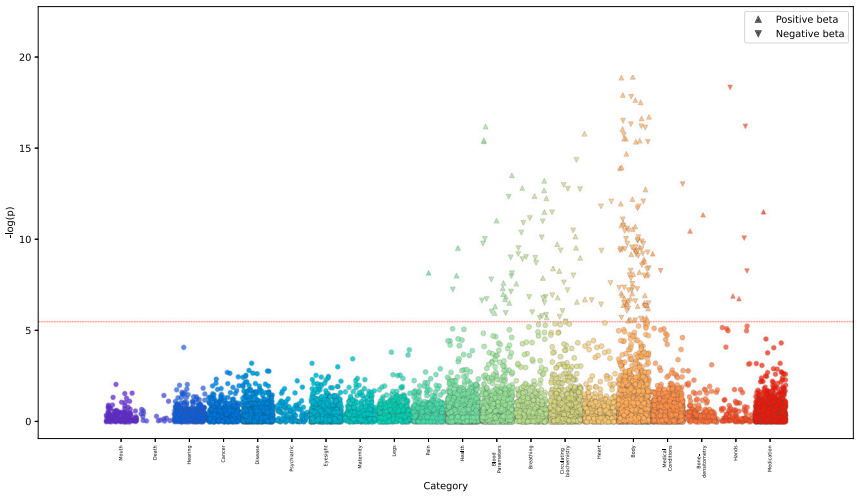

Figure 1. Manhattan plot of aggregate PheWAS results across abdomen, liver, and pancreas age for each of the condensed categories of phenotype fields used in the UKBB from Table S3

Manhattan plot of aggregate PheWAS results across abdomen, liver, and pancreas age for each of the condensed categories of phenotype fields used in the UK Biobank from Table S3, where the y-axis represents the negative logarithm (base 10) of the p values, illustrating the statistical significance of each association.

Enhancing PheWAS studies with downstream annotations and MR

If inputting genotype variables, then PYPE provides downstream annotations for the significant associations that are found in the PheWAS. Specifically, PYPE provides three main downstream functions. First, if the user specifies a gene file (which contains information about the location of genes on each chromosome) and the number of kilobases downstream and upstream to look for genes, PYPE can be used to annotate each variant with which genes are close to it on the chromosome. Secondly, users can choose whether to generate summary files for the significant variants using MyVariantInfo, and MyGeneInfo (RESTful API) queries to a variety of databases that contain functional variant annotations and their clinical significance (such as ClinVar, dbSNP, CADD, etc.) for information about the given variant and its associated genes.[16] PYPE will then use this information to create files for each unique, significant variant, describing its functional consequence and the function of the genes it is close to (if this information exists). Lastly, the user can choose to run a two-sample MR using the genotype data against a second dataset of GWAS results that is either user specified or, if no dataset is provided, queried from the Open GWAS Project using their RESTful API.[17,18] Similar to the TwoSampleMR package, the user has a choice between a variety of MR methods, including inverse-variance weighted regression, Egger regression, various median- and mode-based methods, and the pleiotropy residual sum and outlier method, providing the ability to assess causal relationships as well as identify bias due to pleiotropic effects.[19,20] To verify the accuracy of our code, we have provided a comparison between the output of our MR methods and the outputs from TwoSampleMR in Tables S4–S19. This comparison is based on the data used in the following exploratory analysis.

Applying PYPE for an exploratory analysis

To demonstrate the functionality of our tool, we applied PYPE on the results of abdomen biological age predictors,[21] exploring the genetic architecture of these predictors with cardiometabolic diseases such as type 2 diabetes on an out-of-sample dataset from the UKBB. Using the 16 genetic variants implicated in accelerated abdomen, liver, and pancreas aging, we ran the PheWAS against several broad classes of phenotypes found in the UKBB, including physical lab measurements, self-reported medical conditions, linked health outcomes, blood biochemistry, blood count (parameters), and infectious disease antigens. These classes were then manually split into 20 categories based on their labeling in the UKBB, providing a condensed classification of the many phenotype data fields, which can be found in Table S3. This condensed classification can be tweaked by the user to either further condense the categories, change the categories altogether, or leave them as is, directly from the UKBB website. We ran the PheWAS with sex, age, ethnicity, and all 40 genetic principal components as covariates, generating 15,082 total associations and 290 significant associations ($p < 3.32e{-}06$ Bonferroni corrected with original alpha = 0.05). The significant results can be found in Table S20. Finally, we ran MR to assess the possible causal relationship between accelerated abdomen, liver, and pancreas aging and other phenotypes that have been linked to cardiometabolic diseases.

Figure 2 highlights the percentage of significant associations per category, illustrating which phenotypic categories have the greatest percentage of significant associations. We observe the greatest number of significant associations in the body category, a category that was defined to encapsulate many of the physical lab measures that have to do with adiposity, height, and weight phenotypes. Furthermore, other top categories include breathing and circulating biochemistry, where circulating biochemistry chiefly contains information about cardiometabolic biomarkers such as alanine aminotransferase (ALT), aspartate aminotransferase (AST), gamma-glutamyl transferase, etc. For this demonstration, we will focus on the circulating biochemistry category results for the accelerated-liver-aging-associated variants listed in Table S21, as many of the mentioned enzymes are liver specific.

Out of the various associations in Figure 3, we observed significant associations for high-density lipoprotein cholesterol (HDL-C), AST, and ALT. Here, the association between variant rs13107325 and HDL-C indicates that decelerated aging is associated with lower HDL-C. Using the downstream functionality of PYPE, where the base-pair position of the variant on the chromosome is used to map variants to nearby genes (based on user-specified distances), we see that the associated gene for this variant is *SLC39A8*. Another significant association was found between AST and

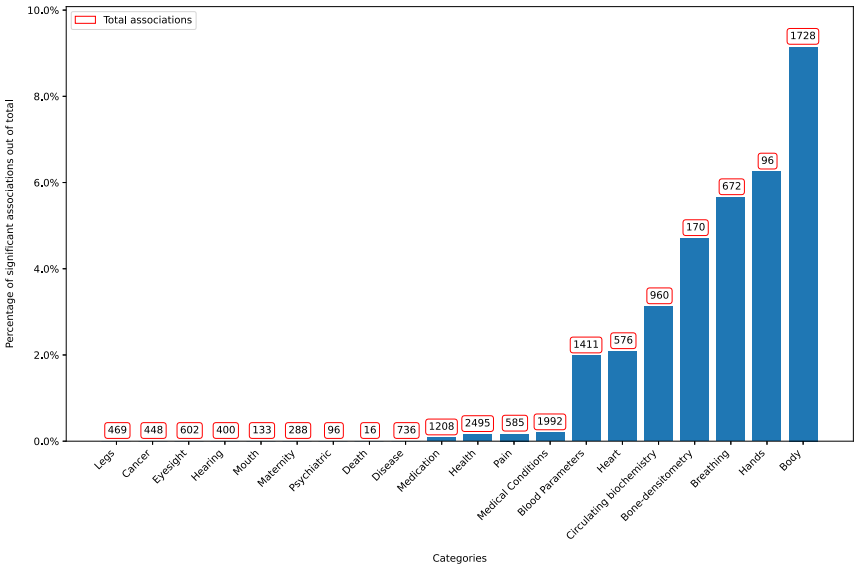

Figure 2. Percentage of each category in terms of the number of significant associations

The number of total associations found is labeled above each category.

variant rs13107325, with the PheWAS results indicating that with decelerated aging, AST levels are higher. Lastly, the PheWAS found a significant association between the variant rs370844658 and ALT, suggesting that with accelerated aging, ALT levels increase. Using the variant-gene mapping functionality, we find that the associated gene for this variant is *EIF2S2*. The full list of these results can be found in Table S22.

After running the PheWAS, annotating the significant results with nearby genes, and visualizing the results in various plot types, we ran MR for the abdomen and liver variants against GWAS associations for a variety of phenotypes. The pancreas variants were not used as there were only 2, and most MR methods require at least 3. Inverse-variance weighted regression was used here to estimate the causal effect of the various age predictors and phenotypes that have known associations with cardiometabolic diseases, finding an association ($p < 0.05$) between accelerated abdomen aging variants and waist circumference variants, as shown in Table 1. We have also provided the results for these variants with all of the other MR methods in Tables S4, S6, S8, S10, S12, S14, S16, and S18.

PYPE simplifies PheWAS studies and enables researchers to focus on result interpretation

We present PYPE as an easy-to-use, customizable, and feature-rich tool for running PheWASs from start to finish. PYPE allows researchers to specify the options they want at all stages of the analysis pipeline, from execution, to visualization, to any downstream analysis. With this tool, we can abstract away many portions of a typical PheWAS, allowing more time to be spent on validating and exploring the consequences of the results generated, in contrast to the aforementioned tools.[7] However, there are some limitations to the current implementation. Chiefly, the tool is currently best optimized for the UKBB dataset, and thus some of the functionalities of the tool (querying for phenotype information from the UKBB showcase) are not suited for other datasets. Furthermore, the visualization capabilities are currently limited to static generation of the plots, and the user has to regenerate the results when they want to change an attribute of the plot. However, with the availability of tools such as PheWEB, the latter issue is less of a problem, as PYPE output can be adapted

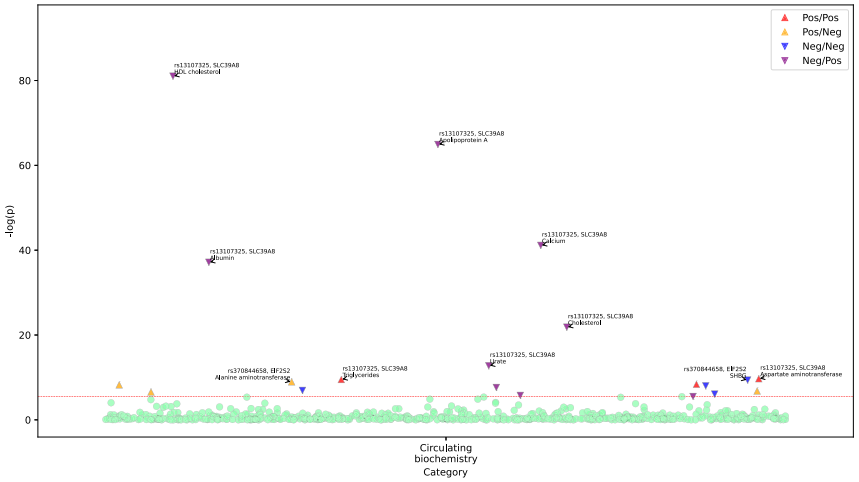

Figure 3. Associations between accelerated liver aging variants and circulating biochemistry biomarkers, where the y-axis represents the negative logarithm (base 10) of the *p* values, illustrating the statistical significance of each association

Note that the directionality of the arrow indicates the sign of the effect size of the PheWAS association, and the color indicates the sign of the effect size of the variant's association with the aging phenotype defined in Le Goallec et al.[21] There, a negative effect size indicated accelerated aging and a positive effect size indicated decelerated aging.

Table 1. Mendelian randomization (inverse-variance weighted) results for genetic variants implicated in accelerated abdomen and liver aging and GWAS from the UKBB (uncorrected for multiple hypotheses with $p < 0.05$)

Predictor	Phenotype	*p* value	Coefficient	Standard error
Abdomen	hemoglobin A1C	0.194	0.027	0.020
Abdomen	body mass index	0.958	0.006	0.116
Abdomen	glucose	0.254	−0.015	0.013
Abdomen	waist circumference	0.001	0.356	0.109
Liver	hemoglobin A1C	0.729	0.015	0.044
Liver	body mass index	0.921	−0.012	0.123
Liver	glucose	0.358	0.007	0.007
Liver	waist circumference	0.110	0.251	0.157

to support interoperability with this tool as well as PheWAS-View, giving the user a variety of visualization choices based on their use case. Lastly, the issue of multiple testing correction of the MR associations is left unaddressed in PYPE, with current literature[22] indicating that conservative approaches to multiple testing correction may be too strict, largely due to the low power of these studies and the fact that the relationships under investigation typically have prior biological support. We acknowledge that PYPE is agnostic to biological claims, and therefore, we recommend that associations are reported with false discovery and/or family-wise error rate control. Future versions of PYPE will provide integration with additional data formats and sources, introduce more comprehensive result annotations such as gene pathway involvement and other known associations to aid in the interpretation of PheWAS findings, and improve the speed of pipeline components. Furthermore, we will add new visualization functions and a wider breadth of MR functions and support more varied association types (i.e., polygenic risk score-phenotype associations), facilitating greater customization for the end user. To keep up with updates to the pipeline, be sure to visit our GitHub, where new releases with feature updates will be posted. Tools such as PYPE are essential for reducing the time researchers spend on the data analysis stage and increasing focus on the result interpretation, a critical stage in the workflow for PheWASs, as it is often difficult to differentiate between spurious and true associations.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Dr. Chirag J. Patel can be reached by email (chirag_patel@hms.harvard.edu).

Materials availability

This study did not generate new materials.

Data and code availability

The data used in the example analysis can be obtained from the UKBB, and the PYPE tool is publicly available in a Github repository (https://github.com/TaykhoomDalal/pype). PYPE is published under the Apache 2.0 license, and supporting documentation can be found at the aforementioned link. The source code is also archived in Zenodo[23] (https://doi.org/10.5281/zenodo.10883968). The hemoglobin A1C, BMI, glucose, and waist circumference variants used in the MR study can be found at the Neale Lab Github (https://github.com/Nealelab/UK_Biobank_GWAS), and the abdomen, liver, and pancreas variants can be found in the paper by Le Goallec et al.[21]

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2024.100982.

ACKNOWLEDGMENTS

We would like to thank the Harvard Medical School research computing group for access and utilization of the O2 cluster. We also want to acknowledge the UKBB for providing us with access to the data they collected under project number 52887. This work was supported by NIH NIEHS R01ES032470.

AUTHOR CONTRIBUTIONS

T.D. and C.J.P. designed the study. T.D. was involved in creating software and literature search and wrote the first version of the manuscript. C.J.P. wrote and revised the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 30, 2023
Revised: November 30, 2023
Accepted: April 8, 2024
Published: May 1, 2024

REFERENCES

1. Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D.R., Roden, D.M., and Crawford, D.C. (2010). PheWAS: Demonstrating the Feasibility of a Phenome-Wide Scan to Discover Gene-Disease Associations. Bioinformatics *26*, 1205–1210. https://doi.org/10.1093/bioinformatics/btq126.
2. Denny, J.C., Bastarache, L., Ritchie, M.D., Carroll, R.J., Zink, R., Mosley, J.D., Field, J.R., Pulley, J.M., Ramirez, A.H., Bowton, E., et al. (2013). Systematic Comparison of Phenome-Wide Association Study of Electronic Medical Record Data and Genome-Wide Association Study Data. Nat. Biotechnol. *31*, 1102–1110. https://doi.org/10.1038/nbt.2749.
3. Diogo, D., Tian, C., Franklin, C.S., Alanne-Kinnunen, M., March, M., Spencer, C.C.A., Vangjeli, C., Weale, M.E., Mattsson, H., Kilpeläinen, E., et al. (2018). Phenome-Wide Association Studies across Large Population Cohorts Support Drug Target Validation. Nat. Commun. *9*, 4285. https://doi.org/10.1038/s41467-018-06540-3.
4. Tyler, A.L., Crawford, D.C., and Pendergrass, S.A. (2016). The Detection and Characterization of Pleiotropy: Discovery, Progress, and Promise. Briefings Bioinf. *17*, 13–22. https://doi.org/10.1093/bib/bbv050.
5. Carroll, R.J., Bastarache, L., and Denny, J.C. (2014). R PheWAS: Data Analysis and Plotting Tools for Phenome-Wide Association Studies in the R Environment. Bioinformatics *30*, 2375–2376. https://doi.org/10.1093/bioinformatics/btu197.
6. Kerley, C.I., Chaganti, S., Nguyen, T.Q., Bermudez, C., Cutting, L.E., Beason-Held, L.L., Lasko, T., and Landman, B.A. (2022). pyPheWAS: A Phenome-Disease Association Tool for Electronic Medical Record Analysis. Neuroinformatics *20*, 483–505. https://doi.org/10.1007/s12021-021-09553-4.
7. Packer, R.J., Williams, A.T., Hennah, W., Eisenberg, M.T., Shrine, N., Fawcett, K.A., Pearson, W., Guyatt, A.L., Edris, A., Hollox, E.J., et al. (2023). DeepPheWAS: An R Package for Phenotype Generation and Association Analysis for Phenome-Wide Association Studies. Bioinformatics *39*, btad073. https://doi.org/10.1093/bioinformatics/btad073.
8. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets. GigaScience *4*, 7. https://doi.org/10.1186/s13742-015-0047-8.
9. Pendergrass, S.A., Dudek, S.M., Crawford, D.C., and Ritchie, M.D. (2012). Visually Integrating and Exploring High Throughput Phenome-Wide

Association Study (PheWAS) Results Using PheWAS-View. BioData Min. *5*, 5. https://doi.org/10.1186/1756-0381-5-5.

10. Gagliano Taliun, S.A., VandeHaar, P., Boughton, A.P., Welch, R.P., Taliun, D., Schmidt, E.M., Zhou, W., Nielsen, J.B., Willer, C.J., Lee, S., et al. (2020). Exploring and Visualizing Large-Scale Genetic Associations by Using PheWeb. Nat. Genet. *52*, 550–552. https://doi.org/10.1038/s41588-020-0622-5.
11. Ebrahim, S., and Davey Smith, G. (2008). Mendelian Randomization: Can Genetic Epidemiology Help Redress the Failures of Observational Epidemiology? Hum. Genet. *123*, 15–33. https://doi.org/10.1007/s00439-007-0448-6.
12. Yavorska, O.O., and Burgess, S. (2017). MendelianRandomization: An R Package for Performing Mendelian Randomization Analyses Using Summarized Data. Int. J. Epidemiol. *46*, 1734–1739. https://doi.org/10.1093/ije/dyx034.
13. Hemani, G., Tilling, K., and Davey Smith, G. (2017). Orienting the Causal Relationship between Imprecisely Measured Traits Using GWAS Summary Data. PLoS Genet. *13*, e1007081. https://doi.org/10.1371/journal.pgen.1007081.
14. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. *12*, e1001779. https://doi.org/10.1371/journal.pmed.1001779.
15. Constantinescu, A.E., Mitchell, R.E., Zheng, J., Bull, C.J., Timpson, N.J., Amulic, B., Vincent, E.E., and Hughes, D.A. (2022). A Framework for Research into Continental Ancestry Groups of the UK Biobank. Hum. Genom. *16*, 3. https://doi.org/10.1186/s40246-022-00380-5.
16. Lelong, S., Zhou, X., Afrasiabi, C., Qian, Z., Cano, M.A., Tsueng, G., Xin, J., Mullen, J., Yao, Y., Avila, R., et al. (2022). BioThings SDK: A Toolkit for Building High-Performance Data APIs in Biomedical Research. Bioinformatics *38*, 2077–2079. https://doi.org/10.1093/bioinformatics/btac017.
17. Pierce, B.L., and Burgess, S. (2013). Efficient Design for Mendelian Randomization Studies: Subsample and 2-Sample Instrumental Variable Estimators. Am. J. Epidemiol. *178*, 1177–1184. https://doi.org/10.1093/aje/kwt084.
18. Elsworth, B., Lyon, M., Alexander, T., Liu, Y., Matthews, P., Hallett, J., Bates, P., Palmer, T., Haberland, V., Davey, S.G., et al. (2020). The MRC IEU OpenGWAS Data Infrastructure. Preprint at bioRxiv. https://doi.org/10.1101/2020.08.10.244293.
19. Bowden, J., Davey Smith, G., Haycock, P.C., and Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genet. Epidemiol. *40*, 304–314. https://doi.org/10.1002/gepi.21965.
20. Verbanck, M., Chen, C.Y., Neale, B., and Do, R. (2018). Detection of Widespread Horizontal Pleiotropy in Causal Relationships Inferred from Mendelian Randomization between Complex Traits and Diseases. Nat. Genet. *50*, 693–698. https://doi.org/10.1038/s41588-018-0099-7.
21. Le, G.A., Diai, S., Collin, S., Prost, J.-B., Vincent, T., and Patel, C.J. (2022). Using Deep Learning to Predict Abdominal Age from Liver and Pancreas Magnetic Resonance Images. Nat. Commun. *13*, 1979. https://doi.org/10.1038/s41467-022-29525-9.
22. Burgess, S., Davey Smith, G., Davies, N.M., Dudbridge, F., Gill, D., Glymour, M.M., Hartwig, F.P., Kutalik, Z., Holmes, M.V., Minelli, C., et al. (2019). Guidelines for Performing Mendelian Randomization Investigations: Update for Summer 2023. Wellcome Open Res. *4*, 186. https://doi.org/10.12688/wellcomeopenres.15555.3.
23. Dalal, T., and Patel, C. (2024). PYPE: A Python pipeline for phenome-wide association and mendelian randomization in investigator-driven phenotypes and genotypes of biobank data. Zenodo. https://doi.org/10.5281/zenodo.10883968.