MethylDetectR – Reference Document

Background

Thank you for your interest in using 'MethylDetectR'. This is an online platform which allows users to obtain estimated values or scores for a variety of human traits including age, body mass index and alcohol consumption based off of their DNA methylation data (1, 2). DNA methylation is a naturally occurring biological process within our cells in which chemical tags (called methyl groups) are added to DNA. The areas in DNA molecules in which these tags are added are called 'CpG sites'. These chemical tags can determine whether a gene is switched on or off – which may have harmful or beneficial effects depending on the normal function of a given gene. Any given person's DNA methylation profile results from a combination of their genetics and their environment. This makes estimating human traits and health from DNA methylation data particularly exciting as one's estimated risk profile may change from one time point to another.

The estimators or 'predictors' of human traits in this platform come from state-of-the-art machine learning methodologies. Typically, we begin by examining around ~27,000, ~450,000 or ~800,000 CpG sites across the genome depending on which technology is used to measure DNA methylation in individuals. Methylation levels are often reported between 0 to 100%. In an individual, a methylation level of 50% means that 50% of their cells or DNA molecules show methylation at that CpG site. We then select a trait of interest, such as body mass index. We take each CpG site in turn and correlate the average methylation level of that CpG site across individuals with our trait of interest. These individuals constitute the training sample. As an example, a researcher might find that 200 CpG sites correlate strongly with body mass index. The strength of the correlation provides a weight for that CpG site. In a separate group of individuals (the test sample), these weights can be applied to the same CpG sites in order to estimate body mass indices of individuals in the test sample. For instance, CpG_1, CpG_2 and CpG_3 may have weightings of 0.2, 0.4 and 0.6, respectively. The methylation levels in a given individual for these CpG sites might be 20%, 40% and 60%. To derive a methylation-based predictor of body mass index, the CpG sites are multiplied by their associated weights until the last CpG site is reached. The sum of these products will return a methylation-based score for body mass index. This often returns an arbitrary value which is not on a typical scale of body mass index. Therefore, a statistical transformation may be applied to return the score to a value of body mass index or alternatively, we can compare how an individual's score relates to scores of other individuals to give a sense of how their predicted body mass index compares to the wider population.

The comparison to other individuals allows for the communication of a key message: the predictors can work well on a population level but do not necessarily hold true on an individual level. For example, an epigenetic age predictor may show a strong correlation with actual age across the entire sample but for a given individual, predicted age may be incorrect by a number of years or even decades. Predictors will become more accurate with larger study sizes and increasingly sophisticated statistical tools.

Calculate Your Scores

Using the App

Briefly, the 'MethylDetectR' platform consists of two applications. The first application, named 'MethylDetectR – Calculate Your Scores', allows users to securely upload blood DNA methylation data and obtain DNAm-based predicted scores (or values) for several human traits, such as smoking behaviour and body mass index. DNA methylation data should be uploaded as an .rds file. We recommend that files of no larger than 500 MB are uploaded in order to allow for fast calculation, however, the software can deal with uploads of up to 3 GB. To make files smaller prior to upload and reduce upload time, users can use a file called 'Truncate_to_these_CpGs.csv' available in our Zenodo repository (https://doi.org/10.5281/zenodo.4646300). This allows you to subset CpG sites in your methylation file to those used by the 'MethylDetectR – Calculate Your Scores' application. A 'SexAgeInfo' file as a .csv file may also be included if the user wishes. This file should 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 another column may list the 'true age' of individuals ('Age' column). This step is useful given that users can subset the input dataset and Generation Scotland by sex in the main 'MethylDetectR' application. Furthermore, if the user chooses to upload true ages of individuals, then the app will use these data to subset the sample according to the age slider. However, if the user does not wish to include this information and leave the 'Age' column in 'SexAgeInfo' blank, then the app will use DNAm-based predicted age for subsetting the sample. If some individuals have missing true ages and others have this information included, then true ages will be used for those who have such data and epigenetic age will be used for those without. Example input files are available at https://doi.org/10.5281/zenodo.4646300. For your own convenience in saving your methylation object, we recommend 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 for the user:

- 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 which 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 from the original training sample. 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.

Link to App: https://shiny.igmm.ed.ac.uk/Calculate_Your_Scores/

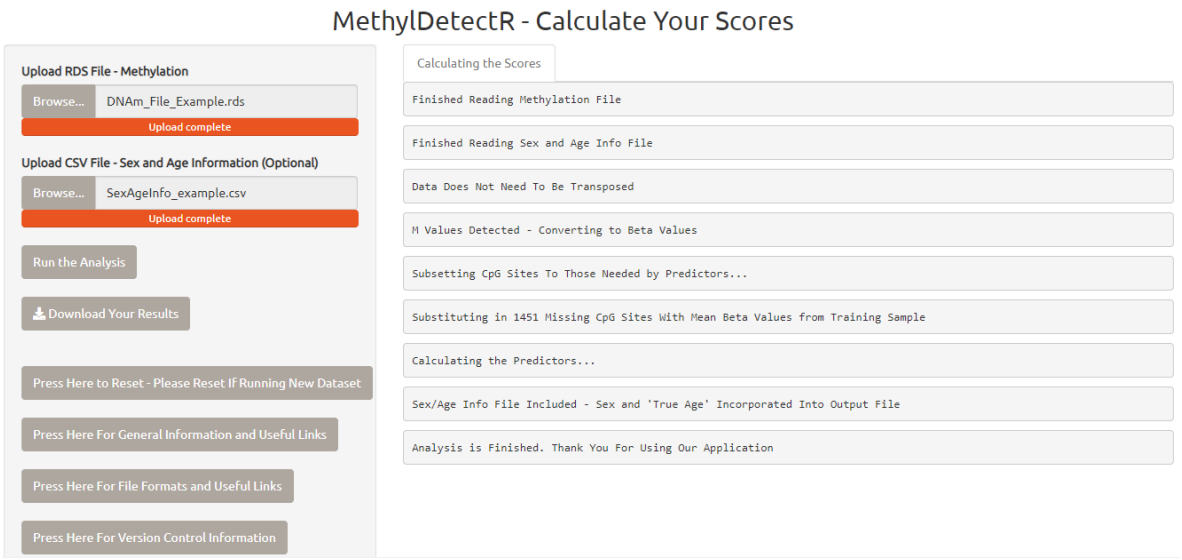

Figure 1. Example of the 'MethylDetectR - Calculate Your Scores' application.

Generating the Predictors without the App

If your file is too large or if you do not wish to upload your dataset to the online application, an R script is provided which also automatically generates the scores. All the user needs to do is call their methylation object 'data' and include an optional 'SexAgeInfo.csv' file if they so wish. This object should be called 'sexageinfo'. The user can download the script and the associated file with predictor information from https://doi.org/10.5281/zenodo.4646300. A reference output file is also available at https://doi.org/10.5281/zenodo.4646300.

MethylDetectR

The main 'MethylDetectR' application has four panels. 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. 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 first panel allows you to select any predictor and view how a selected individual in the input dataset ranks against the remainder of the input dataset (in pink; Figure 2A). Alternatively, if the user uploads an optional file with binary phenotype information, then users may also view how an individual compares against controls (in pink) and cases (in blue) for a trait of interest (Figure 2B). This optional file should contain an 'ID' column with IDs for individuals as in the methylation file and any binary traits of interest for the remainder of the columns with controls coded as '0' and cases as '1'. 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. It is also possible to download

the percentile ranks for every individual in the input dataset, against all other individuals in the dataset, for each trait.

Link to App: https://shiny.igmm.ed.ac.uk/MethylDetectR/

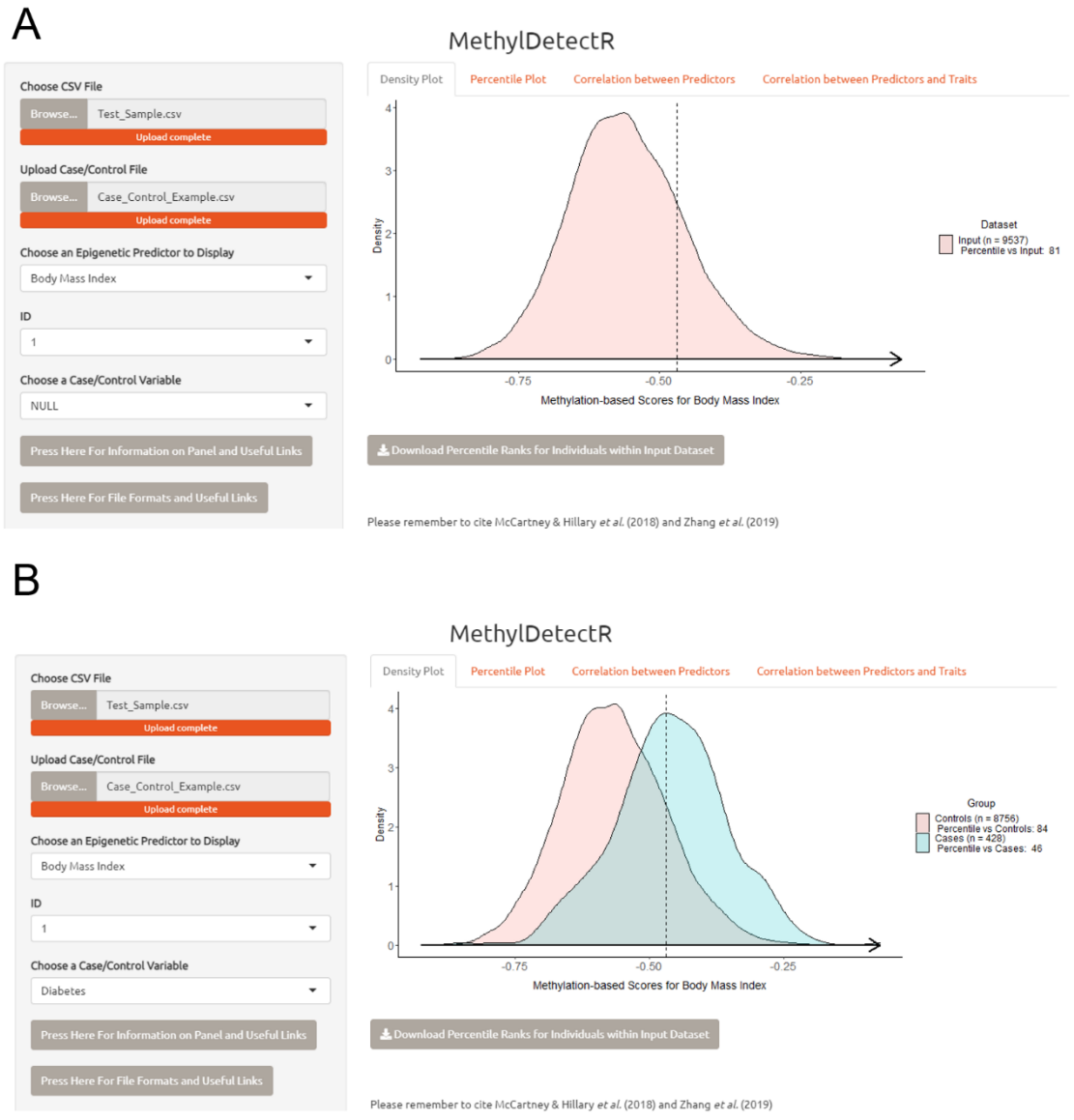

Figure 2. Panel 1 (A) Percentile rank for selected individual against remainder of input dataset is shown. (B) Distribution of selected DNAm-based scores is split according to cases and controls for a chosen binary trait of interest. This binary trait is uploaded optionally and separately from the scores. Percentile ranks against cases and controls are plotted.

The second panel allows users to select multiple traits and simultaneously view the percentile ranks for a given individual against the remainder of the input dataset (Figure 3A). Again, the user can use the sidebar functionality to subset by age range and sex. The user can also choose to view how percentile ranks differ among cases and controls for a selected binary phenotype. The median percentile for cases and interquartile range are plotted for any selected traits of interest (Figure 3B).

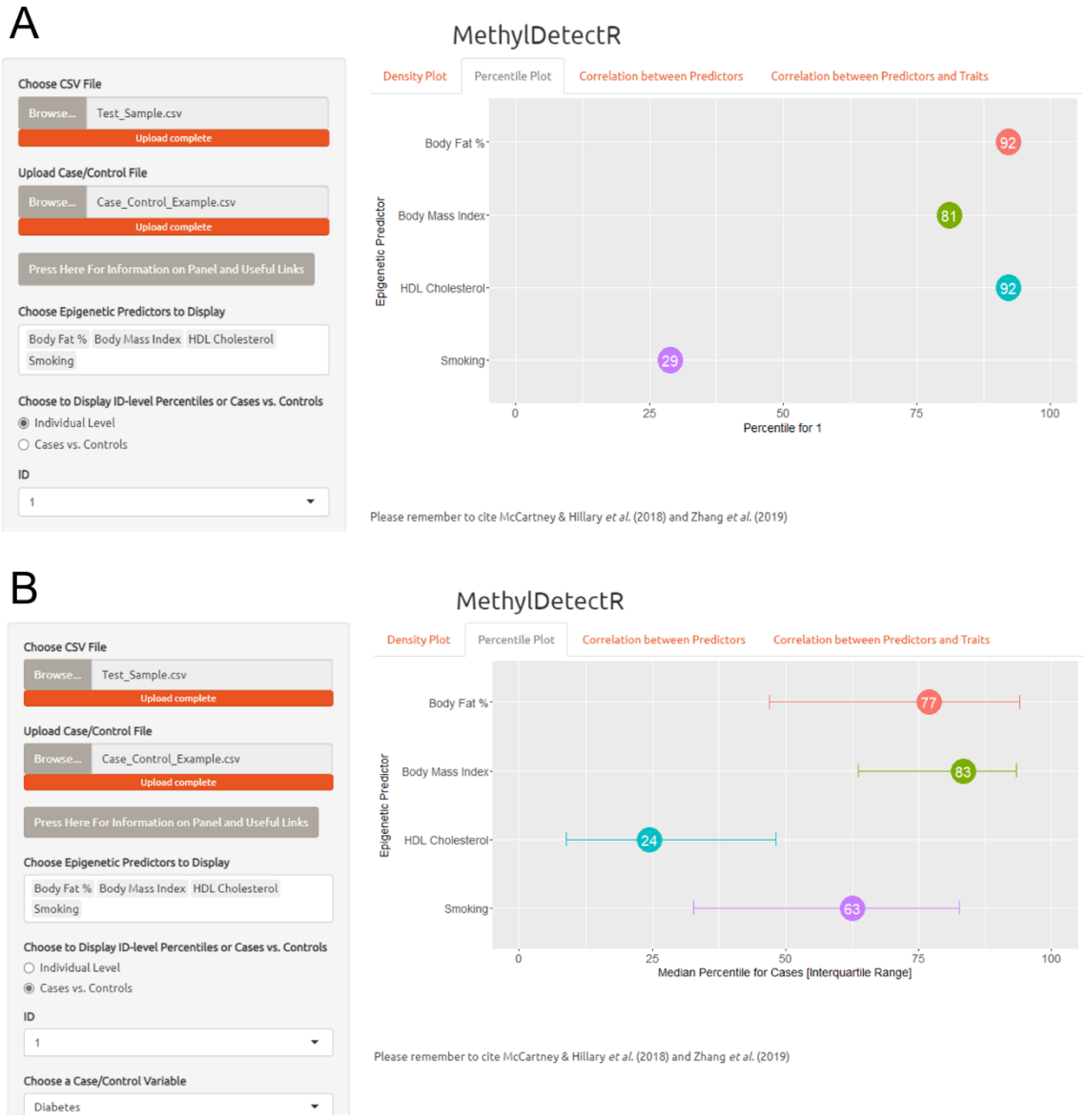

Figure 3. Panel 2 (A) Percentile ranks are plotted simultaneously for selected individual when compared to remainder of input dataset in relation to a number of traits. (B) Median percentile rank for cases are plotted for a number of traits along with interquartile range (horizontal bars).

In the third panel, the user can select multiple methylation-based predictors and view how they correlate with one another in order to get a sense of their interrelationships and better understand the presented results

(Figure 4A). This is possible for both the input dataset and a reference sample of 4,450 individuals, and may be subset by age and sex. This reference sample represents the Generation Scotland study – a Scottish family based health study with rich health, genetic and methylation data (https://www.ed.ac.uk/generation-scotland). It is one of the largest methylation data resources in the world making it well positioned for the implementation of this application. Users can also subset the input dataset by cases or controls, or choose to plot the correlation data for cases and controls simultaneously according to a binary phenotype of interest which they have uploaded (Figure 4B).

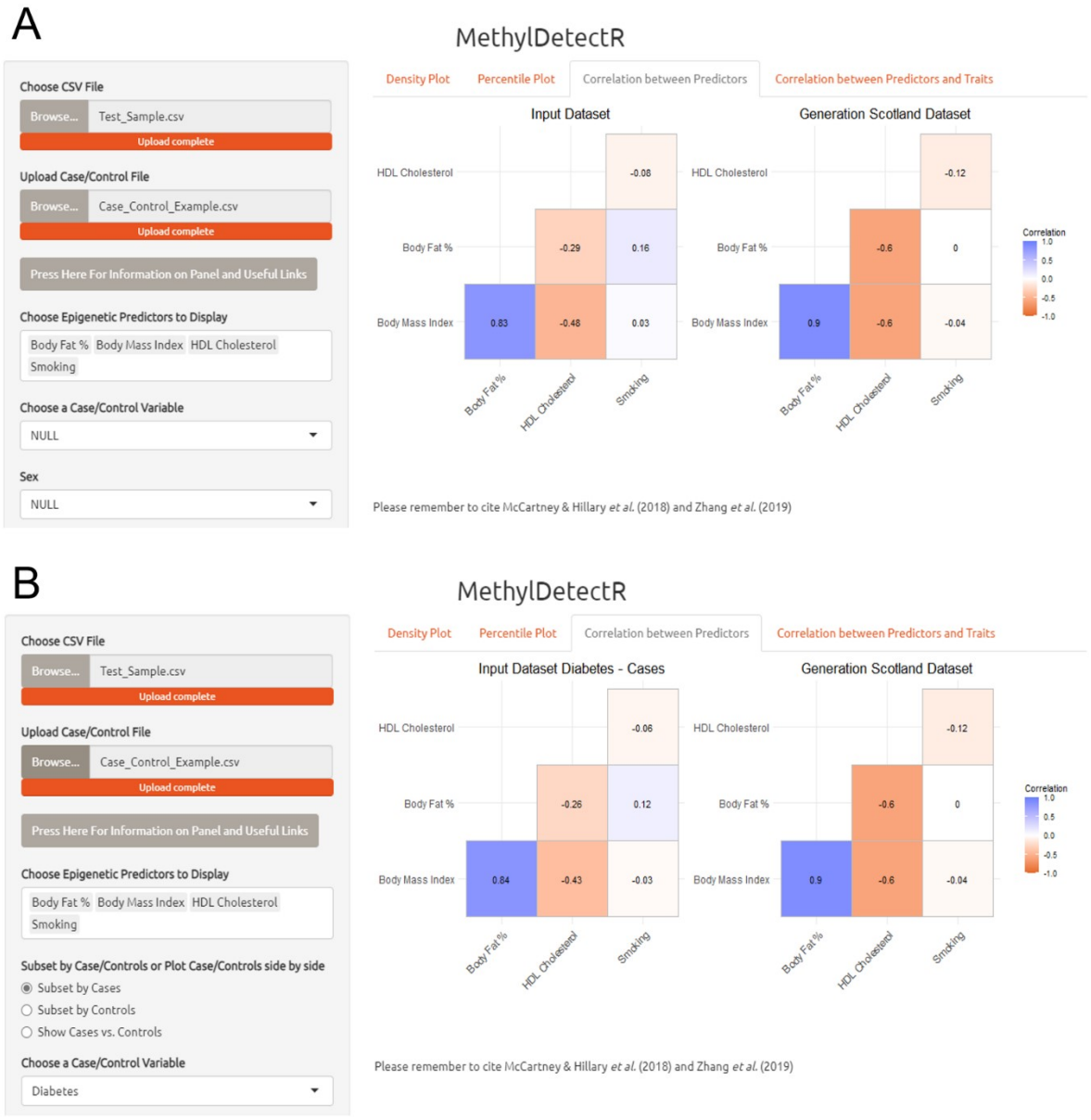

Figure 4. Panel 3 (A) Users can select multiple predictors of interest and simultaneously view the interrelationships between these variables of interest in both the input dataset and a reference sample - Generation Scotland (n =

4,450). (B) The user can choose to subset the input dataset by cases or controls, or choose to visualise the input cases and controls side by side according to a binary phenotype of interest they have uploaded.

In the fourth panel, users can view how well the methylation-based age, lifestyle and biochemical predictors correspond with phenotypic values of their respective traits in Generation Scotland (Figure 5). In this final panel, users can also subset by age and sex to see how the performance of methylation-based predictors vary according to the truncated reference dataset.

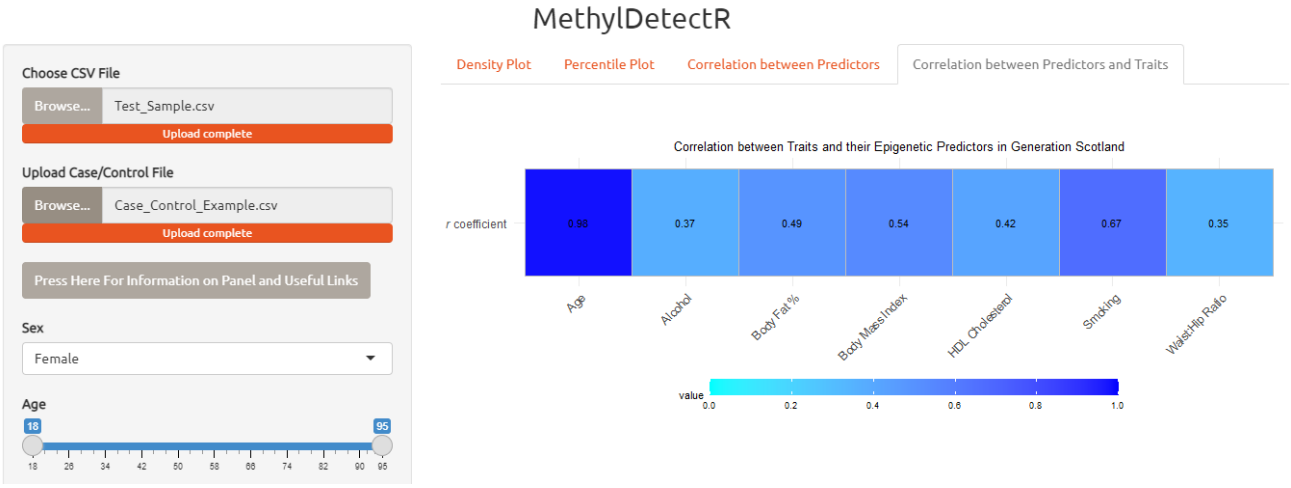

Figure 5. Panel 4 Correlation between methylation-based scores for epigenetic age and lifestyle traits and actual values for the respective traits in Generation Scotland.

Data Privacy and Protection

Please refer to the 'Participant Information Sheet' and 'Participant Consent Statement' documents on the main website. No data are stored or collected. The documents are present to describe the general risk surrounding upload of biological data to online software and the measures taken to mitigate such risks. A full Data Protection Impact Assessment has been conducted for this translational and research tool and was approved by the University of Edinburgh data controller. The design and implementation of the applications are in full accordance with GDPR guidelines. The applications have been designed to ensure maximum possible data privacy and protection. The applications are also hosted on a secure, encrypted server at the Institute of Genetics and Cancer, University of Edinburgh. If you have any further data privacy or security concerns, please contact either robert.hillary@ed.ac.uk or riccardo.marioni@ed.ac.uk.

Contact Details

Robert Hillary – robert.hillary@ed.ac.uk

Riccardo Marioni – riccardo.marioni@ed.ac.uk

Generation Scotland website: https://www.ed.ac.uk/generation-scotland

References

1. McCartney DL, Hillary RF, Stevenson AJ, Ritchie SJ, Walker RM, Zhang Q, et al. Epigenetic prediction of complex traits and death. Genome biology. 2018;19(1):136.
2. Zhang Q, Vallerga CL, Walker RM, Lin T, Henders AK, Montgomery GW, et al. Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. Genome medicine. 2019;11(1):54.