SOFTWARE TOOL ARTICLE

REVISED MethylDetectR: a software for methylation-based health profiling [version 2; peer review: 2 approved]

Robert F. Hillary, Riccardo E. Marioni

Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, Midlothian, EH4 2XU, UK

v2 **First published:** 07 Dec 2020, **5**:283
https://doi.org/10.12688/wellcomeopenres.16458.1
Latest published: 13 Apr 2021, **5**:283
https://doi.org/10.12688/wellcomeopenres.16458.2

Abstract

DNA methylation is an important biological process that involves the reversible addition of chemical tags called methyl groups to DNA and affects whether genes are active or inactive. Individual methylation profiles are determined by both genetic and environmental influences. Inter-individual variation in DNA methylation profiles can be exploited to estimate or predict a wide variety of human characteristics and disease risk profiles. Indeed, a number of methylation-based predictors of human traits have been developed and linked to important health outcomes. However, there is an unmet need to communicate the applicability and limitations of state-of-the-art methylation-based predictors to the wider community. To address this need, we have created a secure, web-based interactive platform called 'MethylDetectR' which automates the calculation of estimated values or scores for a variety of human traits using blood methylation data. These traits include age, lifestyle traits and high-density lipoprotein cholesterol. Methylation-based predictors often return scores on arbitrary scales. To provide meaning to these scores, users can interactively view how estimated trait scores for a given individual compare against other individuals in the sample. Users can optionally upload binary phenotypes and investigate how estimated traits vary according to case vs. control status for these phenotypes. Users can also view how different methylation-based predictors correlate with one another, and with phenotypic values for corresponding traits in a large reference sample (n = 4,450; Generation Scotland). The 'MethylDetectR' platform allows for the fast and secure calculation of DNA methylation-derived estimates for several human traits. This platform also helps to show the correlations between methylation-based scores and corresponding traits at the level of a sample, report estimated health profiles at an individual level, demonstrate how scores relate to important binary outcomes of interest and highlight the current limitations of molecular health predictors.

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 2 (revision) 13 Apr 2021	✓ report	✓ report
version 1 07 Dec 2020	? report	? report

1. **Gemma Sharp**, University of Bristol, Bristol, UK

2. **Charlotte Cecil**, Erasmus MC, Rotterdam, The Netherlands
Erasmus MC, Rotterdam, The Netherlands
Leiden University Medical Center, Leiden, The Netherlands
King's College London, London, UK

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

Keywords
epigenetics, DNA methylation, epidemiology, translation, Shiny, prediction

This article is included in the Generation Scotland gateway.

Corresponding author: Riccardo E. Marioni (riccardo.marioni@ed.ac.uk)

Author roles: Hillary RF: Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Marioni RE**: Conceptualization, Methodology, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: R.E.M has received payment from Illumina for a presentation.

Grant information: This research was supported by funding from the Wellcome 4-year PhD in Translational Neuroscience–training the next generation of basic neuroscientists to embrace clinical research [grant: 108890/Z/15/Z awarded to R.F.H]. R.E.M. is supported by an Alzheimer's Research UK major project grant [grant: ARUK-PG2017B—10]. LBC1936 methylation typing was supported by Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award), Age UK, The Wellcome Trust Institutional Strategic Support Fund, The University of Edinburgh, and The University of Queensland. Lothian Birth Cohort 1921 and 1936 proteomic analyses were supported by a National Institutes of Health (NIH) research grant (R01AG054628). Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates (CZD/16/6) and the Scottish Funding Council (HR03006). DNA methylation profiling of the GS samples was carried out by the Genetics Core Laboratory at the Wellcome Trust Clinical Research Facility, Edinburgh, Scotland and was funded by the Wellcome Trust (Wellcome Trust Strategic Award "STratifying Resilience and Depression Longitudinally" ([STRADL; Reference 104036/Z/14/Z]). Proteomic analyses in STRADL were supported by Dementias Platform UK (DPUK). DPUK funded this work through core grant support from the Medical Research Council [MR/L023784/2].
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: Hillary RF and Marioni RE. **MethylDetectR: a software for methylation-based health profiling [version 2; peer review: 2 approved]** Wellcome Open Research 2021, **5**:283 https://doi.org/10.12688/wellcomeopenres.16458.2

First published: 07 Dec 2020, **5**:283 https://doi.org/10.12688/wellcomeopenres.16458.1

REVISED **Amendments from Version 1**

First, we have updated the three applications associated with the 'MethylDetectR' platform to include 'info' buttons which, when pressed, show information on the presented data, limitations of methylation-based predictors available in 'MethylDetectR' and links to all other elements of the platform to allow for quick and convenient navigation across the platform.

Second, we have removed methylation-based predictors of 27 blood protein levels which were included in the previous version of 'MethylDetectR'. The pipeline for generating these predictors has been refined. We will include the refined predictors for a larger set of proteins once they become available and are published. The predictors for chronological age and six lifestyle and biochemical traits are still available. Indeed, users can use our platform for the quick and convenient generation of methylation-based scores for these traits and interactively view how scores compare across individuals in their dataset.

Third, the discussion on the limitations of methylation-based predictors available in 'MethylDetectR' has been refined and expanded. The aim of this amendment is to emphasise that methylation-based scores cannot make consistently accurate predictions at an individual level, instead, they work well at a population level. This limits their clinical utility, however, they will improve through the employment of larger-scale studies and more refined prediction methods.

Fourth, we have included information on 'Version Control' detailing how and when we will update the 'MethylDetectR' platform. We will update the platform every three months in an effort to include new methylation-based predictors of human traits as they are generated by our group and others. Updates will be managed by the authors Robert F. Hillary and Riccardo E. Marioni. Researchers are invited to contact the authors to discuss the inclusion of new methylation-based predictors in the 'MethylDetectR' platform.

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

Introduction

DNA methylation (DNAm) is an epigenetic mechanism in which methyl groups are added to the genome sequence. Inter-individual variability in DNAm profiles results from differences in both underlying genetics and environmental influences[1]. Factors such as diet, stress and smoking behaviours may influence the process of methylation. Typically, methyl groups are added to cytosine residues in the context of a cytosine-guanine dinucleotide (CpG site)[2]. The addition of these chemical tags can alter whether, and to what extent, a gene is active. In contrast to genetics, these molecular modifications are dynamic, tissue-specific and reversible[3]. Further, methylation at many CpG sites is tissue-specific though some show strong concordance across multiple tissues[4]. In addition, CpG modifications induced through environmental factors, such as smoking, may be reversible or show persistent alterations[5].

Biological data may be harnessed to estimate or predict a variety of human characteristics and disease risk profiles. There is a growing body of evidence demonstrating the effective creation and application of DNAm-based predictors of human traits and health[6–19]. Additionally, methylation-based predictors of traits such as smoking status may provide more accurate measurements than self-reported information, thereby allowing for improved disease prediction and risk stratification[20]. Blood DNAm data is often used as it is minimally-invasive to collect and it provides a good index of the overall health status of the body[21].

Increased training sample sizes and refinements in statistical and machine learning methodologies have improved the accuracy of DNAm-based predictors[22,23]. Furthermore, there has been an increase in the commercialisation and scalability of DNAm assays for direct-to-consumer use or for use in clinical, research or industrial settings[24]. A major goal of using these predictors is to aid in prediction strategies and provide better clinical outcomes for individuals. Therefore, translational platforms for methylation-based health profiling are warranted in order to communicate the applications and limitations of DNAm-based predictors of human traits.

To address this need, we have created a web-based platform called 'MethylDetectR' that allows for an interactive demonstration of state-of-the-art DNAm-based predictors. A demo version of the app which does not require the upload of data is available at: https://shiny.igmm.ed.ac.uk/MethylDetectR_Demo/. The DNAm-based predictors in this platform include a highly accurate predictor of chronological age trained in > 13,000 individuals across 14 cohorts with a root mean squared error of 2.04 years in the original publication[22]. We also include six DNAm-based predictors of lifestyle and biochemical traits: alcohol consumption per week, body fat percentage, body mass index, high-density lipoprotein (HDL) cholesterol, smoking status and waist-to-hip ratio[25]. These predictors were generated in 5,087 individuals who are members of the Generation Scotland: Scottish Family Health Study (GS) which represents one of the largest DNAm resources in the world.

Briefly, the 'MethylDetectR' platform consists of two application named. The first application, named 'MethylDetectR – Calculate Your Scores', allows users to securely upload Illumina 450k or EPIC DNAm array data and obtain blood-based methylation predicted scores (or values) for the aforementioned traits. No data are stored by the application. Furthermore, predicted scores are often returned on arbitrary scales. The use of this application is optional as users may instead use R scripts which we have made publicly available if they so wish or if their DNAm files are too large for upload to the online application (>3 GB) (https://doi.org/10.5281/zenodo.4646300). However, users can access a file called 'Truncate_to_these_CpGs.csv' to subset the list of CpG sites in their DNAm files to those required by the 'MethylDetectR – Calculate Your Scores' application. This should reduce file size and upload time. The second and main application named 'MethylDetectR' allows users to compare DNAm-derived scores for any individual in their input dataset against other individuals in the input dataset. Percentile ranks for individuals in the input dataset may be downloaded. Users can also upload an optional file containing binary phenotypes whereby individuals are coded as '0' for control status and '1' for case status. This information allows users to view how distributions of the DNAm-derived traits vary by cases and controls. Users can also

view how the various DNAm-based predictors correlate with one another in the input dataset and in a separate reference sample. This reference sample comprises 4,450 individuals who are members of the GS study. These individuals are unrelated to each other and distinct from those in the original training samples in which the predictors for age, HDL cholesterol and lifestyle traits were developed. They are also unrelated to those included in the original training sample. Furthermore, information is provided on how well the predicted scores correlate with phenotypic values for corresponding traits that are available in GS. Lastly, the user can subset the input sample by sex, age range or case vs. control status determined by the case-control variables uploaded by the user. Further, the user can subset the GS reference sample by age and sex.

This platform communicates important information relating to the generation and applicability of DNAm-based predictors of human traits and health. The 'MethylDetectR' platform also represents a research tool for fast and automatic generation of DNAm-derived estimates for human traits. This platform can show that DNAm-based scores for traits may correlate well with measured values for a given trait at the level of the cohort. For example, a predictor for epigenetic age correlates strongly with true age with a root mean squared error of 2.04 years[22]. However, this platform also helps to show that predictors may report inaccurate values at an individual level. For instance, although the age predictor correlates well with true age at the level of the cohort, an individual's predicted age may differ from their true age by a number of years or decades. The optional incorporation of binary phenotype data allows users to view how well established or putative risk factors, as estimated by DNAm data, are stratified according to cases and controls for a given trait of interest. Together, the functionalities of 'MethylDetectR' begin to address the translational gap in the development and implementation of molecular-based health predictors by highlighting their performance and limitations in advance of their potential utility in diagnostic and stratification paradigms.

Methods

Implementation

Data protection and privacy. No data are stored in 'MethylDetectR' and are deleted upon closing the applications. Applications are also timed out after three minutes of inactivity and are hosted on patched and secure servers within the Institute of Genetics and Cancer, University of Edinburgh. This research and translational tool complies with GDPR guidelines and has been designed to ensure the highest level of data security and privacy. The 'MethylDetectR' applications and information on their usage are also available at the following website: https://www.ed.ac.uk/centre-genomic-medicine/research-groups/marioni-group/methyldetectr. Information relating to participant consent is also available at this website. Given that no data are stored, this information pertains to general risk surrounding the upload of biological data to online software and the measures taken to mitigate the risk of motivated intruders gaining access to such data.

The 'MethylDetectR' platform. The 'MethylDetectR' platform consists of two applications. The first application is called 'MethylDetectR - Calculate Your Scores'. Users may upload DNAm data as an R object (.rds file) and obtain estimated values or scores for a variety of traits across individuals in their input dataset (https://shiny.igmm.ed.ac.uk/Calculate_Your_Scores/). The upload limit is 3 gigabytes; however, files greater than 500 megabytes may take a considerable amount of time to upload. Users can make these upload files smaller by subsetting to CpG sites used in 'MethylDetectR – Calculate Your Scores'. These CpG sites are available in the 'Truncate_to_these_CpGs.csv' file in Zenodo (Zenodo link). An optional 'SexAgeInfo' file may also be uploaded in order to include sex and age information in the output file. This should be a .csv file and have three columns: one column for the IDs of individuals in the methylation file ('ID' column), one column should list the sex of these individuals written as 'Male' or 'Female' or 'NA' ('Sex' column) and one column should report the actual or chronological age of individuals ('Age' column). This functionality is important given that users can subset the input dataset and GS cohort by sex in the main 'MethylDetectR' application. Furthermore, if true age is included, then the application will use this information to subset the sample according to the age slider function on the sidebar panel. If this information is not uploaded, then epigenetic or predicted age will be used to subset the data by age range. In the case where some individuals have true age available and others have missing data, true age will be used for those who have such data and epigenetic age will be used for those without age data in order to the subset the sample. It is strongly recommended that anonymised or pseudonymised IDs are used where possible. For the user's own convenience in preparing the methylation object, it is recommended that individuals are included as columns and CpG sites as rows. However, this version or a transposed version are accepted and automatically processed by the software. The following features also aid with automation in generating DNAm-based scores for traits:

- Beta values or M values are accepted with the latter converted to beta values by the software.
- Missing methylation values are accepted and mean imputed across input individuals by the software.
- CpG sites that are necessary for the estimation of a trait but are missing in the uploaded dataset are allowed. In this case, each individual in the input dataset receives the mean beta value for a given missing CpG site derived from GS DNAm data. In effect, this gives every individual in the uploaded dataset a constant that brings their score closer to that of the reference sample. In this way, all CpG sites are used for any sample uploaded.

Predicted values or scores for the aforementioned traits can be downloaded as a .csv file which may be uploaded to the main 'MethylDetectR' application. Alternatively, an R script is provided to generate these DNAm-based scores if the user does not wish to upload DNAm data or if the DNAm file is too large for upload (https://doi.org/10.5281/zenodo.4646300[26]).

The main 'MethylDetectR' application can be described in four modules or panels (https://shiny.igmm.ed.ac.uk/MethylDetectR/). In the first panel, users can view DNAm-based scores corresponding to various traits for any individual in their input dataset. The user can interactively compare scores for any individual to the remainder of the input dataset. A .csv file may be downloaded which shows percentile ranks for each individual in the uploaded dataset. Percentile ranks are reported for each estimated trait with the exception of the age predictor which is reported in years. If the user uploads an optional binary phenotype file, they can also view the distributions of the DNAm-based scores according to case vs. control status for a given trait of interest. This file should contain IDs of individuals in the methylation file ('ID' column) as the first column and the remaining columns may contain any names or traits of interest with individuals coded as '0' for controls and '1' for cases. In the second panel, a plot shows the percentile ranks for a given individual for selected traits. Users can also choose to view the spread of percentile ranks for cases versus controls. Here, the median percentile for cases, along with the first and third quartiles (interquartile range), are plotted for each selected DNAm-based estimated trait. In the third panel, the user can view how different DNAm-based predictors correlate with one another in both the input and GS datasets. In the fourth panel, users can view how DNAm-based scores for age, lifestyle and biochemical traits correlate with phenotypic values for these traits in GS participants (n = 4,450). In each panel, users can subset the samples by age range and sex. In panels 1-3, options to subset the input dataset by case vs. control status are present.

Development of a DNAm-based predictor. Readers are referred to a review on the development of DNAm-based scores and the challenges surrounding their generation[27]. To develop 'omics'-based predictors, such as DNAm-based predictors, statistical or machine learning methodologies are commonly applied. In the case of DNAm, the process begins with the quantification of DNAm across individuals using a tissue or cell-type of interest. Many studies focus on blood as it integrates information from various tissues around the body and represents an inexpensive and minimally-invasive approach to gather molecular data. Many cohort studies also have DNA from historic blood samples stored and available to analyse. The collection of saliva and buccal samples is becoming increasingly popular as a relatively low cost and non-invasive method for cohort studies interested in epigenetic epidemiology. The number of CpG sites which are measured depends on the array used but typically includes up to around 800,000 unique sites. Following quantification and quality control, a researcher may wish to study the association between DNAm and a trait of interest, such as smoking status. The average methylation level at a given CpG site across individuals will be correlated with the trait of interest. Methylation levels may be reported between 0-100% for convenience, and a level of 50% means that 50% of cells or DNA molecules within an individual's sample show methylation at that CpG site. One approach is to correlate each CpG site, in turn, with the trait of interest thereby considering each CpG site in isolation. This is approach is referred to as an epigenome-wide or methylome-wide association study (EWAS or MWAS). Alternatively, methods such as penalised regression can be used to model all CpG sites simultaneously producing parsimonious models that account for correlated features/sites (e.g., least absolute shrinkage and selection operator or LASSO regression) or models that apply small weights to all features/sites (e.g., ridge regression). Elastic net regression is a commonly used intermediate of these two approaches. Correlations among CpGs may arise from sites which lie near each other in a genomic region or from a shared environmental influence; for instance, inhalation of cigarette smoke may affect many CpG sites across different chromosomes. A subset of CpG sites may show a strong relationship with the trait of interest and therefore be informative for predicting the trait in other individuals. The strength of the correlation, or association, is represented by an effect size and provides a weighting for that CpG site's importance in predicting the trait. In a separate or test group of individuals, predicted values or scores for the trait can be obtained by multiplying DNAm levels at each informative CpG site by their weight derived from statistical analyses. The sum of these products provides a predicted or estimated value for the trait. A statistical transformation can be applied to return DNAm-based scores on the original scale for the trait, such as pack years for smoking.Alternatively, comparison to other individuals may provide meaning to the DNAm-based scores. In any case, predicted values or scores in the test sample may be correlated with true values for a given trait to provide an index of predictive power.

DNAm-based predictors in 'MethylDetectR'. In 'MethylDetectR', we include DNAm-based predictors of chronological age, and six lifestyle and biochemical traits.

Age predictor

The age predictor was developed by Zhang *et al.* using elastic net regression and best linear unbiased prediction (BLUP)[22,28,29]. The two sets of predictors were both built on the same set of 13,566 training samples spread across 14 cohorts. The age predictor was generated using data from individuals with an age range of 2 to 104 years. The elastic net method selected 514 CpG sites as informative for predicting chronological age whereas the BLUP predictor used all CpG sites (319,607 probes). In 'MethylDetectR', we apply the elastic net predictor owing to the faster computation and superior performance of this age predictor when compared to the BLUP predictor. The elastic net predictor correlates 0.98 with chronological age (root mean squared error = 2.6 years in GS (n = 4,450) and 2.04 in original publication[22]). The age predictor is returned in values of years.

Lifestyle and biochemical traits

Previously, we generated ten predictors of lifestyle and biochemical traits in 5,087 individuals within the GS study using LASSO penalised regression[25,30]. Ten-fold cross-validation was applied and the mixing parameter (alpha) was set to 1. The lambda value corresponding to the minimum mean cross-validation error was selected and applied to generate the optimal models[30]. In a test sample consisting of 875 individuals in the Lothian Birth Cohort 1936 study, DNAm-based predictors

for four of the traits explained greater than 10% of phenotypic variance in their respective trait. These four traits were alcohol consumption, body mass index, high-density lipoprotein cholesterol and smoking behaviour. There were no phenotypic data available for body fat percentage and waist-to-hip ratio in the Lothian Birth Cohort 1936 study. However, these traits were highly correlated with body mass index in GS (correlation coefficients of 0.6 and 0.4, respectively). Therefore, body fat percentage and waist-to-hip ratio were brought forward to the 'MethylDetectR' platform in addition to the four DNAm-based predictors of lifestyle and biochemical traits which demonstrated test R^2 statistics of greater than 10% in the Lothian Birth Cohort 1936 sample. Four other traits demonstrated test R^2 statistics of less than 10%: educational attainment (2.5%), low-density lipoprotein and remnant cholesterol (0.6%), total cholesterol (2.7%) and total-to-high-density lipoprotein cholesterol ratio (4.5%). These four traits were not brought forward to the 'MethylDetectR' platform[25].

In the training sample, alcohol intake was assessed in units per week and was only considered in those who reported that their intake was representative of a normal week. A natural log(units + 1) transformation was applied to reduce skewness. For body mass index, extreme values defined as less than 17 kg/m^2 or greater than 50 kg/m^2 were removed and a natural log transformation was applied. Smoking behaviour was assessed using pack years which is calculated by multiplying the number of packs smoked per day by the number of years the participant has smoked. Current and never smokers were included; ex-smokers were removed owing to complications in adjusting for time since cessation when calculating pack years. To reduce skewness, a natural log(pack years + 1) transformation was applied. In generating the predictors, phenotypes were pre-corrected to remove the influence of age, sex and ancestry using ten genetic principal components. Phenotypic data used to train the predictors were not corrected for cell-type heterogeneity. Further quality control details are available in the original publication[25]. DNAm values at CpG sites were the independent variables (n = 392,843 CpG sites). CpG sites were filtered to include loci present on both the Illumina EPIC and 450k arrays.

Version Control. We will update 'MethylDetectR' every three months to include new DNAm-based predictors of human traits as they are generated by our own group and others. Updates will be managed by Robert F. Hillary or Riccardo E. Marioni. If researchers wish to have their predictors considered for inclusion in 'MethylDetectR', please use the corresponding author email address in this manuscript or the contact details available at the 'MethylDetectR' website (https://www.ed.ac.uk/centre-genomic-medicine/research-groups/marioni-group/methyldetectr). The current and historical versions of 'MethylDetectR' are available in the Zenodo repository, updated versions will also be made available in this repository (https://doi.org/10.5281/zenodo.4646300).

Operation

Software requirements. Both applications are hosted on a secure, patched server hosted at the University of Edinburgh. The applications are developed using Shiny (version 1.4) in R[31]. The version of R used at the time of 'MethylDetectR' development was 3.5.0[32]. For 'MethylDetectR – Calculate Your Scores', the following R packages were utilised: shinyWidgets (version 0.5.4)[33], shinythemes (version 1.1.2)[34], data.table (version 1.12)[35], shinyalert (version 1)[36]. For the main 'MethylDetectR' application, shinyWidgets (version 0.5.4)[33], shinythemes (version 1.1.2)[34], data.table (version 1.12)[35], shinyalert (version 1)[36] were also used, in addition to ggplot2 (version 3.0)[37], dplyr (version 0.7)[38], forcats (version 0.4.0)[39], wesanderson (version 0.3.6)[40], shinycssloaders (version 0.2)[41], magick (version 2.5.0)[42], corrplot (version 0.84)[43], ggcorrplot (version 0.1)[44] and cowplot (version 0.9.4)[45]. Scripts for implementing both applications are available at: https://doi.org/10.5281/zenodo.4646300[26]. The script has been designed to ensure automatic installation of missing CRAN packages which are necessary for the operation of 'MethylDetectR'.

Overview of workflow. The main components of the platform are outlined in the Implementation section and the associated workflow is graphically depicted in Figure 1.

Use cases

MethylDetectR – Calculate Your Scores

The user can upload a DNAm file as an R object (.rds file) to the 'MethylDetectR – Calculate Your Scores' application. Beta values or M values may be used. If M values are detected, these are converted to beta values. It is recommended that individuals are included as columns and CpG sites (derived from Illumina arrays) are included as rows and that file sizes of no greater than 500 MB are uploaded. The application details general information on the platform, as well as information on how to format files and links to all elements of the platform. To make DNAm files smaller prior to upload, users can access a file called 'Truncate_to_these_CpG.csv' which is available in the Zenodo repository. This file allows users to subset CpGs measured in their dataset to those used in 'MethylDetectR – Calculate Your Scores' making files considerably smaller. A 'SexAgeInfo' file may also be uploaded as a .csv file so that data pertaining to the ages and sex of the input individuals are included in the output file along with DNAm-based values for age (epigenetic age), lifestyle and biochemical traits. This file should include a column corresponding to the IDs of individuals in the methylation file ('ID' column), a column that lists the sex of these individuals written as 'Male' or 'Female' or 'NA' ('Sex' column) and a column that reports the chronological or true ages of individuals ('Age' column). Alternatively, the user can merge this information into their DNAm score file after running the 'MethylDetectR – Calculate Your Scores' step. Examples for the DNAm and 'SexAgeInfo' files are available at: https://doi.org/10.5281/zenodo.4646300[26].

The software automatically generates DNAm-based scores for every individual in the dataset (.csv file) and a report for the user is printed on the application detailing quality control steps carried out during the calculation process. For example, the report informs the user whether or not the data had to be transposed, or if M values were converted to beta values (Figure 2).

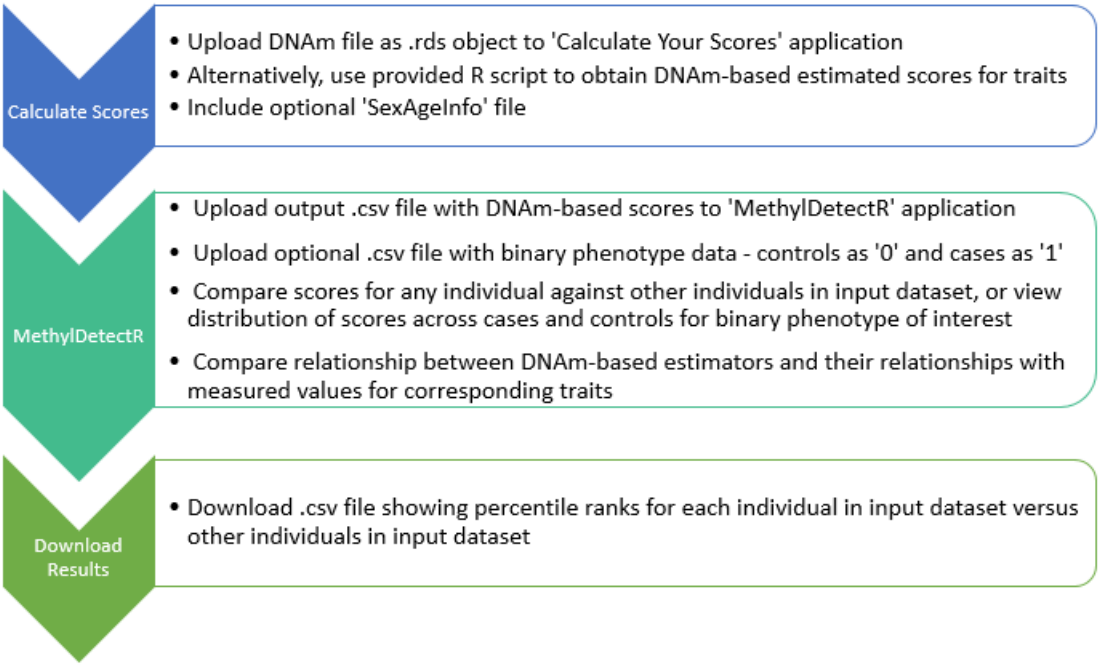

Figure 1. Overview of workflow in 'MethylDetectR' platform. Users upload DNAm data to a secure application named 'MethylDetectR – Calculate Your Scores' or use a provided R script to locally generate DNAm-based values for age and a variety of lifestyle and biochemical traits. DNAm-derived scores are submitted to the 'MethylDetectR' application to view these scores, interactively compare scores within the input sample and optionally view the distribution of scores across cases and controls for uploaded binary phenotypes of interest. CSV files are available to download showing percentile ranks for individuals in the input dataset against other individuals in the uploaded dataset.

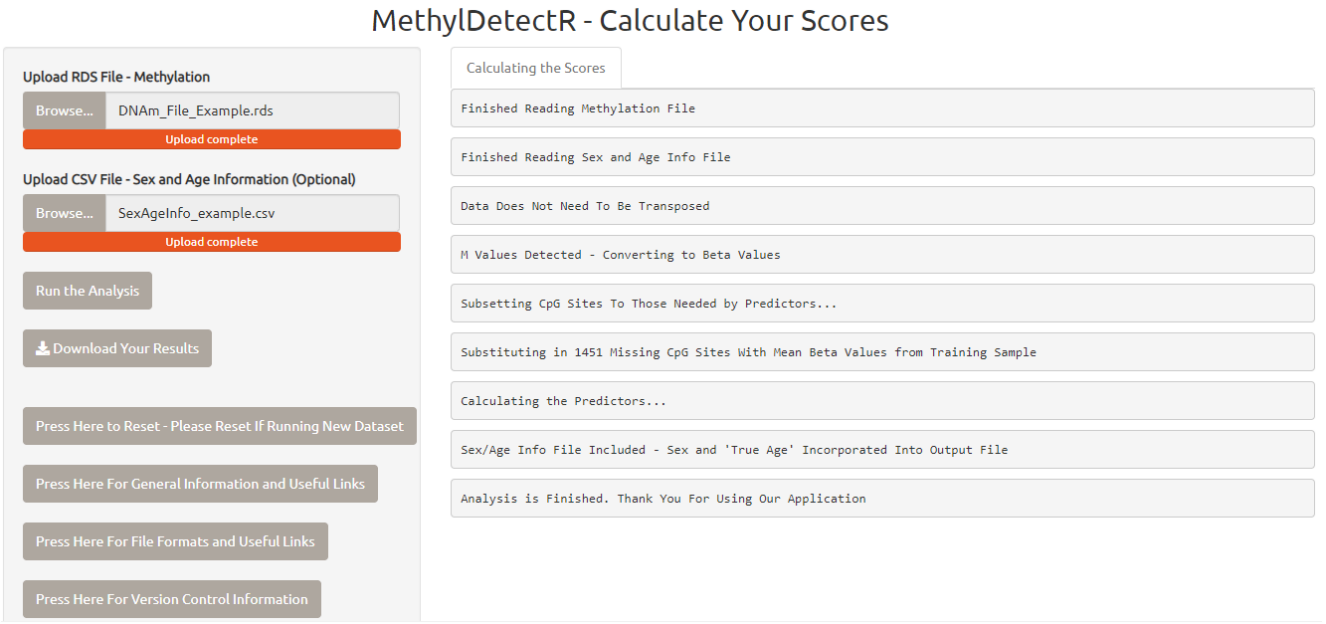

Figure 2. MethylDetectR – Calculate Your Scores Application. An example session for the 'MethylDetectR – Calculate Your Scores' application. In this case, a DNAm dataset, stored as an R file .rds object, has been uploaded along with the optional 'SexAgeInfo'.csv file. A log output has been generated for the user detailing quality control steps which have been carried out in the calculation of DNAm-based predictors. For instance, M values were uploaded and converted to beta values. The resultant output file can be downloaded as a .csv file for upload to the main 'MethylDetectR' application.

Alternatively, the user can download an R script to locally generate DNAm-based scores for the traits. The DNAm object is annotated as 'data' and the 'SexAgeInfo' input is annotated as 'sexageinfo' (https://doi.org/10.5281/zenodo.4646300[26]). In either case, an output .csv file is generated containing DNAm-based scores or values for each trait and for every individual in the input dataset. This output file should be uploaded to the main 'MethylDetectR' application. An example output .csv file showing the correct column names and file structure is available at: https://doi.org/10.5281/zenodo.4646300[26].

MethylDetectR

Panel 1. The output file from either the 'MethylDetectR – Calculate Your Scores' application or a publicly available script should be uploaded to the main 'MethylDetectR' application. Incorrectly assigned column names will be reported to the user, as will files with no individuals or files with non-numeric values. A timeout is triggered following three minutes of inactivity. All panels contain information on the data shown in the panel, and Panel 1 details information on how to format files. Links to all other elements of the platform are shown in each panel via 'info' buttons located in the sidebar panels.

The first panel allows users to choose a predictor of interest and view how a selected individual in the input dataset ranks against the remainder of the input dataset (in pink) in the context of that predictor (Figure 3A). Alternatively, if the user uploads an optional file with binary phenotype information, then users can also subset the data by case vs. control status. In Figure 3B, the user can view where a selected individual's DNAm-based score for body mass index lies along the sample subset by controls (in pink) and cases (in blue) for diabetes. The user can subset to different age ranges and sex in order to see how the selected individual would compare to the truncated sample selection. Users can also download the percentile ranks for every individual in the input dataset when compared against all other individuals in the dataset. Percentile ranks are available for each trait with the exception of age which is reported in years.

Panel 2. In the second panel, users may select multiple traits in order to simultaneously view the percentile ranks for a selected individual in the input dataset when compared against other individuals in the sample (Figure 4A). Furthermore, the user can view how percentile ranks for a given trait vary according to cases and controls for a selected binary phenotype. In Figure 4B, the median percentile for diabetes cases along with the interquartile range (first to third quartile) are plotted for multiple traits, such as body mass index and body fat percentage. Again, the user can use a sidebar functionality to subset by age range and sex.

Panel 3. In the third panel, users can select multiple DNAm-based predictors and view how they correlate with one another in order to visualise their interrelationships and the underlying data structure. This is represented for both the input and GS datasets (Figure 5A). Furthermore, the correlations are updated according to the selected age range and sex. The user can also subset the input dataset to cases, controls or choose to visualise correlation data for cases and controls in the input dataset alongside each other (Figure 5B).

Panel 4. In the fourth and final panel, users can view how well the DNAm-based predictors for age, lifestyle traits and HDL cholesterol correlate with actual values of their respective traits in GS (Figure 6). In this final panel, users can also subset by age range and sex to view how the performance of the predictors varies according to the truncated reference dataset.

Discussion

We have created and implemented the first publicly available online translational platform for methylation-based health profiling. The platform includes a wide variety of traits which are estimated from large-scale DNAm data. These include chronological age, lifestyle traits and biochemical data thereby providing an automatic and comprehensive estimate of individual health profiles from a single blood draw. Users can interactively view how well DNAm-based estimators for various traits perform at an individual level and how DNAm-based estimators stratify according to case and control status for binary phenotypes of interest. The 'MethylDetectR' platform communicates key messages surrounding the development and present limitations of DNAm-based health profiling to the wider research community and public. This is achieved by including 'info' buttons in the sidebar panels of each application that lead to important information for interpreting the presented results, key limitations and general information on DNAm-based scores. Furthermore, the platform is designed to ensure the highest level of data security and safety and is publicly available with open source code and example input and output files. We will continue to update 'MethylDetectR' every three months with the aim of including new DNAm-based predictors of human traits when they come available.

DNAm-based predictors can integrate biological and environmental information to provide important indices of an individual's health status and well-being. These predictors must display high degrees of sensitivity and specificity in order to accurately distinguish individuals on trajectories toward disease and adverse clinical endpoints from those who will remain healthy in a given clinical context. Currently, the DNAm-based predictors in 'MethylDetectR' cannot make consistently accurate predictions at an individual level and therefore cannot yet be reliably applied in a diagnostic or forensic context. Highly-accurate DNAm-based scores can aid in research environments as they may provide more accurate information than self-report data[20]. For instance, the DNAm-based predictor of smoking can provide a more accurate profile of smoking history than responses from participants in questionnaires. Further, the use of DNAm-based scores for a variety of human traits can proxy many phenotypes such as biochemical and lifestyle traits using a single blood draw. Together, these data can help researchers determine relationships between putative risk factors and important health outcomes, and aid in patient stratification paradigms. Further, 'MethylDetectR' serves as an important translational tool showing an interactive, demo version of the platform and substantial information within each application regarding the interpretation and limitations of

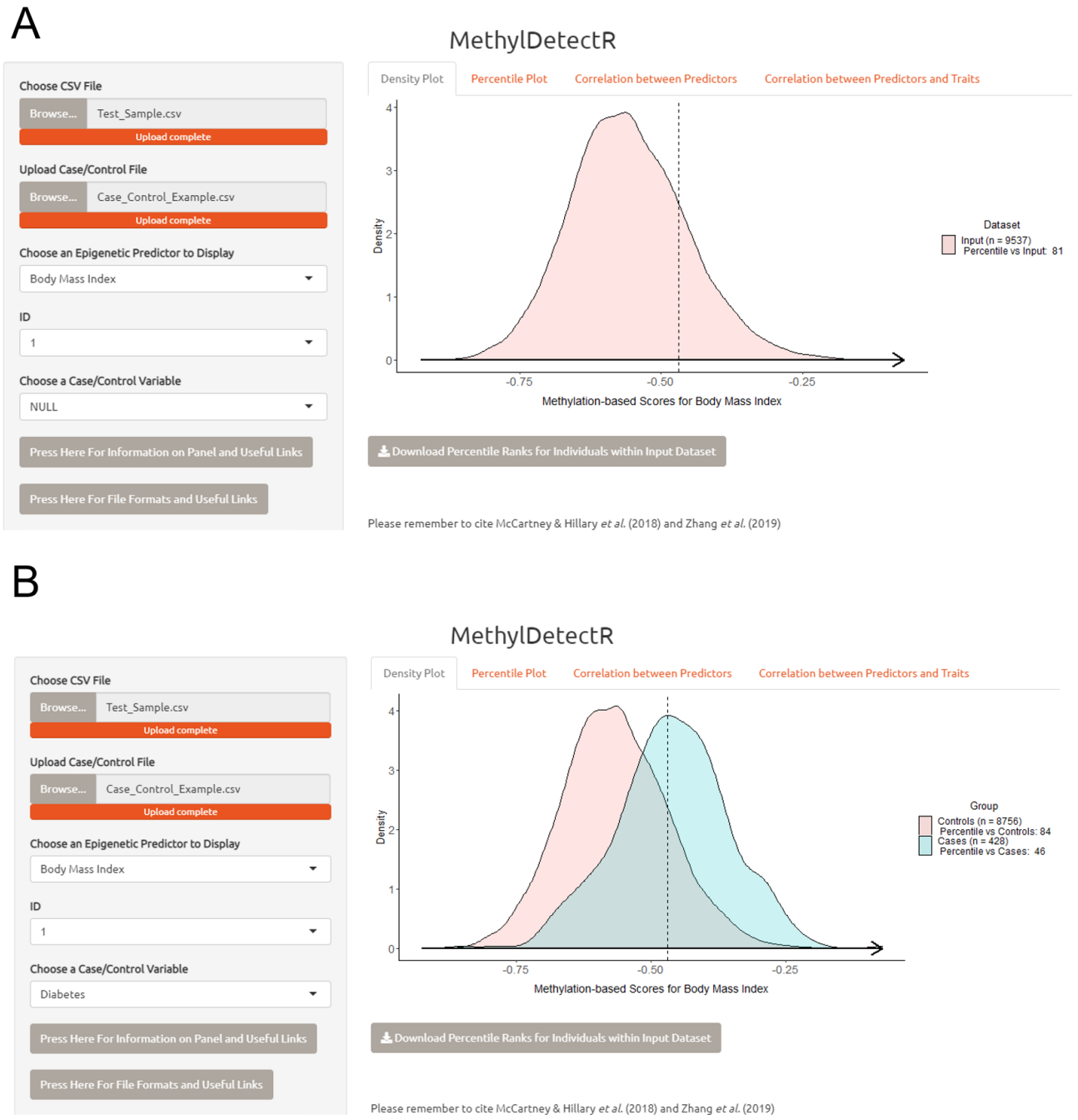

Figure 3. MethylDetectR Application – Panel 1. (**A**) In the first panel, users can select a variable of interest and view how the methylation-based estimate for a chosen individual in the input dataset compares against the remainder of the input dataset (pink). (**B**) The user can subset the scores according to case vs. control status for an uploaded binary phenotype of interest, such as disease status. Here, the distributions of DNAm-based scores for body mass index are plotted for diabetes cases (blue) and controls (pink). Percentile ranks for all individuals in the input dataset when compared against other members of the input dataset may be downloaded as a .csv file in this panel. Percentile ranks are reported for all traits in the output file with the exception of age which is reported in years.

DNAm-based predictors. As these predictors become refined, they may be of clinical value. For instance, a blood-based DNAm test was recently developed that could detect five separate types of cancer up to four years before conventional diagnosis. The assay measured circulating tumour DNA methylation and predicted disease in 88% of post-diagnosis patients, with a specificity of 96%[19].

Distributions of DNAm values and subsequent DNAm-based scores may vary across different methylation datasets. In relation

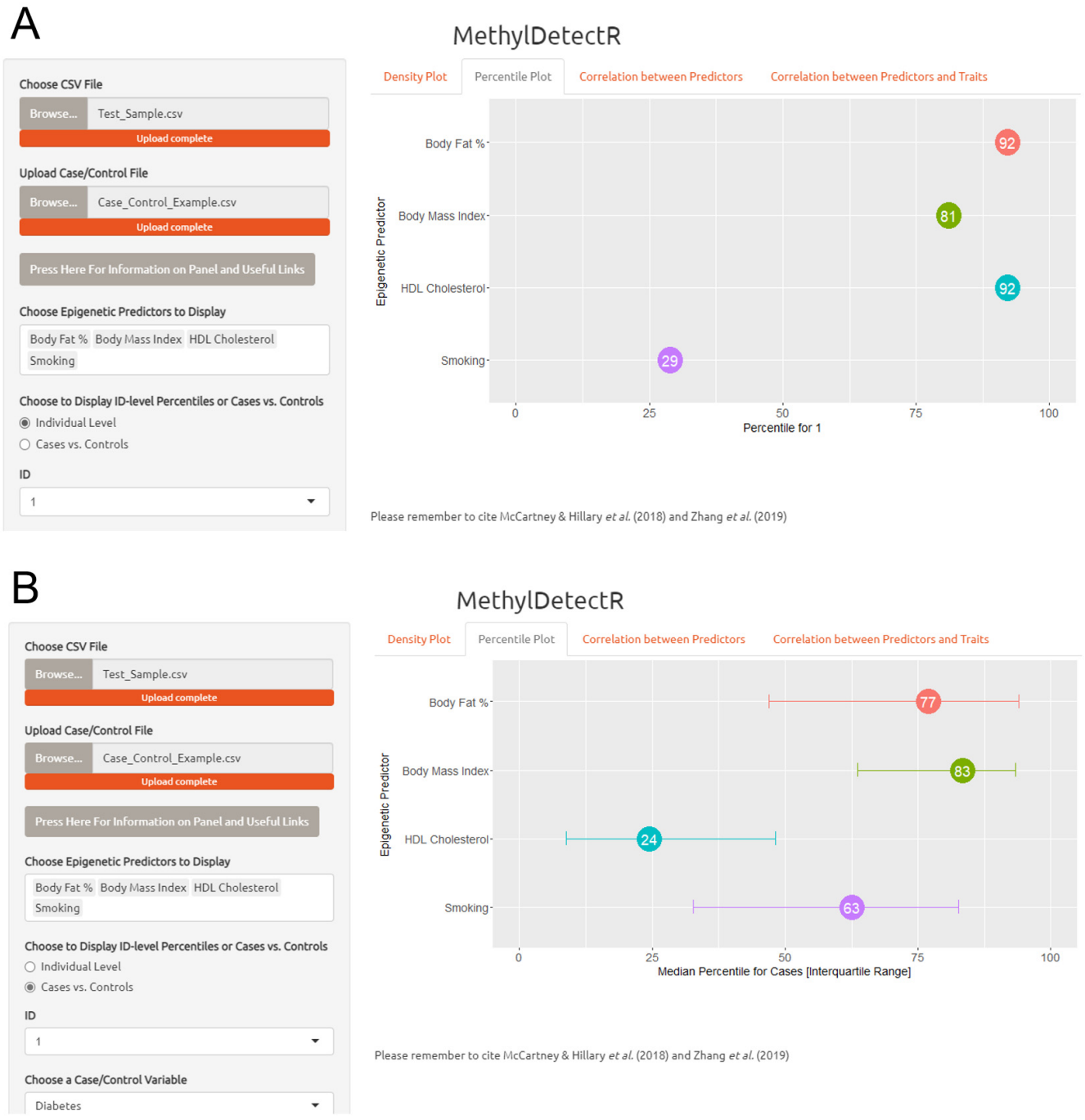

Figure 4. MethylDetectR Application – Panel 2. (**A**) In the second panel, users can select multiple predictors of interest and simultaneously view the percentile rank for a selected individual in relation to these traits when compared against the remainder of the uploaded dataset. The percentile ranks dynamically update according to the selected age range and sex. (**B**) The median percentile ranks for cases are plotted for selected traits of interest. The number shown in the circle reflects the median percentile, and the interquartile range of percentile ranks for cases is shown with the horizontal lines extending from the circle. Here, the median percentile ranks with respect to a number of physical traits are shown for cases of diabetes.

to the biochemical and lifestyle traits, the predictors were generated using an adult sample of individuals with European ancestry. Therefore, it is possible that the predictors may not be generalisable to datasets comprising different age ranges, such as cohorts of children, and individuals with different ancestries. Differences between datasets may also arise from biological differences, for example cases for a given disease may have altered DNAm values for a number of probes relative to controls, or result from technical or normalisation differences. As a result, DNAm-based scores may vary greatly

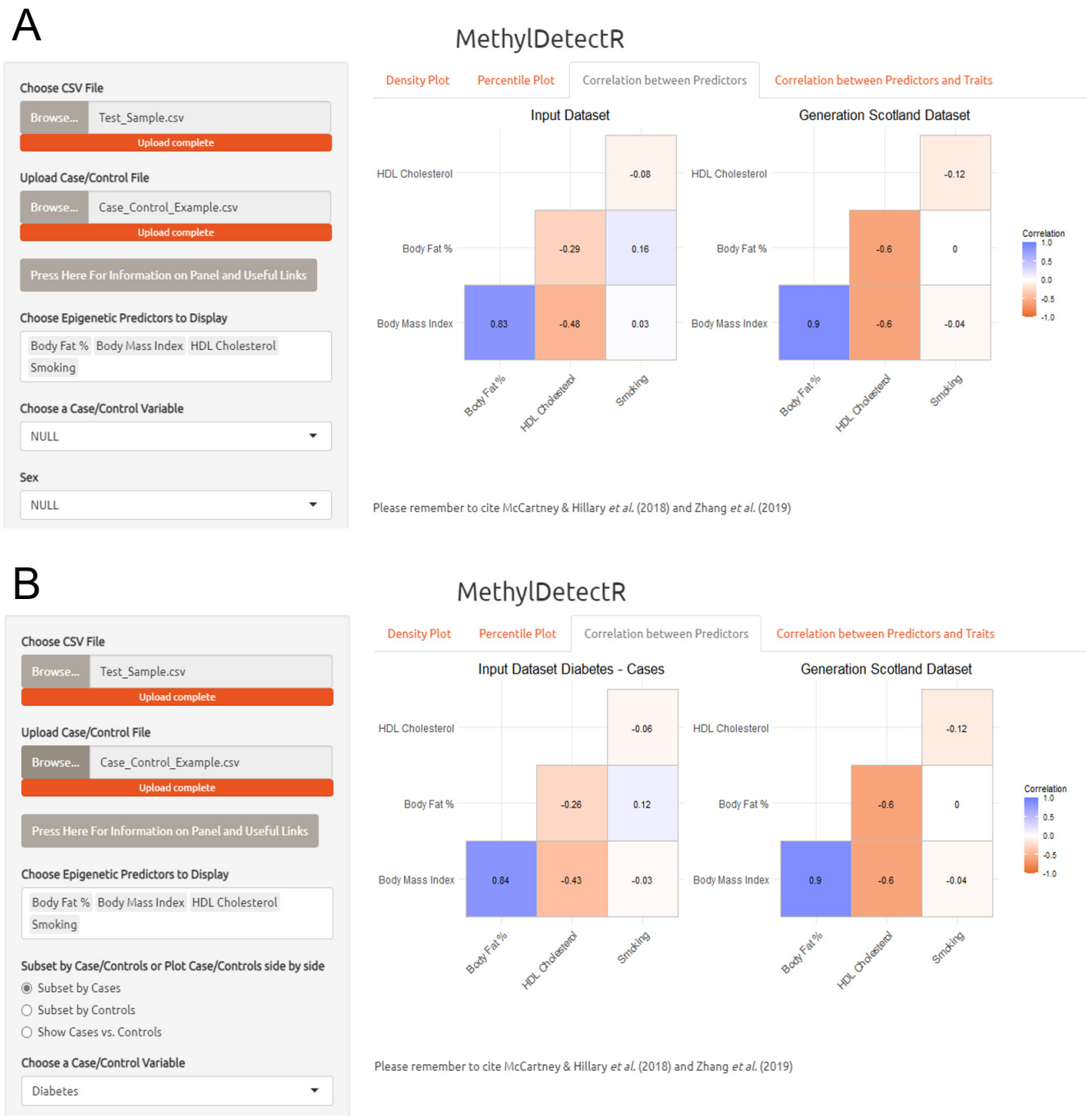

Figure 5. MethylDetectR Application – Panel 3. (**A**) In the third panel, users can select multiple predictors of interest and simultaneously view the interrelationships between these variables of interest in both the input dataset and a reference sample - GS (n = 4,450). (**B**) The user can subset the input dataset according to cases or controls, or choose to view the data structure for cases and controls side by side. The correlation coefficients are updated according to the selected age range and sex.

across datasets and projecting an individual onto a reference sample to view where their DNAm-based score would lie along the reference sample is therefore challenging. Future work will focus on developing methods which can appropriately account for variability across datasets and allow for a projection of individuals onto disparate DNAm samples or datasets.

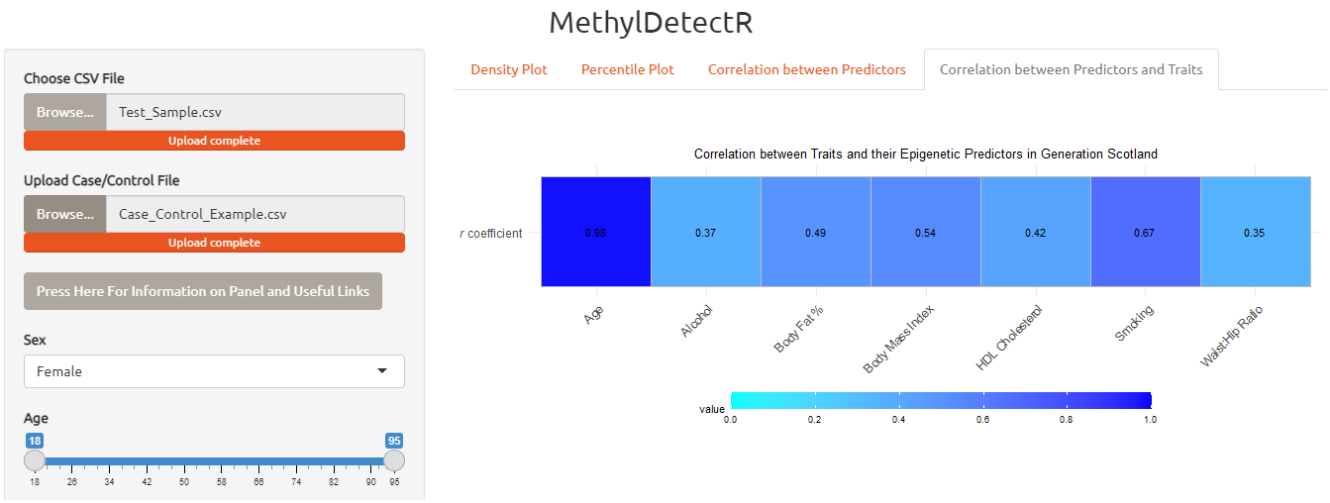

Figure 6. MethylDetectR Application – Panel 4. In the fourth panel, users can view how age, lifestyle traits and high-density lipoprotein (HDL) cholesterol relate to phenotypic or measured values in the GS reference sample (n = 4,450). Correlation coefficients for age, lifestyle traits and HDL cholesterol are updated according to the selected age range and sex. HDL (high-density lipoprotein).

Increased sample sizes through recruitment, consortia or meta-analyses may allow for more sensitive or specific DNAm-based predictors. Advancements in statistical and machine learning approaches used to generate such predictors will also allow for greater accuracy in predicting human traits and health[23]. Furthermore, if the outcomes on which the predictors are trained are inaccurate or possess lots of noise, then the predictors themselves will perform poorly in identifying individuals at risk of disease. Therefore, advancements in understanding disease biology and ways to diagnose or stratify different diseases will help to create well-defined outcomes on which predictors can be trained. This is expected to improve their ability in predicting important health and clinical outcomes. However, stringent ethical frameworks are also necessitated prior to widespread application of molecular-based health profiling in health and forensic contexts[46].

DNAm-based predictors represent one avenue within molecular-based health profiling. Genetics-based predictors of human traits may correlate well with true values for traits, such as human height[47]. However, genetic predictors of disease may often fail to accuracy classify individuals by disease status[48]. Additionally, other 'omics' data have been explored in order to predict human traits or disease. For example, a proteomic signature of age correlates 0.94 with chronological age[49]. Plasma protein-based predictors of disease states, including dementia and cancer, have been explored[50–52]. Lipid-based predictors of human traits have also been developed using plasma samples[53,54]. Complex and common disease states are multifactorial conditions. Therefore, it is likely that composite predictors using various lines of 'omics' data may allow for greater accuracy in predicting disease risk and outcomes when compared to using one line of evidence alone. Furthermore, the incorporation of 'omics' data with clinical or demographic data could provide even more refined predictors of human health and disease[48].

Conclusions

Our platform provides an important translational tool which communicates state-of-the-art developments in relation to DNAm-based predictors of human traits and health. The 'MethylDetectR' platform also represents a research tool for the convenient and secure generation of DNAm-estimated traits for use in clinical and population studies. Importantly, our platform highlights the applicability and limitations surrounding such predictors prior to their potential deployment in clinical assessment and management paradigms.

Reproducibility: All relevant code is available at: https://doi.org/10.5281/zenodo.4646300[26]. The following Research Resource Identifiers (RRIDs) have been generated for 'MethylDetectR – Calculate Your Scores', RRID: SCR_018972, and 'MethylDetectR', RRID: SCR_018973. The limitations surrounding the resultant datasets are that the predictors may work well for risk stratification relative to others in the dataset, but may fail to accurately predict trait information at an individual level.

Data availability

Underlying data

The underlying methylation and phenotypic data used to generate the original predictors cannot be made available due to the data containing information that could compromise participant consent and confidentiality. According to the terms of consent for GS participants, access to data must be reviewed by the GS Access Committee. Applications should be made to access@generationscotland.org. Lothian Birth Cohort 1936 data can be requested from the Lothian Birth Cohort 1936 research team, following completion of a data request application. More information can be found online (http://www.lothianbirthcohort.ed.ac.uk/content/collaboration).

Extended data

Zenodo: MethylDetectR - A Translational Tool for Methylation-Based Health Profiling, https://doi.org/10.5281/zenodo.4646300[26]

This project contains the following extended data:

- DNAm_File_Example.rds. (Example DNAm input file for upload to 'MethylDetectR – Calculate Your Scores'.)
- SexAgeInfo_example.csv. (Example optional 'SexAgeInfo' input file for upload to 'MethylDetectR – Calculate Your Scores'.)
- Truncate_to_these_CpGs.csv. (File for truncating DNAm input file to CpG sites used in 'MethylDetectR – Calculate Your Scores'.)
- MethylDetectR_Test_for_Upload.csv. (Example input file for upload to main 'MethylDetectR' application.)
- MethylDetectR_Case_Control_Example.csv. (Example binary phenotype .csv file with controls coded as '0' and cases coded as '1'.)
- Script_For_User_To_Generate_Scores.R. (R script for user to locally generate DNAm-based estimates of human traits for upload to 'MethylDetectR' application.)
- Predictors_Shiny_by_Groups.csv. (Associated file for use in 'Script_For_User_To_Generate_Scores.R.)

Software availability

Zenodo: MethylDetectR - A Translational Tool for Methylation-Based Health Profiling, https://doi.org/10.5281/zenodo.4646300[26].

This project contains the following scripts:

- MethylDetectR – Calculate Your Scores.R (R script for running of 'MethylDetectR – Calculate Your Scores' application.)
- MethylDetectR.R. (R script for running of 'MethylDetectR' application.)

Reporting guidelines

Open Science Framework: TRIPOD checklist for 'MethylDetectR - a translational platform for methylation-based health profiling', https://doi.org/10.17605/OSF.IO/3MGJT[55]

The TRIPOD checklist is available at: https://www.tripod-statement.org/wp-content/uploads/2020/01/Tripod-Checlist-Prediction-Model-Development.pdf[56].

Acknowledgements

We thank the staff and participants involved in the Lothian Birth Cohort 1936 and Generation Scotland studies for their ongoing commitment and contribution to these studies. We thank Dr Qian Zhang and Prof Peter Visscher for their permission to use the age predictor. We thank Dr Adam Jackson at the Institute of Genetics and Cancer, University of Edinburgh for developing the 'MethylDetectR' website, and for his expertise and advice in preparing the website content. We thank Mr Stephen Cass and Mr Ewan McDowall at the Institute of Genetics and Cancer, University of Edinburgh for their advice on security and safety aspects surrounding the platform and for managing and hosting the secure server from which 'MethylDetectR' operates. We also thank Dr Rena Gertz, University of Edinburgh Data Protection Officer, and Victoria Rowntree, Assistant Data Protection Officer, for their review and approval of our software as GDPR compliant, and for their advice on Participant Information and Informed Consent documentation.

References

1. Jaenisch R, Bird A: **Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals.** *Nat Genet.* 2003; **33**(Suppl): 245–54.
PubMed Abstract | Publisher Full Text

2. Beck S, Rakyan VK: **The methylome: approaches for global DNA methylation profiling.** *Trends Genet.* 2008; **24**(5): 231–7.
PubMed Abstract | Publisher Full Text

3. Bestor TH, Edwards JR, Boulard M: **Notes on the role of dynamic DNA methylation in mammalian development.** *Proc Natl Acad Sci U S A.* 2015; **112**(22): 6796–9.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Hannon E, Lunnon K, Schalkwyk L, *et al.*: **Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes.** *Epigenetics.* 2015; **10**(11): 1024–32.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Joehanes R, Just AC, Marioni RE, *et al.*: **Epigenetic Signatures of Cigarette Smoking.** *Circ Cardiovasc Genet.* 2016; **9**(5): 436–447.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Stevenson AJ, McCartney DL, Hillary RF, *et al.*: **Characterisation of an inflammation-related epigenetic score and its association with cognitive ability.** *Clin Epigenetics.* 2020; **12**(1): 113.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Hillary RF, Stevenson AJ, McCartney DL, *et al.*: **Epigenetic measures of ageing predict the prevalence and incidence of leading causes of death and disease burden.** *Clin Epigenetics.* 2020; **12**(1): 115.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Langdon RJ, Beynon RA, Ingarfield K, *et al.*: **Epigenetic prediction of complex traits and mortality in a cohort of individuals with oropharyngeal cancer.** *Clin Epigenetics.* 2020; **12**(1): 58.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Lu AT, Seeboth A, Tsai PC, *et al.*: **DNA methylation-based estimator of**

telomere length. *Aging (Albany NY).* 2019; **11**(16): 5895–923.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Lu AT, Quach A, Wilson JG, *et al.*: **DNA methylation GrimAge strongly predicts lifespan and healthspan.** *Aging (Albany NY).* 2019; **11**(2): 303–327.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Hannum G, Guinney J, Zhao L, *et al.*: **Genome-wide methylation profiles reveal quantitative views of human aging rates.** *Mol Cell.* 2013; **49**(2): 359–67.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Levine ME, Lu AT, Quach A, *et al.*: **An epigenetic biomarker of aging for lifespan and healthspan.** *Aging (Albany NY).* 2018; **10**(4): 573–91.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Horvath S: **DNA methylation age of human tissues and cell types.** *Genome Biol.* 2013; **14**(10): R115.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Rosen AD, Robertson KD, Hlady RA, *et al.*: **DNA methylation age is accelerated in alcohol dependence.** *Transl Psychiatry.* 2018; **8**(1): 182.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Horvath S, Garagnani P, Bacalini MG, *et al.*: **Accelerated epigenetic aging in Down syndrome.** *Aging Cell.* 2015; **14**(3): 491–5.
PubMed Abstract | Publisher Full Text | Free Full Text

16. McCrory C, Fiorito G, Hernandez B, *et al.*: **Association of 4 epigenetic clocks with measures of functional health, cognition, and all-cause mortality in The Irish Longitudinal Study on Ageing (TILDA).** *BioRxiv.* 2020.
Publisher Full Text

17. Zhao W, Ammous F, Ratliff S, *et al.*: **Education and Lifestyle Factors Are Associated with DNA Methylation Clocks in Older African Americans.** *Int J Environ Res Public Health.* 2019; **16**(17): 3141.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Barbu MC, Shen X, Walker RM, *et al.*: **Epigenetic prediction of major depressive disorder.** *Mol Psychiatry.* 2020.
PubMed Abstract | Publisher Full Text

19. Chen X, Gole J, Gore A, *et al.*: **Non-invasive early detection of cancer four years before conventional diagnosis using a blood test.** *Nat Commun.* 2020; **11**(1): 3475.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Zhang Y, Elgizouli M, Schöttker B, *et al.*: **Smoking-associated DNA methylation markers predict lung cancer incidence.** *Clin Epigenetics.* 2016; **8**: 127.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Salameh Y, Bejaoui Y, El Hajj N, *et al.*: **DNA Methylation Biomarkers in Aging and Age-Related Diseases.** *Front Genet.* 2020; **11**: 171.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Zhang Q, Vallerga CL, Walker RM, *et al.*: **Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing.** *Genome Med.* 2019; **11**(1): 54.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Trejo Banos D, McCartney DL, Patxot M, *et al.*: **Bayesian reassessment of the epigenetic architecture of complex traits.** *Nat Commun.* 2020; **11**(1): 2865.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Dupras C, Beauchamp E, Joly Y, *et al.*: **Selling direct-to-consumer epigenetic tests: are we ready?** *Nat Rev Genet.* 2020; **21**(6): 335–336.
PubMed Abstract | Publisher Full Text

25. McCartney DL, Hillary RF, Stevenson AJ, *et al.*: **Epigenetic prediction of complex traits and death.** *Genome Biol.* 2018; **19**(1): 136.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Hillary: **MethylDetectR - A Translational Tool for Methylation-Based Health Profiling (Version 5.0) [Data set].** Wellcome Open. *Zenodo*. 2020.
http://www.doi.org/10.5281/zenodo.4646300

27. Hüls A, Czamara D: **Methodological challenges in constructing DNA methylation risk scores.** *Epigenetics.* 2020; **15**(1–2): 1–11.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Zou H, Hastie T: **Regularization and variable selection via the elastic net.** *J R Statist Soc B.* 2005; **67**(2): 301–20.
Publisher Full Text

29. Robinson GK: **That BLUP is a good thing: the estimation of random effects.** *Stat Sci.* 1991; **6**(1): 15–32.
Reference Source

30. Friedman J, Hastie T, Tibshirani R: **Regularization paths for generalized linear models via coordinate descent.** *J Stat Softw.* 2010; **33**(1): 1–22.
PubMed Abstract | Free Full Text

31. Chang W, Cheng J, Allaire J, *et al.*: **Shiny: Web Application Framework for R; R package version 1.4. 0.2**. 2020.

32. Team RC: **R version 3.5. 0. R: A language and environment for statistical computing R Foundation for Statistical Computing**. Vienna, Austria. 2018.
Reference Source

33. Perrier V, Meyer F, Granjon D: **shinyWidgets: Custom Inputs Widgets for Shiny**. R package version. 2019.

34. Chang W, Park T, Dziedzic L, *et al.*: **shinythemes: Themes for Shiny.** *R package version.* 2015; **1**(1): 144.

35. Dowle M, Srinivasan A: **data. table: Extension of 'data. frame**. R package version 1.12. 8. 2019.

36. Attali D, Edwards T: **shinyalert: Easily Create Pretty Popup Messages (Modals) in'Shiny'**. R package version 10. 2018.
Reference Source

37. Wickham H: **ggplot2: elegant graphics for data analysis**. springer. 2016.
Reference Source

38. Hadley Wickham RF, Henry L, Müller K: **dplyr: A Grammar of Data Manipulation**. R package version 0.7. 4. 2017.

39. Wickham H: **Tools for working with categorical variables (factors)(R package Version 0.4. 0)[Computer software]**. 2019.

40. Ram K, Wickham H: **wesanderson: A Wes Anderson palette generator**. R package version 0.3. 6. 2018.
Reference Source

41. Sali A: **shinycssloaders: Add CSS Loading Animations to "shiny" Outputs**. R Package Version 02 0. 2017.

42. Ooms J: **Magick: advanced graphics and image-processing in R**. CRAN R package version. 2018; 1.
Reference Source

43. Wei T, Simko V: **R package "corrplot": Visualization of a Correlation Matrix (Version 0.84)**. 2017.
Reference Source

44. Kassambara A: **ggcorrplot: Visualization of a Correlation Matrix using'ggplot2'**. R package version 01. 2016; 1.
Reference Source

45. Wilke CO: **cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2**. R package version 0.9. 4. 2019.

46. Bell CG, Lowe R, Adams PD, *et al.*: **DNA methylation aging clocks: challenges and recommendations.** *Genome Biol.* 2019; **20**(1): 249.
PubMed Abstract | Publisher Full Text | Free Full Text

47. Lello L, Avery SG, Tellier L, *et al.*: **Accurate Genomic Prediction of Human Height.** *Genetics.* 2018; **210**(2): 477–97.
PubMed Abstract | Publisher Full Text | Free Full Text

48. Lewis CM, Vassos E: **Polygenic risk scores: from research tools to clinical instruments.** *Genome Med.* 2020; **12**(1): 44.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Tanaka T, Biancotto A, Moaddel R, *et al.*: **Plasma proteomic signature of age in healthy humans.** *Aging Cell.* 2018; **17**(5): e12799.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Tanaka T, Lavery R, Varma V, *et al.*: **Plasma proteomic signatures predict dementia and cognitive impairment.** *Alzheimers Dement (N Y).* 2020; **6**(1): e12018.
PubMed Abstract | Publisher Full Text | Free Full Text

51. Nedjadi T, Benabdelkamal H, Albarakati N, *et al.*: **Circulating proteomic signature for detection of biomarkers in bladder cancer patients.** *Sci Rep.* 2020; **10**(1): 10999.
PubMed Abstract | Publisher Full Text | Free Full Text

52. Cuvelliez M, Vandewalle V, Brunin M, *et al.*: **Circulating proteomic signature of early death in heart failure patients with reduced ejection fraction.** *Sci Rep.* 2019; **9**(1): 19202.
PubMed Abstract | Publisher Full Text | Free Full Text

53. Mundra PA, Barlow CK, Nestel PJ, *et al.*: **Large-scale plasma lipidomic profiling identifies lipids that predict cardiovascular events in secondary prevention.** *JCI Insight.* 2018; **3**(17): e121326.
PubMed Abstract | Publisher Full Text | Free Full Text

54. Gerl MJ, Klose C, Surma MA, *et al.*: **Machine learning of human plasma lipidomes for obesity estimation in a large population cohort.** *PLoS Biol.* 2019; **17**(10): e3000443.
PubMed Abstract | Publisher Full Text | Free Full Text

55. Hillary RF: **TRIPOD Checklist for MethylDetectR - A Translational Tool For Methylation-Based Health Profiling**. 2020.

56. Collins GS, Reitsma JB, Altman DG, *et al.*: **Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) The TRIPOD Statement.** *Circulation.* 2015; **131**(2): 211–9.

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 27 April 2021

https://doi.org/10.21956/wellcomeopenres.18504.r43463

✔ **Charlotte Cecil**

[1] Department of Child and Adolescent Psychiatry, Erasmus MC, Rotterdam, The Netherlands
[2] Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands
[3] Molecular Epidemiology, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands
[4] Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK

The authors have done an excellent job at revising the manuscript and online application in light of the reviewer comments. I particularly commend the authors for taking the time to amend the app by providing information that will help users to contextualize and interpret the scores - I believe that it is now much improved and have no further comments.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Epigenetic epidemiology; psychiatric epidemiology; child development; developmental psychopathology; early life stress.

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.

Reviewer Report 13 April 2021

https://doi.org/10.21956/wellcomeopenres.18504.r43464

✔ **Gemma Sharp**

MRC Integrative Epidemiology Unit, Population Health Sciences, University of Bristol, Bristol, UK

The authors have updated the manuscript and I am satisfied with the amendments.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Epigenetic epidemiology, R, Shiny.

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.

Version 1

Reviewer Report 16 February 2021

https://doi.org/10.21956/wellcomeopenres.18118.r42132

?

Charlotte Cecil
[1] Department of Child and Adolescent Psychiatry, Erasmus MC, Rotterdam, The Netherlands
[2] Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands
[3] Molecular Epidemiology, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands
[4] Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK

In the present study, the authors present a new publicly available online platform for methylation-based health profiling. The manuscript is clear, well written and addresses an important emerging topic in population epigenetics. The online platform enables the calculation of multiple DNA methylation (DNAm) scores for age, lifestyle factors and protein levels related to neurology or inflammation. Based on these scores, users are able to view percentile ranks for specific individuals as compared to others in the input sample, with the option to subset by case-control status, sex and age. Users are also able to estimate a broader health profile for specific individuals by displaying percentiles for multiple DNAm scores simultaneously. Finally, users can check correlations between their predicted DNAm scores compared to those in the Generation Scotland dataset, as well as correlations between predicted DNAm scores and actual measured traits in Generation Scotland. Overall, I think this is an excellent contribution to the field and an important first step for moving methylation-based predictors into translational applications. I am concerned however, that no practical information is provided, especially within the online platform, to aid users in the correct interpretation of the scores themselves, and the findings generated from input datasets. If the platform is to reach widespread use and fulfil its translational potential, I would anticipate that part of the users may not have a research background and many may not read this manuscript in detail before utilizing the platform. As such, I think it is essential to add

information within the online tool to guide users and ensure that scores are interpreted adequately with full transparency on limitations (e.g. applicability to other age ranges, populations, error around DNAm estimates, meaning of different percentile ranks and how those should be interpreted etc.). I would also like to see in the manuscript a more detailed discussion of potential uses for this platform in different contexts, using practical examples (e.g. screening and risk prediction in research and clinical settings). I believe these steps are necessary in order to ensure that the platform is adequately used without DNAm scores being overly- or incorrectly-interpreted, especially if the hope is that such a platform will be utilized in a clinical context. Specific comments are listed below.

- How 'live' is this online resource? Given the fast-paced developments in population epigenetic methods and prediction tools, how regularly will the resource be updated to take stock of these developments (both in terms of the actual calculation of methylation-based scores, as well as the information provided regarding applicability and limitations)?

- Methods – implementation
  - The authors provide a link to their online interface for calculating DNAm scores. I would anticipate that, at least for research use, many studies would not be allowed to upload individual level DNAm data due to data protection policies, even when anonymized (some institutions have concerns that omics level data can never be truly anonymized), so it is useful that scripts are also provided to run the analyses locally.
  - Is it possible to upload potential covariates, such as cell-type, batch, ancestry etc.? How were these factors, especially cell-type heterogeneity, taken into account in the development of MethylDetecR DNAm scores?
  - Is this tool adequate for input datasets containing samples who may differ in characteristics from those upon which the algorithms were based (e.g. datasets of different ancestry)? Further, could the tool be utilized on child populations, where certain but not all scores would be applicable (e.g. BMI but not alcohol), if the algorithms were developed on adults? And regarding protein levels, how significant is the fact that these scores were developed from a sample of adults in old age? Can we assume that the same relationships between DNAm and protein levels hold across different life stages?

- Demo
  - It would help to add information in the demo about how to interpret the different modules (e.g. a walkthrough, or pop-up bubbles), for example:
    - **First module** how should a user (especially one who may not have a research background) interpret the density plots, what this means when stratified by case-control status, and how to evaluate a given individual's score within these plots? For example, with regards to the smoking DNAm score, what should a user conclude if an individual of interest scores on the 75^{th} percentile overall, and specifically on the 99^{th} percentile within controls and 14^{th} percentile within cases when stratified by smoker case-control status? Should the user interpret this as a sign that the individual in question is (probably) a smoker or ex-smoker? (p.s. I see later in the manuscript that ex-smokers were removed from the data upon which the prediction was based, which makes the above example all the more relevant).
    - **Second module** how should users interpret the percentile ranks for multiple traits? I would expect users, especially for more clinical applications, to look for

a particular threshold with which to guide their interpretation. For example, should anyone with scores above a certain percentile (e.g. 50th, 75th or higher) be considered to have a risky profile (e.g. in terms of DNAm signatures for smoking, body fat etc.)? With regards to the protein DNAm scores, how should a user interpret an individual scoring for example on the 27th percentile for IL6, but the 80th percentile for TNF.alpha? What implications would this have for making assessments (and thus also potential recommendations) regarding this person's inflammation profile?

- On a related note, is there a way to help users gauge how 'reliable' these different scores are? Some DNAm predictors (e.g. smoking) are more reliable and accurate than others, could this be indicated in the platform e.g. by visualizing this in the plots or adding text to aid interpretation (i.e. to what extent could a user interpret the percentiles for different scores with confidence)? I see in page 10 that error bars can be displayed, but (a) I do not see where the option to visualize these is in the online tool and this is also not explained in the manuscript; (b) I am assuming that these error bars are related to the input data, so a bit different from the point of indicating how good the scores themselves are. I assume that this can be partly gauged by the information on correlations in the fourth panel, although this includes only a selection of the DNAm scores.

- **Third panel** how should users interpret differences in correlations between their input dataset and the reference GS dataset? What are the implications of these differences for interpreting their results (e.g. does it mean that scores should be interpreted with more caution or has no bearing on this)? For example, if the correlation between alcohol and smoking DNAm scores in the input dataset is 0.40, but in GS it is only 0.20, what does this mean for the user?

- In the discussion, the authors state that '*the MethylDetecR platform communicates key messages surrounding the development and present limitations of DNAm-based health profiling to the wider research community and public*' - where is this information provided? I cannot see it either in the Demo or Calculate your Scores platforms. This should be provided in the Demo, as the first 'port of entry' users should be using to get familiarized with the platform (and there should be a notice on the Calculate your Scores platform to first check out the Demo, for an example of correct use). This is a good place to walk users through the different functionalities/panels of the platform, and for each panel, to provide key information on interpretation (e.g. using case examples – "here is Joe, he is estimated to be xx years old and score within the xx percentile for yy variable. Based on this, we can conclude that xxx. Limitations to keep in mind are xxx. This information could be used for xxx applications").

- Similarly, in the conclusion, the authors state that "*our platform provides an important translational tool which communicates state-of-the-art developments in relation to DNAm-based predictors of human traits and health… and highlights the applicability and limitations surrounding such predictors prior to their potential deployment in clinical assessment and management paradigms*." While the platform allows to calculate state-of-the-art scores, I cannot see anything related to communication of its uses and limitations.

- Generally, the manuscript refers to a lot of different files and links (e.g. the Demo, the actual platform, all the Zenodo materials). This relies on the idea that users in future will always refer to the manuscript and read it in detail before using the platform. I would suggest having an easy to find, centralized repository for all of these materials, or alternatively, to add the various links in the Demo/Calculate your Scores platforms (which should be linked somehow).

- The discussion is missing a section on the research and clinical applications of this platform. Can the authors provide concrete examples of where and how this information could be useful?

- Minor points
  - P.3: "*In contrast to genetics, these molecular modifications are dynamic, tissue-specific and reversible*" – would qualify this sentence a little bit. While it is correct that, overall, DNAm differ largely across tissues, some sites do show strong correspondence across multiple tissues. Similarly, it is unclear whether all DNAm changes are reversible (e.g. some smoking-related DNAm alterations persisting post-cessation).
  - P. 3: "*Furthermore, there has been a rapid increase in the commercialisation and scalability of DNAm assays for direct-to-consumer use or for use in clinical, research or industrial settings*." Can the authors add references or include a couple of examples?
  - P.3: "*Blood DNAm data …provides a good index of the overall health status of the body*." Can the authors add a reference to support this?
  - P. 4: *"Lastly, the user can subset the reference samples by sex, age range or case vs. control status."* Case-control status on what variables?
  - P. 5: https://www.ed.ac.uk/centre-genomic-medicine/research-groups/
  - marioni-group/methyldetectr the link provided has two attachments, which appear to contain the same information: MethylDetectR Participant Information Sheet and MethylDetectR Participant Consent Statement
  - P. 5: in the explanation of methylation risk scores, the authors could point readers to a new review on the topic, including strengths and challenges of this approach (Huls & Czamara, 2020 *Epigenetics*)[1].
  - P. 6: Age predictor - can the authors specify the age range of the discovery data upon which the predictor was based?
  - P. 6: Can the author provide a reference/specify how cutoffs were selected for their protein analyses (e.g. for training-test ratio, *r* and *p* thresholds etc.).

References

1. Hüls A, Czamara D: Methodological challenges in constructing DNA methylation risk scores. *Epigenetics*. **15** (1-2): 1-11 PubMed Abstract | Publisher Full Text

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?
No

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Epigenetic epidemiology; psychiatric epidemiology; child development; developmental psychopathology; early life stress.

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, however I have significant reservations, as outlined above.

Author Response 30 Mar 2021

Robert Hillary, University of Edinburgh, Edinburgh, UK

Comment 1: I am concerned, however, that no practical information is provided, especially within the online platform, to aid users in the correct interpretation of the scores themselves, and the findings generated from input datasets. If the platform is to reach widespread use and fulfil its translational potential, I would anticipate that part of the users may not have a research background and many may not read this manuscript in detail before utilizing the platform. As such, I think it is essential to add information within the online tool to guide users and ensure that scores are interpreted adequately with full transparency on limitations (e.g. applicability to other age ranges, populations, error around DNAm estimates, meaning of different percentile ranks and how those should be interpreted etc.).

Response: We thank the reviewer for their excellent suggestions towards improving the platform. We agree that the applications would benefit from more general information on what the predictors reflect and what we can conclude from them in light of limitations. This would achieve an appropriate balance between its applicability for expert and non-expert users alike. As a result, we have added 'info' action buttons on each panel of the applications to reveal important information pertaining to the DNAm-based scores and their interpretation. Further, we include links in each application to all other elements of the platform i.e. the other applications, the website, the paper and the Zenodo repository.

Comment 2: I would also like to see in the manuscript a more detailed discussion of potential uses for this platform in different contexts, using practical examples (e.g. screening and risk prediction in research and clinical settings). I believe these steps are

necessary in order to ensure that the platform is adequately used without DNAm scores being overly- or incorrectly-interpreted, especially if the hope is that such a platform will be utilized in a clinical context.

Response: We thank the reviewer for highlighting the need to include more information regarding potential applications of the DNAm-based scores in 'MethylDetectR' in research and clinical settings and the appropriate interpretation of their potential utility. To address this, we have amended Paragraph 2 in the 'Discussion' section to include the following:

"Currently, the DNAm-based predictors in 'MethylDetectR' cannot make consistently accurate predictions at an individual level and therefore cannot yet be reliably applied in a diagnostic or forensic context. Highly-accurate DNAm-based scores can aid in research environments as they may provide more accurate information than self-report data (6). For instance, the DNAm-based predictor of smoking can provide a more accurate profile of smoking history than responses from participants in questionnaires. Further, the use of DNAm-based scores for a variety of human traits can proxy many phenotypes such as biochemical and lifestyle traits using a single blood draw. Together, these data can help researchers determine relationships between putative risk factors and important health outcomes, and aid in patient stratification paradigms. Further, 'MethylDetectR' serves as an important translational tool showing an interactive, demo version of the platform and substantial information within each application regarding the interpretation and limitations of DNAm-based predictors. As these predictors become refined, they may be of clinical value. For instance, a blood-based DNAm test was recently developed that could detect five separate types of cancer up to four years before conventional diagnosis. The assay measured circulating tumour DNA methylation and predicted disease in 88% of post-diagnosis patients, with a specificity of 96%".

Comment 3: How 'live' is this online resource? Given the fast-paced developments in population epigenetic methods and prediction tools, how regularly will the resource be updated to take stock of these developments (both in terms of the actual calculation of methylation-based scores, as well as the information provided regarding applicability and limitations)?

Response: The platform will be continually updated when robust DNAm-based predictors are generated by our groups and others. We have added a section called 'Version Control' within the 'Implementation' subsection in 'Methods'. This section details who will be responsible for managing 'MethylDetectR', and that it will be updated every three months. We also invite other researchers to contact us if they wish to have their DNAm-based predictors added to 'MethylDetectR'.

Version Control. *We will update 'MethylDetectR' every three months to include new DNAm-based predictors of human traits as they are generated by our own group and others. Updates will be managed by Robert F. Hillary or Riccardo E. Marioni. If researchers wish to have their predictors considered for inclusion in 'MethylDetectR', please use the corresponding author email address in this manuscript or the contact details available at the 'MethylDetectR' website (https://www.ed.ac.uk/centre-genomic-medicine/research-groups/marioni-group/methyldetectr). The current and historical versions of 'MethylDetectR' are available in the Zenodo repository, updated versions will also be made available in this repository (*

https://doi.org/10.5281/zenodo.4646300).

Further, we have created 'Version Control' action buttons at the start of each application, this details the history of updates to 'MethylDetectR' and indicates that its different versions can be accessed through our Zenodo repository. We also take this opportunity to highlight that our updated version of 'MethylDetectR' associated with this resubmission temporarily omits protein predictors from the dataset. The reason for this is that we are generating, and aim to publish, predictors for 109 blood protein levels (previously 27 proteins). We are committed to updating and managing 'MethylDetectR' as a live platform with new predictors added every three months. We will also continually monitor the platform to update information presented on the applications and website in light of feedback obtained from users and peers. The following text has been added the following text in Paragraph 1 of the 'Discussion':

"We will continue to update 'MethylDetectR' every three months with the aim of including new DNAm-based predictors of human traits when they come available".

Comment 4: The authors provide a link to their online interface for calculating DNAm scores. I would anticipate that, at least for research use, many studies would not be allowed to upload individual level DNAm data due to data protection policies, even when anonymized (some institutions have concerns that omics level data can never be truly anonymized), so it is useful that scripts are also provided to run the analyses locally.

Response: We thank the reviewer for this comment. We also highlight that we have now included a data protection statement early within the text at the beginning of the 'Methods' section:

*"**Data Protection and Privacy**. No data are stored in 'MethylDetectR' and are deleted upon closing the applications. Applications are also timed out after three minutes of inactivity and are hosted on patched and secure servers within the Institute of Genetics and Cancer, University of Edinburgh. This research and translational tool complies with GDPR guidelines and has been designed to ensure the highest level of data security and privacy. 'MethylDetectR' and information on its usage are also available at the following website: https://www.ed.ac.uk/centre-genomic-medicine/research-groups/marioni-group/methyldetectr. Information relating to participant consent is also available at this website. Given that no data are stored, this information pertains to general risk surrounding the upload of biological data to online software and the measures taken to mitigate the risk of motivated intruders gaining access to such data."*

Comment 5: Is it possible to upload potential covariates, such as cell-type, batch, ancestry etc.? How were these factors, especially cell-type heterogeneity, taken into account in the development of MethylDetectR DNAm scores?

Response: When the predictors for lifestyle and biochemical traits were generated, they were pre-corrected for age, sex as well as genetic principal components to remove the potential influence of ancestry. We did not correct for cell-type heterogeneity in our training data. This is clarified with the addition of the following text under the 'Lifestyle and biochemical traits' subsection in 'Methods':

"In generating the predictors, phenotypes were pre-corrected to remove the influence of age, sex and ancestry using ten genetic principal components. Phenotypic data used to train the predictors were not corrected for cell-type heterogeneity."

We also explored the addition of a 'Covariates' file to the 'Calculate Your Scores' application. When testing this, we felt that it may have been confusing for users given that not all users may wish to add covariates, and may wish to use some covariates to pre-correct DNAm data prior to the calculation of scores and also perhaps use other covariates to residualise the scores themselves. This is likely specific to each study. However, to address this excellent point raised by the reviewer, we have instead added text within the 'Press Here for File Formats and Useful Links' 'info' button on the sidebar panel of the 'Calculate Your Scores' application to highlight that the user may wish to adjust the scores following their download for important covariates dependent on their study design or motivation to use 'MethylDetectR'. We highlight cell-type heterogeneity as a key example of this discussion point:

In 'Calculate Your Scores': "Once you have downloaded the DNAm-based scores, you may wish to adjust them for covariates, such as cell type counts or proportions. Importantly, when the predictors were created, they were trained on phenotypic data that were not adjusted for cell-type heterogeneity. The need for covariate adjustments will be specific to the aims of each study."

Comment 6: Is this tool adequate for input datasets containing samples who may differ in characteristics from those upon which the algorithms were based (e.g. datasets of different ancestry)? Further, could the tool be utilized on child populations, where certain but not all scores would be applicable (e.g. BMI but not alcohol), if the algorithms were developed on adults? And regarding protein levels, how significant is the fact that these scores were developed from a sample of adults in old age? Can we assume that the same relationships between DNAm and protein levels hold across different life stages?

Response: The reviewer raises excellent points of discussion. Our datasets were trained on an adult sample. It is possible that the predictors are not generalisable to different age groups and individuals with different ethnic backgrounds. The protein levels have been omitted temporarily from the platform but we will include a larger set of 109 proteins owing to refinements in the pipeline used to generate the protein predictors. In the future, we will adapt the platform to allow for comparisons of DNAm-based scores across samples. At the moment, this is challenging owing to variability in technical factors such as batch effects. We appreciate that we must acknowledge that the predictors may not be adequate for all age groups and datasets of different ancestry. To address these points, we have amended the text in Paragraph 3 of the 'Discussion' as follows:

"Distributions of DNAm values and subsequent DNAm-based scores may vary across different methylation datasets. In relation to the biochemical and lifestyle traits, the predictors were generated using an adult sample of individuals with European ancestry. Therefore, it is possible that the predictors may not be generalisable to datasets comprising different age ranges, such as cohorts of children, and individuals with different ancestries. Differences between datasets may result also arise from biological differences, for example cases for a given disease may have altered DNAm values for a number of probes relative to controls, or result from technical or normalisation differences. As a result, DNAm-based scores may vary greatly across datasets and projecting an individual onto a reference sample to view where their DNAm-based score would lie

along the reference sample is therefore challenging. Future work will focus on developing methods which can appropriately account for variability across datasets and allow for a projection of individuals onto disparate DNAm samples or datasets."

Comment 7: It would help to add information in the demo about how to interpret the different modules (e.g. a walkthrough, or pop-up bubbles)

Response: We agree with the reviewer that the demo version of the app should include detailed information on what each panel shows along with guidance on the interpretation of the plots. In addition to the other applications, we have now placed in 'info' action buttons on the sidebar of each panel to show information on what is being shown and how it may be interpreted in addition to key limitations.

Comment 8: On a related note, is there a way to help users gauge how 'reliable' these different scores are? Some DNAm predictors (e.g. smoking) are more reliable and accurate than others, could this be indicated in the platform e.g. by visualizing this in the plots or adding text to aid interpretation (i.e. to what extent could a user interpret the percentiles for different scores with confidence)? I see in page 10 that error bars can be displayed, but (a) I do not see where the option to visualize these is in the online tool and this is also not explained in the manuscript; (b) I am assuming that these error bars are related to the input data, so a bit different from the point of indicating how good the scores themselves are. I assume that this can be partly gauged by the information on correlations in the fourth panel, although this includes only a selection of the DNAm scores.

Response: We agree with the reviewer that information on the performance of the DNAm-based scores must be outlined. Panel 4 is designed to indicate how well DNAm-based scores correlate with phenotypic values of the traits in the large Generation Scotland sample. We have added appropriate text in 'info' buttons in the 'MethylDetectR' application and the demo version to remedy this issue and clarify the accuracy of the predictors in a large sample. Information on how well the predictors perform in the input dataset would require the upload of phenotypic data to the platform which is discouraged given data privacy concerns. The error bars in the plot are misleading in that they reflect interquartile ranges for the percentiles ranks attributed to cases for each predictor – this is clarified in the figure legend and in the 'info' button on the relevant panel. The figure legend of Figure 4B has been amended as follows:

"(B) The median percentile ranks for cases are plotted for selected traits of interest. The number shown in the circle reflects the median percentile, and the interquartile range of percentile ranks for cases is shown with the horizontal lines extending from the circle. Here, the median percentile ranks with respect to a number of physical traits are shown for cases of diabetes."

Users of the 'MethylDetectR' and 'MethylDetectR – Demo' apps are also referred to Panel 4 within the information on the first panel in order to quickly indicate where they can go to view performance indicators of the predictors.

Comment 9: Third panel how should users interpret differences in correlations between their input dataset and the reference GS dataset? What are the implications of these differences for interpreting their results (e.g. does it mean that scores should be interpreted

with more caution or has no bearing on this)? For example, if the correlation between alcohol and smoking DNAm scores in the input dataset is 0.40, but in GS it is only 0.20, what does this mean for the user?

Response: We thank the reviewer for highlighting this. As the reviewer indicates, the correlation structure can have important implications for interpreting how well the predictors perform across the input dataset and Generation Scotland participants. To resolve the concerns raised by the reviewer, we have added information in the application on Panel 3 accessible by clicking an 'info' button. This informs users that if the correlation structure differs greatly from that in Generation Scotland, then it may indicate that the general pattern of DNAm-based scores in their input dataset may differ with those in Generation Scotland. It may indicate that the characteristics of the two datasets differ from one another. Indeed, the performance of predictors in their own dataset may not be comparable to that of Generation Scotland.

Comment 10: In the discussion, the authors state that 'the MethylDetectR platform communicates key messages surrounding the development and present limitations of DNAm-based health profiling to the wider research community and public' - where is this information provided? I cannot see it either in the Demo or Calculate your Scores platforms. This should be provided in the Demo, as the first 'port of entry' users should be using to get familiarized with the platform (and there should be a notice on the Calculate your Scores platform to first check out the Demo, for an example of correct use). This is a good place to walk users through the different functionalities/panels of the platform, and for each panel, to provide key information on interpretation (e.g. using case examples – "here is Joe, he is estimated to be xx years old and score within the xx percentile for yy variable. Based on this, we can conclude that xxx. Limitations to keep in mind are xxx. This information could be used for xxx applications").

Response: We agree with the reviewer and add information in the demo and main applications which can be accessed upon clicking an 'info' action button on each panel. This will allow the user, whatever their background, to immediately become familiarised with the calculation and the limitations of DNAm-based scores. Further, these action buttons provide a walkthrough of each panel. We thank the reviewer for their excellent and very helpful suggestions on how this information should be best portrayed.

Comment 11: Similarly, in the conclusion, the authors state that "our platform provides an important translational tool which communicates state-of-the-art developments in relation to DNAm-based predictors of human traits and health… and highlights the applicability and limitations surrounding such predictors prior to their potential deployment in clinical assessment and management paradigms." While the platform allows to calculate state-of-the-art scores, I cannot see anything related to communication of its uses and limitations.

Response: As above, we have ensured to include this information within the applications themselves to resolve this shortcoming in the previous version of the platform.

Comment 12: Generally, the manuscript refers to a lot of different files and links (e.g. the Demo, the actual platform, all the Zenodo materials). This relies on the idea that users in future will always refer to the manuscript and read it in detail before using the platform. I

would suggest having an easy to find, centralized repository for all of these materials, or alternatively, to add the various links in the Demo/Calculate your Scores platforms (which should be linked somehow).

Response: We agree with the reviewer that the various components of the platform should be easier to access. It was our aim to achieve this using the website. Necessarily, all files must be deposited to Zenodo, and example .csv files are not appropriate for the website given issues surrounding access i.e. text-to-speech conversion on our website. Based on the suggestion of the reviewer, we have included links to all elements of the platform, i.e. the other applications, paper, website and repository, within each application. This will make navigating the platform quick and easy for the user.

Comment 13: The discussion is missing a section on the research and clinical applications of this platform. Can the authors provide concrete examples of where and how this information could be useful?

Response: We thank the reviewer for highlighting this opportunity to improve communication of the potential utility of DNAm-based predictors. In addition to concerns raised by Reviewer 1, we acknowledge that, at present, the clinical applications of the DNAm-based scores are limited given that they cannot faithfully provide accurate estimates at an individual level. As above, we have added the following text in Paragraph 3 of the 'Discussion' section to reflect these concerns:

"Currently, the DNAm-based predictors in 'MethylDetectR' cannot make accurate predictions at an individual level and therefore cannot yet be reliably applied in a diagnostic or forensic context. Highly-accurate DNAm-based scores can aid in research environments as they may provide more accurate information than self-report data (6). For instance, the DNAm-based predictor of smoking can provide a more accurate profile of smoking history than responses from participants in questionnaires. Further, the use of DNAm-based scores for a variety of human traits can proxy many phenotypes such as biochemical and lifestyle traits using a single blood draw. Together, these data can help researchers determine relationships between putative risk factors and important health outcomes, and aid in patient stratification paradigms. Further, 'MethylDetectR' serves as an important translational tool showing an interactive, demo version of the platform and substantial information within each application regarding the interpretation and limitations of DNAm-based predictors. As these predictors become refined, they may be of clinical value. For instance, a blood-based DNAm test was recently developed that could detect five separate types of cancer up to four years before conventional diagnosis. The assay measured circulating tumour DNA methylation and predicted disease in 88% of post-diagnosis patients, with a specificity of 96%."

Minor points

Comment 14: P.3: "In contrast to genetics, these molecular modifications are dynamic, tissue-specific and reversible" – would qualify this sentence a little bit. While it is correct that, overall, DNAm differ largely across tissues, some sites do show strong correspondence across multiple tissues. Similarly, it is unclear whether all DNAm changes are reversible (e.g. some smoking-related DNAm alterations persisting post-cessation).

Response: We agree with the reviewer that this statement is not wholly accurate. As a result we have amended the text in Paragraph 1 of the 'Introduction' section as follows:
"The addition of these chemical tags can alter whether, and to what extent, a gene is active. In contrast to genetics, these molecular modifications are dynamic (3). Further, methylation at many CpG sites is tissue-specific though some show strong concordance across multiple tissues (4). In addition, CpG modifications induced through environmental factors, such as smoking, may be reversible or show persistent alterations (5)."

Comment 15: P. 3: "Furthermore, there has been a rapid increase in the commercialisation and scalability of DNAm assays for direct-to-consumer use or for use in clinical, research or industrial settings." Can the authors add references or include a couple of examples?

Response: To remedy this, we have added the following reference and removed 'rapid' from this sentence:
https://www.nature.com/articles/s41576-020-0215-2

"Furthermore, there has been an increase in the commercialisation and scalability of DNAm assays for direct-to-consumer use or for use in clinical, research or industrial settings (new reference to https://www.nature.com/articles/s41576-020-0215-2)."

Comment 16: P.3: "Blood DNAm data …provides a good index of the overall health status of the body." Can the authors add a reference to support this?

Response: We have amended the text to include the following reference in Paragraph 2 of the 'Introduction' section:

"Blood DNAm data is often used as it is minimally-invasive to collect and it provides a good index of the overall health status of the body (new reference to https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7076122/)".

Comment 17: P. 4: "Lastly, the user can subset the reference samples by sex, age range or case vs. control status." Case-control status on what variables?

Response: Here, we refer to the case-control variables uploaded by the user. To make this clearer, we have amended the text as follows at the end of Paragraph 5 of the 'Introduction' section:
"Lastly, the user can subset the input sample by sex, age range or case vs. control status determined by the case-control variables uploaded by the user. Further, the user can subset the GS reference sample by age and sex."

Comment 18: P. 5: https://www.ed.ac.uk/centre-genomic-medicine/research-groups/marioni-group/methyldetectr the link provided has two attachments, which appear to contain the same information: MethylDetectR Participant Information Sheet and MethylDetectR Participant Consent Statement.

Response: We sincerely thank the reviewer for spotting this error. The documents were accidentally duplicated on the website. This has been remedied to include the correct Participant Consent Statement.

Comment 19: P. 5: in the explanation of methylation risk scores, the authors could point readers to a new review on the topic, including strengths and challenges of this approach (Huls & Czamara, 2020 Epigenetics)1.

Response: We thank the reviewer for highlighting this review, we have included the following text under the section 'Development of a DNAm-based predictor' in 'Methods' to allow for its addition:

"Readers are referred to review on the development of DNAm-based scores and the challenges surrounding their generation (new reference to Huls & Czamara, 2020)."

Comment 20: P. 6: Age predictor - can the authors specify the age range of the discovery data upon which the predictor was based?

Response: We have added this information in the 'Methods' section under the 'Age Predictor' sub-section:
"The age predictor was generated using data from individuals with an age range of 2 to 104 years."

Comment 21: P. 6: Can the author provide a reference/specify how cutoffs were selected for their protein analyses (e.g. for training-test ratio, r and p thresholds etc.).

Response: We have removed the blood protein predictors in the current version of 'MethylDetectR' but will include 109 blood protein predictors in the next update.

Competing Interests: R.E.M has received payment from Illumina for a presentation.

Reviewer Report 22 January 2021

https://doi.org/10.21956/wellcomeopenres.18118.r42039

?

Gemma Sharp
MRC Integrative Epidemiology Unit, Population Health Sciences, University of Bristol, Bristol, UK

This paper describes a very useful and novel platform for generating DNA methylation-based predictors of traits from 450k/EPIC array data. The plots produced by the app are attractive and it is very pleasing to see that all code has been provided so that users can reproduce and adapt these outputs if they wish. I look forward to using the platform and thank and congratulate the authors on their great work.

I have a few more specific comments and some ideas for how the paper and/or platform could

possible be improved:

The paper does a very good job of describing the potential for DNA methylation as an indicator or predictor of human traits and health. The motivation for doing this from a clinical point of view is well-argued.

The need for the platform is clear. It saves the user a lot of time and effort in generating DNAm-based predictors of traits in their own data. While generating a DNAm-based predictor of epigenetic age has been possible for several years using the web-based platform developed by Horvath *et al*., this platform is unique in that it also includes six DNAm-based predictors of lifestyle and biochemical traits (alcohol consumption, body fat percentage, BMI, HDL cholesterol, smoking status and waist-to-hip ratio) and 27 predictors of blood protein levels related to inflammatory or neurological processes/diseases. The lifestyle predictors were generated in a large sample of individuals. The 27 protein predictors were generated in a much smaller sample, but replicated in an independent subset.

The platform consists of two applications, but it was a little unclear to me what these were. The first application is MethylDetectR - Calculate Your Scores. Is the second application "the main MethylDetectR application"?

Users must upload their 450k or EPIC array data to the Calculate Your Scores application. The authors say that no data are stored by the application near the start of the paper. Later, in the Methods section, they provide more information on this and this section settled my concerns about data sharing of individual level data. Given that DNA methylation array data can potentially be used to identify individuals, data managers, legal teams and researchers may be understandably reluctant to upload data to an external website. Although it seems that this is probably a very low risk, it could be work highlighting this helpful section with a specific heading on data protection, and/or moving it to earlier in the paper.

I agree that the utility of DNAm-based predictors lies in population-level research and are inaccurate and therefore of less use in predicting values at the individual level. If this platform can help illustrate that, it will be a useful teaching and learning aid. I suppose there's a concern that users may ignore that warning though and use the platform to make inferences about individuals. Are these caveats explained clearly within the app? It doesn't help that the app appears to provide predictions at the individual level, such as "we predict that you are 60 years old!" and I would strongly suggest changing this. Similarly in the discussion there is a paragraph that outlines the potential for these scores to be used in diagnosis/prognosis and forensics, which would require accurate prediction at the individual level. I suggest that this paragraph includes a caution that this is not possible using the scores generated by MethylDetectR. The issue with differences in distribution across datasets is explained well in the next paragraph, but I think the point (that MethylDetectR scores cannot be used to make predictions at the individual level) needs making explicitly.

The upload limit is 3gb but files greater than 500mb "may take a considerable amount of time to upload". Is there any way to prepare files that could reduce this limitation? Does the website require data for every CpG on the array or just a subset that are useful in generating the scores? Should/could files be compressed?

I was glad to see that an R script is provided for users who do not wish to/cannot upload DNAm data. The link provided goes to a repository of code and data. Would it be more appropriate/possible to release this as an R package?

The demo version of the app is useful - please could it be mentioned nearer the start of the paper?

The interface for the app itself would benefit from extra information being included on how files should be formatted, file size restrictions, etc. This would be more helpful for users than having to refer back to this paper or other instructions. Similarly, the plots generated in the MethylDetectR app are useful but some information on each tab explaining what this plot is showing is needed.

Do the authors plan to update the app to include predictors for more traits as and when they are discovered?

A small point about Figure 1 and Figure 2: I don't find that these figures add anything to the paper. For example, Figure 2 could be more clearly presented as a bulleted list, no extra information is conveyed by having the information in circles (and all information is already provided in the main text).

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?
Partly

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: Epigenetic epidemiology, R, Shiny.

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, however I have significant reservations, as outlined above.

Author Response 30 Mar 2021

Robert Hillary, University of Edinburgh, Edinburgh, UK

Comment 1: The need for the platform is clear. It saves the user a lot of time and effort in generating DNAm-based predictors of traits in their own data. While generating a DNAm-based predictor of epigenetic age has been possible for several years using the web-based platform developed by Horvath *et al*., this platform is unique in that it also includes six DNAm-based predictors of lifestyle and biochemical traits (alcohol consumption, body fat percentage, BMI, HDL cholesterol, smoking status and waist-to-hip ratio) and 27 predictors of blood protein levels related to inflammatory or neurological processes/diseases. The lifestyle predictors were generated in a large sample of individuals. The 27 protein predictors were generated in a much smaller sample, but replicated in an independent subset.

Response: We thank the reviewer for their encouraging comments regarding this platform. We take this opportunity to highlight that our updated version of 'MethylDetectR' temporarily omits protein predictors from the dataset. The reason for this is that we are generating, and aim to publish, predictors for a larger set 109 blood protein levels including refinements in the pipeline used to generate the protein predictors. We are committed to updating and managing 'MethylDetectR'. We will update 'MethylDetectR' every three months. We will include these 109 new predictors when they become available, likely in the next update. Owing to the temporary omission of blood protein predictors, references to these predictors in the text and figures have been removed. The manuscript has been updated accordingly throughout.

Comment 2: The platform consists of two applications, but it was a little unclear to me what these were. The first application is MethylDetectR - Calculate Your Scores. Is the second application "the main MethylDetectR application"?

Response: We thank the reviewer for highlighting this issue and agree that this must be clarified. The reviewer is correct, the first application is 'MethylDetectR – Calculate Your Scores'. However, the use of this application is optional as users may wish to download the scripts to generate their own scores. Alternatively, the user may have to use the provided scripts if their DNAm file is too large for upload. Though, we now include an additional .csv file 'Truncate_to_these_CpGs.csv' to allow users to subset their DNAm file to those CpG sites used in the 'MethylDetectR – Calculate Your Scores' application. This all results in the main 'MethylDetectR' application being the 'second' application in the context of the entire platform and its pipeline. We have clarified this in the main text with the addition of the following text:

"Briefly, the 'MethylDetectR' platform consists of two applications. The first application named 'MethylDetectR – Calculate Your Scores' allows users to securely upload Illumina 450k or EPIC DNAm array data and obtain blood-based methylation predicted scores (or values) for the aforementioned traits. No data are stored by the application. Furthermore, predicted scores are often returned on arbitrary scales. The use of this application is optional as users may instead use R scripts which we have made publicly available if they so wish or if their DNAm files are too large for upload to the online application (>3 GB) (https://doi.org/10.5281/zenodo.4646300). However, users can access a file called 'Truncate_to_these_CpGs.csv' to subset the list of CpG sites in their DNAm files to those required by the 'MethylDetectR – Calculate Your Scores' application. This

should reduce file size and upload time.
The second and main application named 'MethylDetectR' allows users to compare DNAm-derived scores for any individual in their input dataset against other individuals in the input dataset. Percentile ranks for individuals in the input dataset may be downloaded."

Comment 3: Users must upload their 450k or EPIC array data to the Calculate Your Scores application. The authors say that no data are stored by the application near the start of the paper. Later, in the Methods section, they provide more information on this and this section settled my concerns about data sharing of individual level data. Given that DNA methylation array data can potentially be used to identify individuals, data managers, legal teams and researchers may be understandably reluctant to upload data to an external website. Although it seems that this is probably a very low risk, it could be work highlighting this helpful section with a specific heading on data protection, and/or moving it to earlier in the paper.

Response: We strongly agree with the reviewer that data protection and privacy concerns are of utmost importance in relation to this platform. We also agree that a dedicated section should be placed earlier in the manuscript discussing data protection and privacy aspects of the platform. As a result, we have added the following text at the top of the 'Methods' section giving data protection and privacy aspects their own dedicated section to appropriately reflect its importance:

*"**Data Protection and Privacy**. No data are stored in 'MethylDetectR' and are deleted upon closing the applications. Applications are also timed out after three minutes of inactivity and are hosted on patched and secure servers within the Institute of Genetics and Cancer, University of Edinburgh. This research and translational tool complies with GDPR guidelines and has been designed to ensure the highest level of data security and privacy. 'MethylDetectR' and information on its usage are also available at the following website: https://www.ed.ac.uk/centre-genomic-medicine/research-groups/marioni-group/methyldetectr. Information relating to participant consent is also available at this website. Given that no data are stored, this information pertains to general risk surrounding the upload of biological data to online software and the measures taken to mitigate the risk of motivated intruders gaining access to such data."*

Comment 4: I agree that the utility of DNAm-based predictors lies in population-level research and are inaccurate and therefore of less use in predicting values at the individual level. If this platform can help illustrate that, it will be a useful teaching and learning aid. I suppose there's a concern that users may ignore that warning though and use the platform to make inferences about individuals. Are these caveats explained clearly within the app? It doesn't help that the app appears to provide predictions at the individual level, such as "we predict that you are 60 years old!" and I would strongly suggest changing this. Similarly, in the discussion, there is a paragraph that outlines the potential for these scores to be used in diagnosis/prognosis and forensics, which would require accurate prediction at the individual level. I suggest that this paragraph includes a caution that this is not possible using the scores generated by MethylDetectR. The issue with differences in distribution across datasets is explained well in the next paragraph, but I think the point (that MethylDetectR scores cannot be used to make predictions at the individual level) needs

making explicitly.

Response: We agree with the reviewer that the platform provides an opportunity to communicate the message that DNAm-based predictors work well at a population level, but cannot make accurate predictions at an individual level. This has major implications for their use in clinical settings. We, therefore, take the following actions. First, we change the text in the first panel of 'MethylDetectR' and 'MethylDetectR – Demo' to include the following message: ***"Epigenetic Clock (Zhang) age estimate: X years*"*** *and include underneath* ***"*All methylation-based predictors in 'MethylDetectR' can make inaccurate predictions at an individual level, limiting clinical utility. However, they can work well at the population level and will improve further with more refined prediction methods and larger-scale studies".*** *This ensures that the message is communicated explicitly at the first instance of the application.*

Second, we amend text in Paragraph 2 of the 'Discussion' section to mention that DNAm-based scores in 'MethylDetectR' cannot be used in diagnosis/prognosis and forensics:

"Currently, the DNAm-based predictors in 'MethylDetectR' cannot make consistently accurate predictions at an individual level and therefore cannot be reliably applied in a diagnostic or forensic context."

Third, we remove the following text from the same paragraph:

"DNAm-based predictors of human traits, such as chronological age and body mass index, may also aid in forensic contexts."

Comment 5: The upload limit is 3gb but files greater than 500mb "may take a considerable amount of time to upload". Is there any way to prepare files that could reduce this limitation? Does the website require data for every CpG on the array or just a subset that are useful in generating the scores? Should/could files be compressed?

Response: We recognise this as a current limitation of the software. To resolve this issue, we provide a list of CpG sites ('Truncate_to_these_CpGs.csv') which the user can use to truncate their DNAm dataset prior to upload. This is included in the Zenodo repository and also clearly detailed in the 'Press Here to Format Files and Useful Links' action button in the 'Calculate Your Scores' application.

Comment 6: I was glad to see that an R script is provided for users who do not wish to/cannot upload DNAm data. The link provided goes to a repository of code and data. Would it be more appropriate/possible to release this as an R package?

Response: We thank the reviewer for this excellent suggestion. At the moment, we feel that the provided code and applications achieve a similar goal as an R package. Nonetheless, in the future, we aim to refine the platform by developing methods that can allow for comparisons of DNAm-based scores across different cohorts. Currently, this is challenging owing to variability in technical factors across different DNAm datasets, such as batch effects. As a result, the application allows only for the calculation of DNAm-based scores within the input dataset. Furthermore, we hope to release an R package that can allow for the projection of DNAm-based

scores onto other cohorts, such as those in Generation Scotland and allow for visual comparisons between distributions of scores in the input cohort and those within Generation Scotland.

Comment 7: The demo version of the app is useful - please could it be mentioned nearer the start of the paper?

Response: The demo version of the app is now mentioned at the start of the paper in Paragraph 4 of the 'Introduction' section. It was previously mentioned in the 'Methods' section.

Comment 8: The interface for the app itself would benefit from extra information being included on how files should be formatted, file size restrictions, etc. This would be more helpful for users than having to refer back to this paper or other instructions. Similarly, the plots generated in the MethylDetectR app are useful but some information on each tab explaining what this plot is showing is needed.

Response: We thank the reviewer for suggesting this useful addition to the platform. We have now included 'info' action buttons on each panel of 'MethylDetectR', 'MethylDetectR – Demo' and on 'MethylDetectR – Calculate Your Scores'. These action buttons provide information on what the panels show and how to format files. We also include links to the main website and links to the other apps so navigation between the different components of the platform is simplified.

Comment 9: Do the authors plan to update the app to include predictors for more traits as and when they are discovered?

Response: Yes, we will continue to update the applications and platform to include a larger set of DNAm-based predictors of protein levels. We will add in predictors of 109 blood proteins and include robust DNAm-based predictors of human traits generated in our group and others. We will update 'MethylDetectR' every three months. To clarify this, we have added a 'Version Control' paragraph to the 'Implementation' subsection of 'Methods':

*"**Version Control.** We will update 'MethylDetectR' every three months to include new DNAm-based predictors of human traits as they are generated by our own group and others. Updates will be managed by Robert F. Hillary or Riccardo E. Marioni. If researchers wish to have their predictors considered for inclusion in 'MethylDetectR', please use the corresponding author email address in this manuscript or the contact details available at the 'MethylDetectR' website (https://www.ed.ac.uk/centre-genomic-medicine/research-groups/marioni-group/methyldetectr). The current and historical versions of 'MethylDetectR' will be available in the Zenodo repository, updated versions will also be made available in this repository (https://doi.org/10.5281/zenodo.4646300)".*

We have also added the following text to Paragraph 1 of the 'Discussion':

"We will continue to update 'MethylDetectR' every three months with the aim of including new DNAm-based predictors of human traits when they come available".

Comment 10: A small point about Figure 1 and Figure 2: I don't find that these figures add

anything to the paper. For example, Figure 2 could be more clearly presented as a bulleted list, no extra information is conveyed by having the information in circles (and all information is already provided in the main text).

Response: We agree with the reviewer that these figures are not essential for the paper or the platform. As a result, we have removed them and updated figure annotations accordingly.

Competing Interests: R.E.M has received payment from Illumina for a presentation.