SOFTWARE TOOL ARTICLE

REVISED The EWAS Catalog: a database of epigenome-wide association studies [version 2; peer review: 2 approved]

Thomas Battram[1,2], Paul Yousefi[1,2], Gemma Crawford[1,2], Claire Prince[1,2], Mahsa Sheikhali Babaei[1,2], Gemma Sharp[1,2], Charlie Hatcher[1,2], María Jesús Vega-Salas[3], Sahar Khodabakhsh[3], Oliver Whitehurst[2], Ryan Langdon[1,2], Luke Mahoney[2], Hannah R. Elliott[1,2], Giulia Mancano[1,2], Matthew A. Lee[1,2], Sarah H. Watkins[1,2], Abigail C. Lay[4], Gibran Hemani[1,2], Tom R. Gaunt[1,2], Caroline L. Relton[1,2], James R. Staley[1,2*], Matthew Suderman[1,2*]

[1]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, BS8 1TH, UK
[2]Bristol Medical School, University of Bristol, Bristol, BS8 1TH, UK
[3]Centre for Exercise, Nutrition and Health Sciences, University of Bristol, Bristol, BS8 1TH, UK
[4]Bristol Renal, Translational Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

[*] Equal contributors

v2 **First published:** 04 Feb 2022, **7**:41
https://doi.org/10.12688/wellcomeopenres.17598.1

Latest published: 31 May 2022, **7**:41
https://doi.org/10.12688/wellcomeopenres.17598.2

Abstract

Epigenome-wide association studies (EWAS) seek to quantify associations between traits/exposures and DNA methylation measured at thousands or millions of CpG sites across the genome. In recent years, the increase in availability of DNA methylation measures in population-based cohorts and case-control studies has resulted in a dramatic expansion of the number of EWAS being performed and published. To make this rich source of results more accessible, we have manually curated a database of CpG-trait associations (with $p<1x10^{-4}$) from published EWAS, each assaying over 100,000 CpGs in at least 100 individuals. From January 7, 2022, The EWAS Catalog contained 1,737,746 associations from 2,686 EWAS. This includes 1,345,398 associations from 342 peer-reviewed publications. In addition, it also contains summary statistics for 392,348 associations from 427 EWAS, performed on data from the Avon Longitudinal Study of Parents and Children (ALSPAC) and the Gene Expression Omnibus (GEO). The database is accompanied by a web-based tool and R package, giving researchers the opportunity to query EWAS associations quickly and easily, and gain insight into the molecular underpinnings of disease as well as the impact of traits and exposures on the DNA methylome. The EWAS Catalog data extraction team continue to update the database monthly and we encourage any EWAS authors to upload their summary statistics to our website. Details of how to upload data can be found here:

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 31 May 2022		
version 1 04 Feb 2022	✓ view	✓ view

1. **John W. Holloway**, University of Southampton, Southampton, UK
University Hospital Southampton, Southampton, UK
2. **Harold Snieder**, University Medical Center Groningen, Groningen, The Netherlands

Any reports and responses or comments on the article can be found at the end of the article.

http://www.ewascatalog.org/upload.

The EWAS Catalog is available at http://www.ewascatalog.org.

Keywords
EWAS, epigenome-wide, epigenetics, database, ALSPAC

This article is included in the Avon Longitudinal Study of Parents and Children (ALSPAC) gateway.

Corresponding author: Thomas Battram (thomas.battram@bristol.ac.uk)

Author roles: Battram T: Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Yousefi P**: Data Curation, Formal Analysis, Investigation, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Crawford G**: Data Curation, Writing – Review & Editing; **Prince C**: Data Curation, Writing – Review & Editing; **Sheikhali Babaei M**: Data Curation, Writing – Review & Editing; **Sharp G**: Conceptualization, Writing – Review & Editing; **Hatcher C**: Data Curation, Writing – Review & Editing; **Vega-Salas MJ**: Data Curation, Writing – Review & Editing; **Khodabakhsh S**: Data Curation, Writing – Review & Editing; **Whitehurst O**: Data Curation, Writing – Review & Editing; **Langdon R**: Data Curation, Writing – Review & Editing; **Mahoney L**: Data Curation, Writing – Review & Editing; **Elliott HR**: Data Curation, Writing – Review & Editing; **Mancano G**: Data Curation, Writing – Review & Editing; **Lee MA**: Data Curation, Writing – Review & Editing; **Watkins SH**: Data Curation, Writing – Review & Editing; **Lay AC**: Data Curation, Writing – Review & Editing; **Hemani G**: Data Curation, Funding Acquisition, Supervision, Writing – Review & Editing; **Gaunt TR**: Conceptualization, Resources, Software, Supervision, Writing – Review & Editing; **Relton CL**: Conceptualization, Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing; **Staley JR**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Suderman M**: Conceptualization, Data Curation, Formal Analysis, Methodology, Project Administration, Resources, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome (203746, PhD studentship). This work was also supported by the BBSRC and ESRC (grant number ES/N000498/1) and by the UK Medical Research Council (MC_UU_00011/1, MC_UU_00011/4, MC_UU_12013/1, MC_UU_12013/2 and MC_UU_12013/4), which funds a Unit at the University of Bristol, where the majority of the authors work. The UK Medical Research Council and Wellcome (Grant: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. Methylation data in the ALSPAC cohort were generated as part of the UK BBSRC funded (BB/I025751/1 and BB/I025263/1) Accessible Resource for Integrated Epigenomic Studies (ARIES, http://www.ariesepigenomics.org.uk/).The phenotype collection was also in part funded by The British Heart Foundation (SP/07/008/24066), Roche Diagnostics and the National Institute for Health Research (NF-SI-0611-10196). A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Battram T, Yousefi P, Crawford G *et al.* **The EWAS Catalog: a database of epigenome-wide association studies [version 2; peer review: 2 approved]** Wellcome Open Research 2022, **7**:41 https://doi.org/10.12688/wellcomeopenres.17598.2

First published: 04 Feb 2022, **7**:41 https://doi.org/10.12688/wellcomeopenres.17598.1

REVISED **Amendments from Version 1**

Reviewers' comments were addressed. In summary:

- Instructions on how users can upload data to The EWAS Catalog database was added to the text.
- Information on the frequency of published data collection (monthly) was added to the text and the website (http://www.ewascatalog.org/about/).
- Information regarding the tissues used for the GEO analysis and justification for use of surrogate variables as covariates was added to the text.

None of the figures, tables, extended data, or supplementary data were changed.

Any further responses from the reviewers can be found at the end of the article

Introduction

EWAS assess associations between traits of interest and DNA methylation across the genome[1–3]. These associations may be used to gain mechanistic insights into disease and developmental processes or serve as molecular biomarkers in prediction applications[1–3]. Giving researchers easy access to the data will likely improve understanding of complex traits and may yield other translational benefits.

The EWAS Atlas has previously collated well-curated EWAS on traits in an online database and makes annotated CpG site-level results accessible via a website[4]. Other databases are available but are limited to certain diseases (e.g. MethHC[5]). Ideally, a database of EWAS results will provide summary statistics, including effect estimates, standard errors, and p-values in an easily accessible manner, so that researchers can explore associations without having to retrieve the original article. For example, allowing comparison of effect estimates between studies or a look-up of specific associations to evaluate replication. For completeness, such a database should also, where possible, provide summary statistics for all potentially true associations beyond those passing conservative significance thresholds, but publications rarely report sub-threshold lists of associations. The contents of EWAS Atlas have to-date been restricted to published associations.

We therefore aimed to improve upon current databases to 1) provide all relevant summary statistics from a range of EWAS and 2) allow easy and programmatic access to results. To this end we have produced The EWAS Catalog, a manually curated database of currently published EWAS with additional data from 387 EWAS performed in ALSPAC[6,7], and 40 EWAS performed with publicly available data from the GEO database. The process and data inclusion are summarized in Figure 1. The EWAS Catalog also enables users to upload results, which go through manual and automated checks ensuring the data meets the standards of the database, allowing collection of results not necessarily reported in publications.

Methods

Implementation

The EWAS Catalog web app was built using the Django Python package (https://djangoproject.com). The data is stored in a combination of a MySQL database and fast random access files[8], and can be queried via the website or the R package.

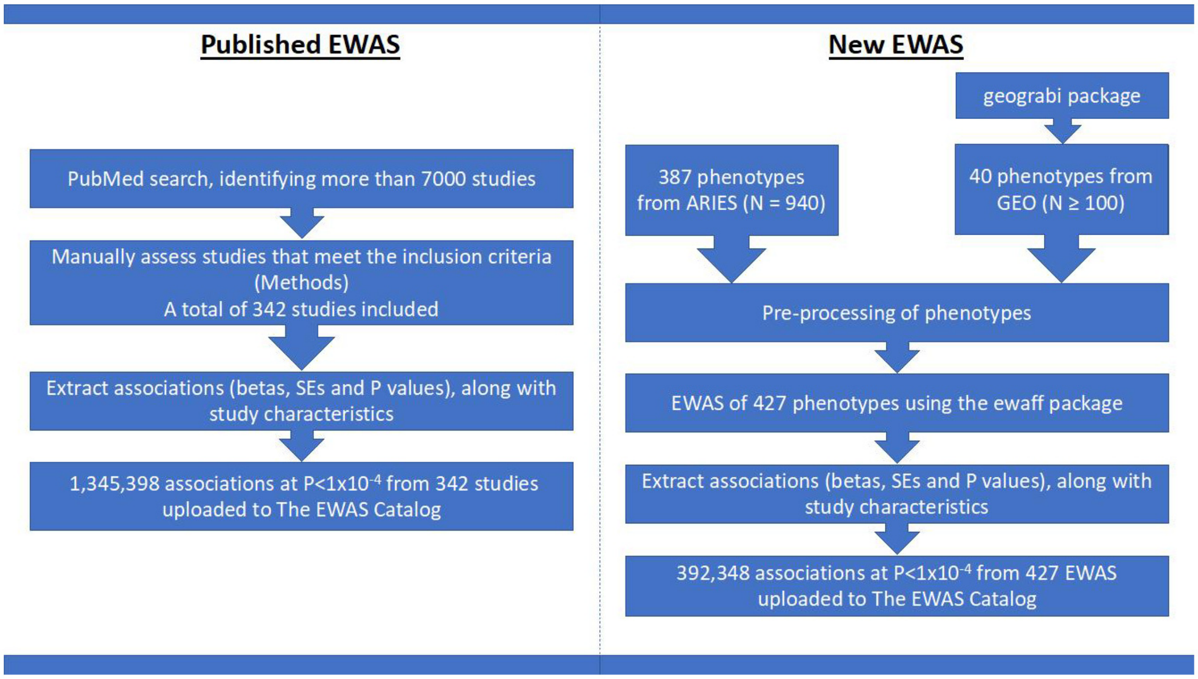

Published EWAS
PubMed search, identifying more than 7000 studies
Manually assess studies that meet the inclusion criteria (Methods)
A total of 342 studies included
Extract associations (betas, SEs and P values), along with study characteristics
1,345,398 associations at P<1x10-4 from 342 studies uploaded to The EWAS Catalog
New EWAS
geograbi package
387 phenotypes from ARIES (N = 940)
40 phenotypes from GEO (N ≥ 100)
Pre-processing of phenotypes
EWAS of 427 phenotypes using the ewaff package
Extract associations (betas, SEs and P values), along with study characteristics
392,348 associations at P<1x10-4 from 427 EWAS uploaded to The EWAS Catalog

Figure 1. Project flowchart. On the left is a brief description of how we assembled CpG-phenotype associations from published works and on the right is a brief description of the EWAS performed using individual level data.

Overview of publication data extraction

To identify publications, we perform periodic literature searches in PubMed using the terms: “epigenome-wide” OR “epigenome wide” OR “EWAS” OR “genome-wide AND methylation” OR “genome wide AND methylation”.

Our criteria for inclusion of EWAS are as follows:

1. The EWAS was performed using data from over 100 humans.
2. The analysis contains over 100,000 CpG sites
3. DNA methylation data is genome-wide (not a candidate gene study)
4. Results are not duplicated from a previous study
5. CpG-trait associations at $p<1x10^{-4}$ are reported

These criteria and the variables extracted are documented on the website. Briefly, extracted variables included: the exposure variable, the outcome variable, the covariates, tissue, sample size, age, sex, reported ancestry or ethnicity, CpG IDs, effect estimates, standard errors, p-values. To unify representation of traits, they were mapped to Experimental Factor Ontology (EFO) terms, which were manually extracted from the European Bioinformatics Institute database.

The EWAS Catalog data extraction team extracts data from newly published EWAS monthly.

EWAS study data

Avon Longitudinal Study of Parents and Children (ALSPAC). EWAS were conducted for 387 continuous and binary traits (*Extended data*[9]) using DNA methylation measured in peripheral blood of middle-aged ALSPAC mothers (N = 940). The trait data were extracted from information collected at the same sampling point blood was drawn for DNA methylation assays. Quality control steps for the traits and information on the cohort are in the Extended methods section.

GEO datasets

Full EWAS results were also estimated for studies that did not report complete summary statistics in their initial publication but where complete DNA methylation and trait of interest information were publicly available through the GEO database. We used the geograbi R package to query GEO for experiments matching inclusion criteria (described above) and extract data for EWAS re-analysis. The query was performed using the geograbi.retrieve.datasets() function on 12 October 2020 and identified 136 experiments with 32,555 samples meeting The EWAS Catalog inclusion criteria where DNA methylation and phenotype information could be successfully extracted. GEO identifies publications corresponding to all database records by PubMed ID and we accessed these for all retained GEO datasets to identify the original variable of interest. We aimed to replicate the original published analysis from the available GEO data in order to generate a full set of summary statistics to be included in The EWAS Catalog. However, of our 136 putative GEO studies, only 34 (25%), which represented 40 EWAS, contained sufficient information to replicate the original analysis. The main reason for study exclusion at this stage was for missing phenotype information. Half of the 40 EWAS, measured DNA methylation data in whole blood, and there was a range of tissues used for the other EWAS, including saliva, brain, skin, colon. This data is available as part of the downloadable meta-data from The EWAS Catalog website. Both original published results and the full re-analysed GEO results have been included in The EWAS Catalog database. A list of all 40 traits with corresponding citations is provided as *Underlying data*[10].

Details on the statistical analyses for EWAS performed specifically for The EWAS Catalog can be found in the Extended methods section. The full summary statistics for these results can be found on the following Zenodo projects: https://doi.org/10.5281/zenodo.4672645, https://doi.org/10.5281/zenodo.4672754.

As of January 7, 2022, The EWAS Catalog contained 1,737,746 associations from 2,686 EWAS.

Extended methods

Avon Longitudinal Study of Parents and Children (ALSPAC). Pregnant women residing in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992 were invited to take part in the study. The initial number of pregnancies enrolled was 14,541 (for these at least one questionnaire has been returned or a “Children in Focus” clinic had been attended by 19 July 1999). Of these initial pregnancies, there was a total of 14,676 foetuses, resulting in 14,062 live births and 13,988 children who were alive at 1 year of age. Full details of the cohort have been published previously[6,7]. The EWAS performed for The EWAS Catalog were done so using DNA methylation measured in peripheral blood of ALSPAC mothers in middle age (N = 940), generated as part of the Accessible Resource for Integrated Epigenomics Studies (ARIES) project[11].

All continuous and binary phenotypes were extracted from the same timepoint that blood was drawn for DNA methylation assays. A list of the phenotypes can be found in the *Extended data*[9].

Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Consent for biological samples has been collected in accordance with the Human Tissue Act (2004). Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time. The study website contains details of all the data that are available through a fully searchable data dictionary and variable search tool.

Preparing phenotype data from ALSPAC and GEO for EWAS

For continuous traits we defined outliers as follows:

$$Outlier < LQ + 3 * IQR \qquad Outlier > UQ + 3 * IQR,$$

where LQ = lower quartile, IQR = interquartile range, UQ = upper quartile. Any outliers were set to missing, then all phenotypes with 100 or more non-missing values were kept for further analysis. To ensure all phenotypes were approximately normally distributed, each distribution was examined and transformed as required. Log-transformations were performed on right-skewed variables. Square-roots and cube-roots were used to try and approximate normality if log-transformation did not produce an approximately normal distribution. To produce approximately normally distributed data for left-skewed variables, they were squared.

EWAS statistical analyses. For all EWAS performed specifically for the EWAS Catalog, linear regression models were fit with DNA methylation as the outcome, coded as numbers between 0 and 1, and the trait as the exposure. For EWAS using ARIES participant data, covariates included age, the top 10 ancestry principal components, and 20 surrogate variables (SVs). For EWAS using GEO data, 20 SVs were included as covariates. Other covariates were considered, but SVs only were used for two reasons: 1) to help automate the process and 2) because covariates used in the original EWAS were not included with many GEO datasets. SVs were included in our EWAS models to capture unmeasured confounded factors, especially batch effects and cell composition differences. SVs were originally developed to help identify batch effects[12] and are commonly used in EWAS to do this[13], but they've also been shown to capture cell composition differences[13,14].

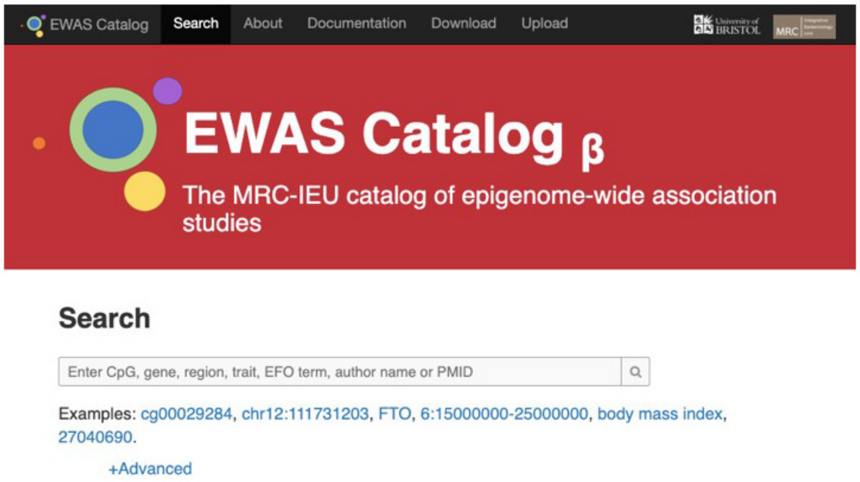
Analyses were conducted in R (Version 3.6.2). The smartsva package[15] was used to create SVs and the ewaff R package was used to conduct the EWAS; all p-values are two-sided.

Results

The database can be queried at www.ewascatalog.org. The website provides a simple user interface with a search bar to explore the database as well as documentation on the catalog contents and how to cite its use (Figure 2). Basic queries may include a CpG identifier, gene symbol, genome region, trait, author name or PubMed ID. Query submission will then lead to an intermediate 'splash page' providing options for more specific queries. For example, a query for a specific trait would lead to a 'splash page' listing that trait, related traits, and all studies of that trait. Selecting one of these leads to a list of relevant EWAS associations, including CpG ID, trait, sample size, publication, and association (effect size and p-value) (Figure 3). This information, along with further details such as reported ancestry, outcome, exposure units and tissue analyzed, are available for download as a tab-delimited text file. Alternatively, advanced queries are also supported wherein both a CpG identifier, gene symbol or genomic region are specified along with a trait, author name or PubMed ID. These queries are more specific and lead directly to a list of relevant EWAS associations.

The catalog can also be queried programmatically using the "ewascatalog" R package. Installation instructions and examples are available at its Github repository. Once installed, the database can be queried directly in R using the "ewascatalog()" function similarly to the website. By supplying the function with a CpG site, gene, genome position or trait, the function returns the same output as is downloadable from the website.

To upload data to The EWAS Catalog database, authors need to simply fill out a short form describing the study (inputting



EWAS Catalog β

The MRC-IEU catalog of epigenome-wide association studies

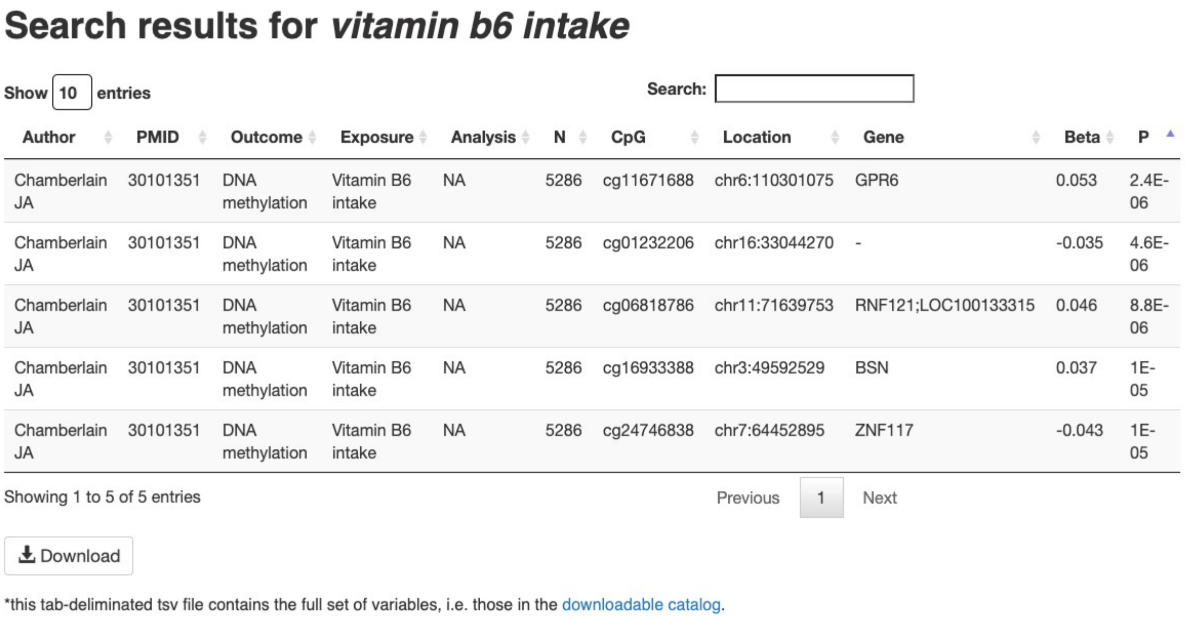
Search

Enter CpG, gene, region, trait, EFO term, author name or PMID

Examples: cg00029284, chr12:111731203, FTO, 6:15000000-25000000, body mass index, 27040690.

+Advanced

Figure 2. The EWAS Catalog home page. From here users can search the database, view documentation, and navigate to pages that allow for download of the full database and upload of user results. An example of results can be found in Figure 3.



Search results for *vitamin b6 intake*

Author	PMID	Outcome	Exposure	Analysis	N	CpG	Location	Gene	Beta	P
Chamberlain JA	30101351	DNA methylation	Vitamin B6 intake	NA	5286	cg11671688	chr6:110301075	GPR6	0.053	2.4E-06
Chamberlain JA	30101351	DNA methylation	Vitamin B6 intake	NA	5286	cg01232206	chr16:33044270	-	-0.035	4.6E-06
Chamberlain JA	30101351	DNA methylation	Vitamin B6 intake	NA	5286	cg06818786	chr11:71639753	RNF121;LOC100133315	0.046	8.8E-06
Chamberlain JA	30101351	DNA methylation	Vitamin B6 intake	NA	5286	cg16933388	chr3:49592529	BSN	0.037	1E-05
Chamberlain JA	30101351	DNA methylation	Vitamin B6 intake	NA	5286	cg24746838	chr7:64452895	ZNF117	-0.043	1E-05

Showing 1 to 5 of 5 entries

*this tab-deliminated tsv file contains the full set of variables, i.e. those in the downloadable catalog.

Figure 3. Example of results from The EWAS Catalog website. These results can be extracted by clicking the "Download" button at the bottom of the figure. This download will include extra study information, such as age, sex and reported ancestry of study participants.

the meta-data variables listed in the Overview of publication data extraction section) and then upload the EWAS summary statistics (effect estimates, standard errors, and P values). The link to this online form and the location of where to upload the summary statistics can be obtained by emailing "ewascatalog@outlook.com", and one of our team will promptly respond with all the necessary details.

Discussion/conclusions

The EWAS Catalog provides a database of summary statistics from currently published EWAS and an additional 427 currently unpublished EWAS. This database has similar aims to the EWAS Atlas but has additional data sources, provides extra useful information and a user upload option. The EWAS Catalog team will continue to collate and upload newly published EWAS and perform additional EWAS on available datasets, whilst encouraging EWAS authors to upload their own summary data. We are currently working to incorporate additional functionality to allow users to systematically compare their own EWAS findings to EWAS already in the database.

Data availability

Underlying data

The EWAS Catalog URL: http://www.ewascatalog.org

All published summary statistics at $p<1\times10^{-4}$ are available on the website. Any additional statistics or data associated with publications can be obtained by following links to the publications provided by The EWAS Catalog website. The full summary statistics from all EWAS conducted within ALSPAC, GEO and from any uploaded data can be found here: https://zenodo.org/communities/ewas-catalog. The original GEO data can be found on the GEO website (https://www.ncbi.nlm.nih.gov/geo/) using the accession IDs provided as underlying data (https://doi.org/10.5281/zenodo.5905938)[10] and The EWAS Catalog website or R package.

ALSPAC data is accessed through a system of managed open access.

The steps below highlight how to apply for access to the data included in this software tool article and all other ALSPAC data. The data presented in this article are linked to ALSPAC project number B3259, please quote this project number during your application. The ALSPAC variable codes highlighted in the dataset descriptions can be used to specify required variables.

1. Please read the ALSPAC access policy (PDF, 627kB) which describes the process of accessing the data and samples in detail, and outlines the costs associated with doing so.
2. You may also find it useful to browse our fully searchable research proposals database, which lists all research projects that have been approved since April 2011.
3. Please submit your research proposal for consideration by the ALSPAC Executive Committee. You will receive a response within 10 working days to advise you whether your proposal has been approved.

If you have any questions about accessing data, please email alspac-data@bristol.ac.uk.

The ALSPAC data management plan describes in detail the policy regarding data sharing, which is through a system of managed open access.

The project contains the following underlying data:

Zenodo: The EWAS Catalog manuscript: Underlying data https://doi.org/10.5281/zenodo.5905938[10]

Extended data

This project contains the following extended data:

- A table of the 387 traits for which EWAS were conducted using data from ARIES along with the sample sizes for each of the EWAS: The EWAS Catalog manuscript: Extended data (https://doi.org/10.5281/zenodo.5905767)[9]

Software availability

Source code available from: https://github.com/MRCIEU/ewascatalog

R package available from: https://github.com/MRCIEU/ewascatalog-r

Archived R package code as at time of publication: https://doi.org/10.5281/zenodo.5519348

License: MIT

Acknowledgements

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

This publication is the work of the authors and TB, MS and JS will serve as guarantors for the contents of this paper.

References

1. Relton CL, Smith GD: **Epigenetic Epidemiology of Common Complex Disease: Prospects for Prediction, Prevention, and Treatment.** *PLoS Med.* 2010; **7**(10): e1000356.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Mill J, Heijmans BT: **From promises to practical strategies in epigenetic epidemiology.** *Nat Rev Genet.* 2013; **14**(8): 585–594.
PubMed Abstract | Publisher Full Text
3. Rakyan VK, Down TA, Balding DJ, *et al.*: **Epigenome-wide association studies for common human diseases.** *Nat Rev Genet.* 2011; **12**(8): 529–541.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Li M, Zou D, Li Z, *et al.*: **EWAS Atlas: A curated knowledgebase of epigenome-wide association studies.** *Nucleic Acids Res.* 2019; **47**(D1): D983–D988.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Huang WY, Hsu SD, Huang HY, *et al.*: **MethHC: A database of DNA methylation and gene expression in human cancer.** *Nucleic Acids Res.* 2015; **43**(Database issue): D856–61.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Fraser A, Macdonald-Wallis C, Tilling K, *et al.*: **Cohort profile: The avon longitudinal study of parents and children: ALSPAC mothers cohort.** *Int J Epidemiol.* 2013; **42**(1): 97–110.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Boyd A, Golding J, Macleod J, *et al.*: **Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children.** *Int J Epidemiol.* 2013; **42**(1): 111–127.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Li H: **Tabix: Fast retrieval of sequence features from generic TAB-delimited files.** *Bioinformatics.* 2011; **27**(5): 718–9.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Battram T, *et al.*: **The EWAS Catalog manuscript: Extended data [Data set].** *Zenodo.* 2022.
http://www.doi.org/10.5281/zenodo.5905767
10. Battram T, *et al.*: **The EWAS Catalog manuscript: Underlying data [Data set].** *Zenodo.* 2022.
http://www.doi.org/10.5281/zenodo.5905938
11. Relton CL, Gaunt T, McArdle W, *et al.*: **Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES).** *Int J Epidemiol.* 2015; **44**(4): 1181–1190.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Leek JT, Storey JD: **Capturing heterogeneity in gene expression studies by surrogate variable analysis.** *PLoS Genet.* 2007; **3**(9): 1724–35.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Teschendorff AE, Relton CL: **Statistical and integrative system-level analysis of DNA methylation data.** *Nat Rev Genet.* 2018; **19**(3): 129–147.
PubMed Abstract | Publisher Full Text
14. Kong Y, Rastogi D, Seoighe C, *et al.*: **Insights from deconvolution of cell subtype proportions enhance the interpretation of functional genomic data.** *PLoS One.* 2019; **14**(4): e0215987.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Chen J, Behnam E, Huang J, *et al.*: **Fast and robust adjustment of cell mixtures in epigenome-wide association studies with SmartSVA.** *BMC Genomics.* 2017; **18**(1): 413.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 1

Reviewer Report 11 May 2022

https://doi.org/10.21956/wellcomeopenres.19460.r50168

✓ **Harold Snieder**

Department of Epidemiology, University Medical Center Groningen, Groningen, The Netherlands

This is a clear software tool article introducing the EWAS catalog. I have a few questions and suggestions the authors may want to consider.

1. The 5th inclusion criterion specifies that CpG trait associations at p<1x10^{-4} need to have been reported in order for the study results to be included in the EWAS catalog. What is the justification for this criterion (make it more explicit) and is it wise to apply it strictly? I'm afraid interesting studies may be missed that have not reported CpG trait associations at p<1x10^{-4}.

2. It should be made more attractive for researchers to upload full EWAS summary statistics to the EWAS catalog, so it becomes the go-to repository for scientists that like to build on earlier work through making use of available EWAS summary stats and perform meta-analyses, just like the GWAS catalog is for GWA studies. As such, please mention that it is one of the aims of the EWAS catalog to encourage scientists to upload their EWAS summary stats, make the upload procedure as user friendly as possible and provide more information on this procedure in the current article.

3. The catalog is periodically updated with new publications, but with which frequency? Every month?

4. Were all the GEO EWAS conducted in peripheral blood or also other tissues?

5. For the new EWAS analyses in ALSPAC and GEO, why was DNA methylation specified as the outcome rather than the reverse. Please indicate briefly pros and cons.

6. Are the 20 surrogate variables (SVs) expected to sufficiently capture technical covariates such as batch effects and cell type distribution? A supporting reference would help.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: (Epi)Genetic epidemiology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 26 May 2022

Thomas Battram, University of Bristol, UK

This is a clear software tool article introducing the EWAS catalog. I have a few questions and suggestions the authors may want to consider.

We thank the reviewer for their positive comments, and helpful suggestions and questions. See our responses to each one individually below.

○ *The 5th inclusion criterion specifies that CpG trait associations at p<1x10-4 need to have been reported in order for the study results to be included in the EWAS catalog. What is the justification for this criterion (make it more explicit) and is it wise to apply it strictly? I'm afraid interesting studies may be missed that have not reported CpG trait associations at p<1x10-4.*

It is true that CpG-trait associations at P>1x10^{-4} may be of interest. Ideally, we would store full-summary statistics that could be searched. Unfortunately, we don't have the capacity to store such a large dataset. Hence, when inputting newly published EWAS into our database, we contact the authors and offer the service of uploading full summary statistics to Zenodo when uploading their data to The EWAS Catalog. We have added a link to these data in the "Download" section of the website to improve their visibility. For studies that report results in their paper at P>1x10^{-4}, we do not ask for the results to be uploaded to Zenodo as the data can be accessed via the original paper. We provide PubMed IDs that make it easy to link EWAS to their publications. Further, we have found that it is rare that individuals report

EWAS associations at $P>1x10^{-4}$, thus the majority of data in The EWAS Catalog database is all that could be taken from each study. Hopefully this will change in the future and people will report full summary statistics. We plan on continuing to upload these full summary statistics to Zenodo.

The $P<1x10^{-4}$ threshold is somewhat arbitrary, but it does have some justification. It corresponds to 80% power to detect a 10% difference in DNA methylation in a sample set of size of 100 (50 per group) assuming a standard deviation of 0.1 for DNA methylation. Given that this ignores adjustment for multiple tests (typical EWAS perform hundreds of thousands of tests) and that 10% methylation differences are rare to observe in an EWAS, we consider this a very permissive threshold.

- *It should be made more attractive for researchers to upload full EWAS summary statistics to the EWAS catalog, so it becomes the go-to repository for scientists that like to build on earlier work through making use of available EWAS summary stats and perform meta-analyses, just like the GWAS catalog is for GWA studies. As such, please mention that it is one of the aims of the EWAS catalog to encourage scientists to upload their EWAS summary stats, make the upload procedure as user friendly as possible and provide more information on this procedure in the current article.*

We thank the reviewer for the great suggestion. Currently, we try to encourage users to upload EWAS summary statistics by contacting the authors of new EWAS every month and asking if they would upload. We've attempted to make the process as simple as possible by providing a short online form to complete and an upload link for summary statistics. To increase visibility of the platform and to remind others that they can upload summary statistics, we've now made a Twitter profile for The EWAS Catalog (https://twitter.com/ewascatalog). And will shortly begin regularly tweeting about catalog updates and relevant highlights from the literature. We've also added these lines to the end of the abstract:

"The EWAS Catalog data extraction team continue to update the database monthly and we encourage any EWAS authors to upload their summary statistics to our website. Details of how to upload data can be found here: http://www.ewascatalog.org/upload."

And we have added these lines to the end of the results:

"To upload data to The EWAS Catalog database, authors need to simply fill out a short form describing the study (inputting the meta-data variables listed in the Overview of publication data extraction section) and then upload the EWAS summary statistics (effect estimates, standard errors, and P values). The link to this online form and the location of where to upload the summary statistics can be obtained by emailing "ewascatalog@outlook.com", and one of our team will promptly respond with all the necessary details."

- *The catalog is periodically updated with new publications, but with which frequency? Every month?*

The catalog is updated monthly. This information has now been added to the website, (http://www.ewascatalog.org/about/). We've further incorporated a line on the front page of the website that indicates when the data was added. We also added this line to the end of the "Overview of publication data extraction" section:
"The EWAS Catalog data extraction team extracts data from newly published EWAS

monthly.”

- ○ *Were all the GEO EWAS conducted in peripheral blood or also other tissues?*

Great question. The GEO EWAS were conducted in multiple tissues. We have added the lines below the “EWAS study data” section to indicate the tissues used. This information can also be seen when downloading the data from The EWAS Catalog database.

“Half of the 40 EWAS, measured DNA methylation data in whole blood, and there was a range of tissues used for the other EWAS, including saliva, brain, skin, colon. This data is available as part of the downloadable meta-data from The EWAS Catalog website.”

- ○ *For the new EWAS analyses in ALSPAC and GEO, why was DNA methylation specified as the outcome rather than the reverse. Please indicate briefly pros and cons.*

This modelling decision was entirely practical. Due to the large number of EWAS conducted, we did not want to hypothesise a direction of effect for each trait. However, evidence from previous studies suggests effects are more likely to go from trait to DNA methylation rather than the other way around (Wahl et al. Nature, 2017; Min et al. Nature Genetics, 2021).

For the EWAS run using the data from ALSPAC, blood was drawn to measure DNA methylation at the same time as phenotypes were measured. Therefore, knowing whether a change in DNA methylation occurred before trait variation is difficult. For the GEO data, it was unclear from the majority datasets when the DNA methylation was measured.

From a statistical point of view, the set of top associations identified by an EWAS should be identical whether or not DNA methylation is the outcome. Having DNA methylation as the outcome is also convenient as it ensures that the outcome is always continuous. Of course, effect sizes will differ in magnitude and interpretation, but we consider this less important than identifying top CpG site associations.

- ○ *Are the 20 surrogate variables (SVs) expected to sufficiently capture technical covariates such as batch effects and cell type distribution? A supporting reference would help.*

Surrogate variables will capture the factors in the data that contribute most to DNA methylation variation independent of the trait of interest. Previous studies have shown that batch effects are major components of DNA methylation variation, effects that surrogate variable analysis was first designed to identify. Cell proportions have also been shown to contribute greatly to DNA methylation variability and surrogate variables have been shown to partially capture this source of variability as well. We’ve now added the following to the manuscript at the end of the “Extended methods” section:

“SVs were included in our EWAS models to capture unmeasured confounded factors, especially batch effects and cell composition differences. SVs were originally developed to help identify batch effects (Leek and Storey, Plos Genetics, 2007) and are commonly used in EWAS to do this (Teschendorff and Relton, Nature Reviews Genetics, 2018), but they’ve also been shown to capture cell composition differences (Teschendorff and Relton Nature Reviews Genetics, 2018; Kong et al. Plos One, 2019).”

Competing Interests: No competing interests were disclosed.

Reviewer Report 14 March 2022

https://doi.org/10.21956/wellcomeopenres.19460.r49012

✓ **John W. Holloway**

[1] Human Development and Health, Faculty of Medicine, University of Southampton, Southampton, UK

[2] NIHR Southampton Biomedical Research Centre, University Hospital Southampton, Southampton, UK

Battram *et al.* describe the construction of an on-line database, the “EWAS catalogue”, that provides a tool to allow researchers to query EWAS associations quickly and easily. The database is populated with data from 342 published studies along with data generated *de novo* from the ARIES epigenetics dataset of middle-aged women in ALSAPC (N=367 phenotypes) and EWAS data generated from geo datasets (N=40).

The documentation is clear and the website well presented. Importantly, an option is provided for authors to upload their own datasets, allowing for further enrichment of the database going forward.

The database presented is complementary to the EWAS atlas (Li M *et al*. 2019), especially given the inclusion of the ALSPAC and Geo data.

Have the authors considered providing an API interface to allow integration of the EWAS catalogue into other tools? What about a trait enrichment tool so researchers can check for enrichment of a set of CpGs identified (e.g. associated with an exposure) with disease associated CpGs in the catalogue?

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genetic and Epigenetic epidemiology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 26 May 2022

Thomas Battram, University of Bristol, UK

Battram et al. describe the construction of an on-line database, the "EWAS catalogue", that provides a tool to allow researchers to query EWAS associations quickly and easily. The database is populated with data from 342 published studies along with data generated de novo from the ARIES epigenetics dataset of middle-aged women in ALSAPC (N=367 phenotypes) and EWAS data generated from geo datasets (N=40).

The documentation is clear and the website well presented. Importantly, an option is provided for authors to upload their own datasets, allowing for further enrichment of the database going forward.

The database presented is complementary to the EWAS atlas (Li M et al. 2019), especially given the inclusion of the ALSPAC and Geo data.

We thank the reviewer for their positive comments and excellent suggestions. See our responses below.

- *Have the authors considered providing an API interface to allow integration of the EWAS catalogue into other tools?*

Yes, we currently have an API to access The EWAS Catalog from R. It is still experimental so we haven't advertised it yet. It can be installed from Github: https://github.com/MRCIEU/ewascatalog-r.

- *What about a trait enrichment tool so researchers can check for enrichment of a set of CpGs identified (e.g. associated with an exposure) with disease associated CpGs in the catalogue?*

This is an excellent idea. We have begun work on implementing this tool and have a version that works offline. We are preparing a manuscript describing results obtained using that tool. We hope to provide website access to the tool soon.

Competing Interests: No competing interests were disclosed.