Bioinformatics, 39(4), 2023, btad073
https://doi.org/10.1093/bioinformatics/btad073
Advance Access Publication Date: 6 February 2023
Applications Note

Genetics and population analysis

DeepPheWAS: an R package for phenotype generation and association analysis for phenome-wide association studies

Richard J. Packer [1,2,*], Alex T. Williams[1], William Hennah[3,4,5], Micaela T. Eisenberg[6], Nick Shrine[1], Katherine A. Fawcett[1], Willow Pearson[1], Anna L. Guyatt [1], Ahmed Edris[1,7], Edward J. Hollox[6], Mikko Marttila[3], Balasubramanya S. Rao[8], John Raymond Bratty[8], Louise V. Wain[1,2], Frank Dudbridge [1] and Martin D. Tobin[1,2]

[1]Department of Population Health Sciences, University of Leicester, Leicester LE1 7RH, UK, [2]Leicester National Institute of Health and Care Research Biomedical Research Centre, Glenfield Hospital, Leicester LE5 4PW, UK, [3]Orion Pharma, Espoo 02200, Finland, [4]Neuroscience Center, HiLIFE, University of Helsinki, Helsinki 00100, Finland, [5]Institute for Molecular Medicine FIMM, HiLIFE, University of Helsinki, Helsinki 00100, Finland, [6]Department of Genetics and Genome Biology, University of Leicester, Leicester LE1 7RH, UK, [7]Department of Bioanalysis, Faculty of Pharmaceutical Sciences, Ghent University, Gent 9000, Belgium and [8]Orion Pharma, Nottingham NG1 2GB, UK

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on August 30, 2022; revised on January 6, 2023; editorial decision on January 29, 2023; accepted on February 3, 2023

Abstract

Summary: DeepPheWAS is an R package for phenome-wide association studies that creates clinically curated composite phenotypes and integrates quantitative phenotypes from primary care data, longitudinal trajectories of quantitative measures, disease progression and drug response phenotypes. Tools are provided for efficient analysis of association with any genetic input, under any genetic model, with optional sex-stratified analysis, and for developing novel phenotypes.

Availability and implementation: The DeepPheWAS R package is freely available under GNU general public licence v3.0 from at https://github.com/Richard-Packer/DeepPheWAS.

Contact: richard.packer@leicester.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Phenome-wide association studies (PheWASs) can be used to better understand the pleiotropic effects of genetic variants (Tyler *et al.*, 2016) and to inform drug development through target identification, target validation and use of variants that mimic drug effects to assess likely drug efficacy, safety and drug repurposing opportunities (Diogo *et al.*, 2018; Gill *et al.*, 2019; Khosravi *et al.*, 2019). PheWASs comprise two stages—phenotype generation and statistical association tests. There have been two widely applicable methods for phenotype generation: PHESANT (Millard *et al.*, 2018) and PheWAS-R (Carroll *et al.*, 2014). PHESANT creates phenotypes by extracting study-specific questionnaire and measurement data alongside linked hospital records in UK Biobank. PheWAS-R combines related international classification of disease version 9 and 10 (ICD-9/ICD-10) codes into clinically relevant groups termed phecodes. Both tools provide regression analysis for per-variant PheWAS for generated phenotypes using generalized linear models in R and produce Manhattan plots. Online PheWAS resources such as Open Targets Genetics (Ghoussaini *et al.*, 2021) do not perform new statistical tests. Instead, they are repositories for existing results from phenotypes generated by one of the two above tools or by individual genome-wide association studies (GWAS).

These tools are useful, but have several key gaps:

1. The phenotypes generated rely on a single data field or coding ontology and do not take advantage of all available data, such as primary care data;
2. Existing approaches do not provide tools for developing new phenotypes;

3. For running per-variant PheWAS, running each regression model in R is computationally inefficient and can result in inflated type I error for low-frequency variants with a case–control imbalance (Ma *et al.*, 2013);
4. Online resources such as Open Targets Genetics have limited flexibility. For example, they accept only single nucleotide polymorphisms (SNPs) and retrieve results only for genetic models tested. The user cannot specify when to use new data fields or updates to existing data fields (e.g. updated health records), and the user cannot specify their preferred statistical approach and outputs, such as false discovery rate (FDR).

The platform we have developed, DeepPheWAS, addresses both phenotype generation and efficient association testing while incorporating the following developments that are not yet available in any single current platform or online resource:

i. Clinically curated composite phenotypes for selected health conditions that integrate different data types (including primary and secondary care data) to study phenotypes not well captured by current classification trees;
ii. Integration of quantitative phenotypes from primary care data, such as pathology records and clinical measures;
iii. Integration of disease progression phenotypes, longitudinal trajectories of quantitative measures and drug response measures;
iv. Clinically curated phenotype selection for traits that are extremely highly correlated;
v. Efficient association testing, and type-1 error control using PLINK 2 firth fall-back regression.
vi. Flexible tests of additive, dominant, recessive and genotypic models;
vii. Inclusion of complex variants, such as copy number variants with a wide range of copy numbers (multiallelic CNVs);
viii. Ability to test genetic risk scores;
ix. Creation of phenotypes in sex-specific strata to run a sex-stratified PheWAS;
x. Providing tools for generating novel phenotypes using a simple phenotype mapping process.

2 Application of DeepPheWAS to UK biobank

2.1 Analysis of quantitative phenotypes

Our package can be applied to quantitative phenotypes derived from numerous data sources, including primary care data. For example, we created a phenotype using recorded levels of blood sodium in primary care records that is not yet included in any PheWAS platform. We applied DeepPheWAS to rs7193778 (nearest genes *NFAT5* and *TERF2*), previously associated with urate levels (Köttgen *et al.*, 2013). Our PheWAS shows various associations which are currently not documented in GWAS Catalog, most strongly with blood sodium levels (Supplementary Fig. S1, Supplementary Table S1).

2.2 Highly correlated traits

We applied our DeepPheWAS approach to rs2912062 (nearest genes *ANGPT2* and *AGPAT5*), shown to be associated with carotid intima-media thickness (IMT) (Strawbridge *et al.*, 2020), a phenotype not currently available in any PheWAS platform. By selecting a single representative measure taken from many individual measurements, DeepPheWAS can collapse highly correlated quantitative traits into single measures (in this case carotid IMT maximum and carotid IMT mean), reducing redundancy and improving power. We recapitulated known GWAS findings (Supplementary Fig. S2, Supplementary Table S2).

2.3 Flexibility with choice of genetic model

To demonstrate the flexibility of DeepPheWAS we assessed the *SERPINA1* Z allele [rs28929474(T)] under additive and recessive genetic models. Patterns of association differed between models, including opposite estimated effect directions for this variant on forced expiratory volume in 1 s (FEV_1) and forced vital capacity (FVC), consistent with previous findings (Fawcett *et al.*, 2021a) (Supplementary Figs S3 and S4, Supplementary Tables S3 and S4).

2.4 Association tests for complex structural variation

Human genomic variation includes variants which have more categories than SNPs. For example, the diploid human copy number of *CCL3L1* ranges from 0 to 8 in UK Biobank participants (Fawcett *et al.*, 2022) (Supplementary Fig. S5). In such situations, association testing may be based on the measured copy number or on user-specified collapsed categories, requiring a flexible platform. We used DeepPheWAS to test association with *CCL3L1* copy number (coded 0–8) under a linear additive model; no associations reached an FDR threshold of 1% (Supplementary Fig. S6, the top 5 associations are shown in Supplementary Table S5), this recapitulates findings from earlier studies (Adewoye *et al.*, 2018; Carpenter *et al.*, 2011; Field *et al.*, 2009; Urban *et al.*, 2009).

2.5 Genetic risk scores, composite and disease-progression phenotypes

Genetic risk scores (GRS) aggregate multiple SNPs, providing improved power for studying phenotypic associations, but cannot be specified in online PheWAS platforms. We performed a PheWAS using a 279-variant GRS for FEV_1/FVC (Shrine, 2019), which showed association ($FDR < 1\%$) with 47 traits including increased risk of clinical COPD and clinical asthma with a higher score of FEV_1/FVC reducing alleles (Supplementary Fig. S7, Supplementary Table S6). Furthermore, the composite phenotypes generated by the DeepPheWAS platform (e.g. P2020 Asthma and P2054 COPD) were consistently more strongly associated with the GRS for FEV_1/FVC than the relevant Phecodes alone. We also show significant association with the novel disease-progression phenotypes: exacerbation of COPD and age-of-onset of COPD both of which are unavailable in existing PheWAS resources and have published GWAS results.

2.6 Sex-stratified PheWAS

We applied DeepPheWAS to rs12777332 (nearest genes *CASP7* and *NRAP*) and rs7697189 (nearest gene *HHIP*). We replicate association for risk of cataract in women only (Supplementary Fig. S8, Supplementary Table S7) for rs12777332 (Choquet *et al.*, 2021), and for increased effect on FEV_1 in men compared to women (Supplementary Figs S9 and S10, Supplementary Table S8) for rs7697189 (Fawcett *et al.*, 2021b).

3 Implementation

DeepPheWAS is an R package that can be run on high-performance computing clusters and requires R 4.1.0 and PLINK 2.0. DeepPheWAS is optimized for UK Biobank data and is expected to be interoperable with the UK Biobank Research Analysis Platform, further details on required data can be seen in the supplement and on https://richard-packer.github.io/DeepPheWAS_site/.

4 Availability

The DeepPheWAS R package is freely available under GNU general public licence v3.0 from at https://github.com/Richard-Packer/DeepPheWAS.

5 Conclusion

Here, we present DeepPheWAS, an R package that facilitates phenome-wide association studies while addressing several limitations of existing approaches. This includes the ability to analyse a broader range of phenotypes derived from large-scale electronic healthcare records, more informative composite phenotypes, greater flexibility in the type of genetic variation that can be studied and assessing associations with genetic risk scores.

Acknowledgements

This research has been conducted using UK Biobank Resource under Application Number 43027. This research used the SPECTRE High Performance Computing Facility at the University of Leicester.

Funding

This work was supported by Orion Pharma who funded a research collaboration with the University of Leicester, which supported the development of DeepPheWAS; Wellcome Trust Institutional Strategic Support Fund [204801/Z/16/Z], BHF Accelerator Award [AA/18/3/34220] to A.L.G.; GSK/Asthma + Lung UK Chair in Respiratory Research (C17-1) to L.V.W.; Wellcome Trust Investigator Award [WT202849/Z/16/Z to M.D.T.]; M.D.T. holds an National Institute for Health and Care Research Senior Investigator Award. The respiratory research was partially supported by BREATHE—The Health Data Research Hub for Respiratory Health [MC_PC_19004]. The research was partially supported by the National Institute for Health and Care Research Leicester Biomedical Research Centre; views expressed are those of the author(s) and not necessarily those of the NHS, the National Institute for Health and Care Research or the Department of Health. This research was funded in part by the Wellcome Trust. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Conflict of interest: M.D.T., R.P., A.T.W., F.D. and L.V.W. received funding from Orion Pharma within the scope of the submitted work and from GSK for collaborative research projects outside of the submitted work. W.H., B.S.R., M.M. and R.B. are salaried employees of Orion Pharma within the scope of the submitted work. L.V.W. has performed consultancy for Galapagos and had travel paid for by Genetech outside the scope of this work.

Data availability

The data underlying this article are available from the UK Biobank to all approved researchers https://www.ukbiobank.ac.uk/. Data were accessed under approved application 43027.

References

Adewoye,A.B. *et al.* (2018) Human CCL3L1 copy number variation, gene expression, and the role of the CCL3L1-CCR5 axis in lung function. *Wellcome Open Res.*, **3**, 13.

Carpenter,D. *et al.* (2011) Accuracy and differential bias in copy number measurement of CCL3L1 in association studies with three auto-immune disorders. *BMC Genomics*, **12**, 418.

Carroll,R.J. *et al.* (2014) R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics*, **30**, 2375–2376.

Choquet,H. *et al.*; 23andMe Research Team (2021) A large multiethnic GWAS meta-analysis of cataract identifies new risk loci and sex-specific effects. *Nat. Commun.*, **12**, 3595.

Diogo,D. *et al.* (2018) Phenome-wide association studies across large population cohorts support drug target validation. *Nat. Commun.*, **9**, 4285.

Fawcett,K.A. *et al.* (2022) Exome-wide analysis of copy number variation shows association of the human leukocyte antigen region with asthma in UK Biobank. *BMC Med. Genomics*, **15**, 119.

Fawcett,K.A. *et al.* (2021a) Pleiotropic associations of heterozygosity for the *SERPINA1* Z allele in the UK biobank. *ERJ Open Res.*, **7**, 00049-2021.

Fawcett,K.A. *et al.* (2021b) Variants associated with HHIP expression have sex-differential effects on lung function. *Wellcome Open Res.*, **5**, 111.

Field,S.F. *et al.* (2009) Experimental aspects of copy number variant assays at CCL3L1. *Nat. Med.*, **15**, 1115–1117.

Ghoussaini,M. *et al.* (2021) Open targets genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.*, **49**, D1311–D1320.

Gill,D. *et al.* (2019) Use of genetic variants related to antihypertensive drugs to inform on efficacy and side effects. *Circulation*, **140**, 270–279.

Khosravi,A. *et al.* (2019) Active repurposing of drug candidates for melanoma based on GWAS, PheWAS and a wide range of omics data. *Mol. Med.*, **25**, 30.

Köttgen,A. *et al.*; MAGIC Consortium (2013) Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nat. Genet.*, **45**, 145–154.

Ma,C. *et al.*; GoT2D Investigators. (2013) Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants: joint and meta-analysis of low-count variants. *Genet. Epidemiol.*, **37**, 539–550.

Millard,L.A. *et al.* (2018) Software application profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. *Int. J. Epidemiol.*, **47**, 29–35.

Shrine,N. *et al.* (2019) New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat. Genet.*, **51**, 481–493.

Strawbridge,R.J. *et al.* (2020) Carotid intima-media thickness: novel loci, sex-specific effects, and genetic correlations with obesity and glucometabolic traits in UK Biobank. *Arterioscler. Thromb. Vasc. Biol.*, **40**, 446–461.

Tyler,A.L. *et al.* (2016) The detection and characterization of pleiotropy: discovery, progress, and promise. *Brief. Bioinform.*, **17**, 13–22.

Urban,T.J. *et al.* (2009) CCL3L1 and HIV/AIDS susceptibility. *Nat. Med.*, **15**, 1110–1112.