I made this template based on thesisdown to comply with the University of Bristol regulations

Thomas Battram
MRC Integrative Epidemiology Unit
Bristol Medical School
Faculty of Health Sciences
University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Population Health Sciences in the Faculty of Health Sciences

Bristol Medical School, September 2020, Word Count:

Abstract

My abstract will go here and it will be a solid abstract. Full of the things that go in abstracts. Such as numbers, acronyms, other words, and lots of punctuation.

It will have multiple paragraphs too!

Acknowledgements

THIS IS WHERE YOU THANK PEOPLE!!!

Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Signed

Dated

Table of Contents

List of Tables

List of Figures

Preface

This template is based on (and in many places copied directly from) the Reed College LaTeX template, but hopefully it will provide a nicer interface for those that have never used TeX or LaTeX before. Using *R Markdown* will also allow you to easily keep track of your analyses in **R** chunks of code, with the resulting plots and output included as well. The hope is this *R Markdown* template gets you in the habit of doing reproducible research, which benefits you long-term as a researcher, but also will greatly help anyone that is trying to reproduce or build onto your results down the road.

Hopefully, you won't have much of a learning period to go through and you will reap the benefits of a nicely formatted thesis. The use of LaTeX in combination with *Markdown* is more consistent than the output of a word processor, much less prone to corruption or crashing, and the resulting file is smaller than a Word file. While you may have never had problems using Word in the past, your thesis is likely going to be about twice as large and complex as anything you've written before, taxing Word's capabilities. After working with *Markdown* and **R** together for a few weeks, we are confident this will be your reporting style of choice going forward.

Why use it?

R Markdown creates a simple and straightforward way to interface with the beauty of LaTeX. Packages have been written in **R** to work directly with LaTeX to produce nicely formatting tables and paragraphs. In addition to creating a user friendly interface to LaTeX, *R Markdown* also allows you to read in your data, to analyze it and to visualize it using **R** functions, and also to provide the documentation and commentary on the results of your project. Further, it allows for **R** results to be passed inline to the commentary of your results. You'll see more on this later.

Who should use it?

Anyone who needs to use data analysis, math, tables, a lot of figures, complex cross-references, or who just cares about the final appearance of their document should use *R Markdown*. Of particular use should be anyone in the sciences, but the user-friendly nature of *Markdown* and its ability to keep track of and easily include figures, automatically generate a table of contents, index, references, table of figures, etc. should make it of great benefit to nearly anyone writing a thesis project.

For additional help with bookdown Please visit the free online bookdown reference guide.

Chapter 1

Introduction

Chapter 2

Methods

Chapter 3

The EWAS Catalog: a database of epigenome-wide association studies

3.1 Abstract

Epigenome-wide association studies (EWAS) seek to understand the link between patterns of DNA methylation, the addition of a methyl group to a DNA molectule that may change how the molecule interacts with other cellular factors, at thousands or millions of sites across the genome to various traits and exposures. In recent years, the increase in availability of NDA methylation measures in population-based cohorts and case-control studies has resulted in a dramatic increase in the number of EWAS being performed and published. To make this rich source of molecular data more accessible, a manually curated database has been made containing CpG-trait associations (at p < 1x10^-4) from published EWAS, each as-

saying over 100,000 CpGs in at least 100 individuals. The database currently contains these associations from over 150 published EWAS as well as full summary statistics for over 180 million association tests of 418 EWAS in the Avon Longitudinal Study of Parents and Children (ALSPAC) and the Gene Expression Omnibus (GEO). It is accompanied by a web-based tool and R package that allow these associations to be easily queried. This database will give researchers the opportunity to quickly and easily query EWAS associations to gain insight into the molecular underpinnings of disaese as well as the impact of traits and exposures on the DNA methylome. The EWAS Catalog is available at: `http://www.ewascatalog.org`.

3.2 Introduction

Epigenome-wide association studies (EWAS) aim to assess the associations between phenotypes of interest and DNA methylation across the genome (Mill & Heijmans, 2013; Rakyan, Down, Balding, & Beck, 2011; Relton & Davey Smith, 2010). These associations may then be used for disease diagnosis or prediction (Mill & Heijmans, 2013; Rakyan et al., 2011; Relton & Davey Smith, 2010). Also, unlike genetic variants, changes in DNA methylation are responsive to the environment and so may be targeted for treatment. EWAS of smoking (Joehanes et al., 2016), body mass index (BMI) (Wahl et al., 2017) and aging (Horvath, 2013) have shown that various exposures are related to large perturbations in DNA methylation across the genome. Furthermore, a paper recently estimated that over 60% of the total proportion of BMI variation was captured by DNA methylation at about 150 CpG sites (Banos et al., 2018). In recent years, there has been a dramatic increase in the number of EWAS being performed and published due to technological advancements making it possible to measure DNA methylation at hundreds

of thousands of CpG sites cheaply and effectively. Giving researchers easy access to EWAS outputs will help them gain insight into the molecular underpinnings of disease as well as the impact of traits and exposures on the DNA methylome. Furthermore, current collections of summary statistics have already proven useful to various fields, for example the GWAS Catalog (Buniello et al., 2019) has been cited over 2000 times in papers contributing to new methods and exploring the genetic architecture of a plethora of traits.

To our knowledge there is currently only one database that has collated well-curated EWAS on traits (not just diseases) in an online database accessible to researchers, EWAS Atlas (Li et al., 2019). Other databases are available but are limited to certain diseases (e.g. MethHC (Huang et al., 2015)). EWAS Atlas provides a simple-to-use website with annotated CpG sites and information on traits. Ideally a database of EWAS results will provide summary statistics, including betas, standard errors and p-values where provided from publications, in an easily accessible manner, this enables researchers to explore various aspects of the published data without having to retrieve the published article. For example, they might compare effect estimates between studies in the database or check to see if their results are replicated in another published study. To our knowledge the EWAS atlas platform does not enable users to download effect estimates and standard errors. Further, there is currently only published data on the platform, not full summary statistics from EWAS.

We aimed to improve upon current databases to 1) allow easy and programmatic access to summary statistics for downstream analyses by researchers and 2) provide full summary statistics from a range of EWAS conducted in multiple cohorts. To this end we have produced the EWAS catalog, a manually curated database of currently published EWAS, 378 EWAS performed in the Avon Longitudinal Study

of Parents and Children (ALSPAC) (Boyd et al., 2013; Fraser et al., 2013) and 40 EWAS performed from data from the Gene Expression Omnibus (GEO) database. The process and data inclusion are summarised in Supplementary Figure 1.

3.3 Methods

3.3.1 Implementation

The EWAS Catalog web app was built using the Django Python package (`https://djangoproject.com`). The data is stored in a combination of MySQL databases and fast random access files (Li, 2011) and can be queried via the web app or the R package (www.github.com/jrs95/ewascatalog).

3.3.2 Overview of publication data extraction

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

Our criteria for inclusion of a study into the EWAS catalog are as follows: 1. The EWAS performed must contain over 100 humans 2. The analysis must contain over 100,000 CpG sites 3. The DNA methylation data must be genome-wide 4. The study must include previously unpublished EWAS summary statistics

We extracted CpG-phenotype associations from studies at P < 1x10-4. All these criteria along with the variables extracted are documented on the website (www.ewascatalog.org/documentation). Briefly, the variables extracted included: the trait, exposure, outcome, covariates, tissue, sample size, age, sex, ancestry, CpGs, betas, standard errors, P values. Experimental factor ontology (EFO) terms were mapped to traits to unify representation of these traits. These EFO terms

were manually entered after looking up the trait in the European Bioinformatics Institute database (www.ebi.ac.uk/efo).

Based on these criteria, from 3rd July 2019, the EWAS catalog contained 540,699 associations from 159 studies.

3.3.3 Overview of GEO data extraction

To recruit additional datasets suitable for new EWAS analysis, we used the geograbi R package (`https://github.com/yousefi138/geograbi`) to both query GEO for experiments matching the EWAS Catalog inclusion criteria (described above) and extract relevant DNA methylation and phenotype information. The query was performed on 20 March 2019 and identified 148 such experiments with 32,845 samples where DNA methylation and phenotype information could be successfully extracted. From these, we aimed to repeat the analyses performed in the publications linked by PubMed IDs to each GEO record. Thus, we looked up the corresponding full texts for each dataset and identified the main variables of interest. Of our 148 putative GEO studies, only 34 (23%) contained sufficient information to replicate the original analysis.

3.3.4 EWAS methods

Avon Longitudinal Study of Parents and Children (ALSPAC) EWAS were conducted for 378 continuous and binary traits in peripheral blood DNA methylation of ALSPAC mothers in middle age (N = 940), generated as part of the Accessible Resource for Integrated Epigenomics Studies (ARIES) project (Relton et al., 2015). The traits were extracted from the same time that blood was drawn for DNA methylation assays. Quality control steps for the phenotypes along with information on the cohort can be found in the Supplementary Material. For all traits, linear

regression models were fitted with DNA methylation, coded as numbers between 0 and 1, as the outcome and the phenotype as the exposure. Covariates included age, the top 10 ancestry principal components, and 20 surrogate variables.

3.3.5 GEO datasets

EWAS were performed using 30 datasets, containing 36 traits were extracted from GEO using the geograbi R package (`https://github.com/yousefi138/geograbi`). Information about the quality control of the data can be found in the Supplementary Material and a list of all the traits with corresponding citations is provided in Supplementary Table 1. For all traits, linear regression models were fitted with DNA methylation as the outcome and the phenotype as the exposure. Twenty surrogate variables were included as covariates. Other covariates were considered, but surrogate variables 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 of the GEO datasets. Statistical analyses were conducted in R (Version 3.3.3). The smartsva package (Chen et al., 2017) was used to create surrogate variables and the ewaff R package (`https://github.com/perishky/ewaff`) was used to conduct the EWAS, all p-values are two-sided.

3.3.6 Database interface and use

There are two ways to access this large, curated database: through the main website www.ewascatalog.org or by using the R package "ewascatalog". The website provides a simple user interface, which resembles that of the GWAS catalog (Buniello et al., 2019), whereby there is a single search bar to explore the database and links to tabs that contain documentation on the contents and how to cite its use (Figure 1). Users may enter a CpG, gene, genome position or trait into the search bar and

it will rapidly return detail for relevant EWAS associations, including CpG, trait, sample size, publication and association (effect and P value) (Figure 1). This information along with additional information such as ancestry, outcome, exposure units, and tissue analysed are available for download as a tab-separated value (tsv) text file. Unlike other EWAS databases, we provide the option of downloading summary results for both the user's search and for the entire database.

Figure 1. Using the EWAS catalog. At the top of the figures is the home page URL, ewascatalog.org. Below that are examples of three types of searches possible: 1. CpG sites, 2. genes and 3. traits. Finally, the results are displayed after searching the catalog for "Depression". Circled in red is the download button, this button enables the user to download the results of their search as a tab-separated value file. This file will contain the information shown on the website as well as additional analysis information.

The R package, along with installation instructions and examples are available at `https://github.com/jrs95/ewascatalog`. Once installed, the database can be queried directly in R using the "ewascatalog()" function similar to the website: simply supply the function with a CpG site, gene, genome position or trait and the function returns the same output as is downloadable from the website.

3.4 Discussion and future developments

The EWAS catalog provides an easily accessible database of summary statistics from currently published EWAS along with full summary statistics from 418 EWAS from the ALSPAC cohort and from the GEO database. This database has a similar aim to the EWAS Atlas but additionally provides full summary statistics

when available and extra information pertinent to using the data for further research. The EWAS catalog team will continue to collate and upload newly published EWAS and further increase the number of full summary statistics on the website by performing additional EWAS on available datasets and by inviting EWAS authors to provide full summary statistics. We are currently working additional functionality to allow users to easily and systematically compare their EWAS findings to EWAS in the database. We believe that with this full summary data, we can make greater strides into discovering the epigenetic architecture of traits.

Chapter 4

Properties of EWAS

Here is a reference to Caroline's paper: (Relton & Davey Smith, 2010)

Chapter 5

m2

Chapter 6

EWAS-GWAS comparison

Chapter 7

DNAm-lung cancer MR

Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Appendix A

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

In the main Rmd file

In Chapter ??:

Appendix B

The Second Appendix, for Fun

References

Banos, D. T., McCartney, D. L., Battram, T., Hemani, G., Walker, R. M., Morris, S. W., ... Robinson, M. R. (2018). Bayesian reassessment of the epigenetic architecture of complex traits. *bioRxiv*, 450288. http://doi.org/10.1101/450288

Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., ... Smith, G. D. (2013). Cohort profile: The 'Children of the 90s'-The index offspring of the avon longitudinal study of parents and children. *International Journal of Epidemiology*, *42*(1), 111–127. http://doi.org/10.1093/ije/dys064

Buniello, A., Macarthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., ... Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gky1120

Chen, J., Behnam, E., Huang, J., Moffatt, M. F., Schaid, D. J., Liang, L., & Lin, X. (2017). Fast and robust adjustment of cell mixtures in epigenome-wide association studies with SmartSVA. *BMC Genomics*. http://doi.or

g/10.1186/s12864-017-3808-1

Fraser, A., Macdonald-wallis, C., Tilling, K., Boyd, A., Golding, J., Davey smith, G., ... Lawlor, D. A. (2013). Cohort profile: The avon longitudinal study of parents and children: ALSPAC mothers cohort. *International Journal of Epidemiology*. http://doi.org/10.1093/ije/dys066

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology*, *14*(10), R115. http://doi.org/10.1186/gb-2013-14-10-r115

Huang, W. Y., Hsu, S. D., Huang, H. Y., Sun, Y. M., Chou, C. H., Weng, S. L., & Huang, H. D. (2015). MethHC: A database of DNA methylation and gene expression in human cancer. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gku1151

Joehanes, R., Just, A. C., Marioni, R. E., Pilling, L. C., Reynolds, L. M., Mandaviya, P. R., ... London, S. J. (2016). Epigenetic Signatures of Cigarette Smoking. *Circulation: Cardiovascular Genetics*, *9*(5), 436–447. http://doi.org/10.1161/CIRCGENETICS.116.001506

Li, H. (2011). Tabix: Fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*. http://doi.org/10.1093/bioinformatics/btq671

Li, M., Zou, D., Li, Z., Gao, R., Sang, J., Zhang, Y., ... Zhang, Z. (2019). EWAS Atlas: A curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gky1027

Mill, J., & Heijmans, B. T. (2013). From promises to practical strategies in epigenetic epidemiology. *Nature Reviews Genetics*, *14*(8), 585–594. http://doi.org/10.1038/nrg3405

Rakyan, V. K., Down, T. A., Balding, D. J., & Beck, S. (2011). Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics*, *12*(8), 529–541. http://doi.org/10.1038/nrg3000

Relton, C. L., & Davey Smith, G. (2010). Epigenetic Epidemiology of Common Complex Disease: Prospects for Prediction, Prevention, and Treatment. *PLoS Medicine*, *7*(10), e1000356. http://doi.org/10.1371/journal.pmed.1000356

Relton, C. L., Gaunt, T., McArdle, W., Ho, K., Duggirala, A., Shihab, H., ... Davey Smith, G. (2015). Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *International Journal of Epidemiology*, *44*(4), 1181–1190. Retrieved from http://dx.doi.org/10.1093/ije/dyv072

Wahl, S., Drong, A., Lehne, B., Loh, M., Scott, W. R., Kunze, S., ... Chambers, J. C. (2017). Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature*, *541*(7635), 81–86. http://doi.org/10.1038/nature20784

Abbreviations

ACR - acronym

aACR - another acronym