**ENIGMA-Epigenetics Working Group**

**Protocol for cortical EWAS**

25 August 2023

The following codes were written by Xinyang Yu and Sylvane Desrivières, with revisions and comments provided by Antoine Weihs. If you have any questions related to these analyses, please feel free to contact us at sylvane.desrivieres@kcl.ac.uk and xinyang.1.yu@kcl.ac.uk.

The following scripts includes 2 sections: methylation data quality check (section 1); epigenome-wide association analysis (EWAS) of cortical structures (section 2).

**Files required for these analyses are:**

**Methylation data** – The following data, generated by ENIGMA-Epigenetics QC protocol, will be used:

"./Quan-norm.rda" (Beta values after quantile normalization)

"./fast\_svd.rda" (Principal components of beta values)

"./cellcount.rda" (Estimated cell type proportion)

"./RGset.rda" (Raw Dataset after QC)

**Cortical measures** – These files will be used for conducting EWAS with cortical measures. If possible, please use the cortical outputs processed with FreeSurfer version 5.3.0. Otherwise, please provide the version of FreeSurfer software that was used to process the data along with the results. Assuming that the FreeSurfer output is available, the files should be named according to the following examples and adhere to the specified format.

* For cortical thickness, which should include 34 ROIs based on Desikan-Killiany atlas, please name the files accordingly and ensure they follow the following format.
  + “lh.aparc.thickness.csv”
  + “rh.aparc.thickness.csv”

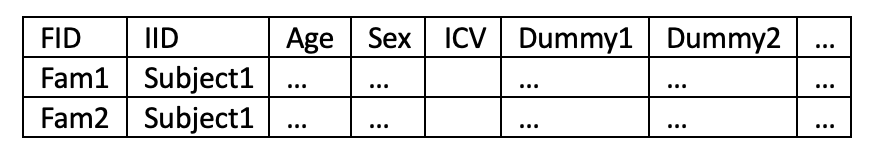
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SubjectID | lh\_bankssts\_thickness | lh\_caudalanteriorcingulate\_thickness | lh\_caudalmiddlefrontal\_thickness | lh\_cuneus\_thickness | lh\_entorhinal\_thickness | … |
| Subject1 | … | … | … | … | … | … |
| Subject2 | … | … | … | … | … | … |

* For cortical surface area, which should include 34 ROIs based on Desikan-Killiany atlas, please name the files accordingly and ensure they follow the following format.
  + “lh.aparc.area.csv”
  + “rh.aparc.area.csv”

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SubjectID | lh\_bankssts\_area | lh\_caudalanteriorcingulate\_area | lh\_caudalmiddlefrontal\_area | lh\_cuneus\_area | lh\_entorhinal\_area | … |
| Subject1 | … | … | … | … | … | … |
| Subject2 | … | … | … | … | … | … |

**Covariates file** – The same files used for ENIGMA2 will also be used. This comma-delimited (.csv) text file should contain the following columns: SubjID, Age, Sex (Male=1 and Female=2), and ICV. Additional columns for dummy covariates (i.e., a covariate to control for different acquisitions sites, if applicable) are optional. In addition, the first 4 principal components of the beta value (i.e., from "**./fast\_svd.rda**"), and the first two components of estimated cell-type proportion (i.e., from "**./cellcount.rda**"), will also be included as control variables. The relevant code of combining data will be provided.

* If your cohort has only healthy controls, only patients, and twin study, the final covariates file should be named “**CorticalCovariates.csv**”, and have the following format. The values for ICV (total intracranial volume) can be extracted from the “EstimatedTotalIntraCranialVol” in the FreeSurfer output.



* If your cohort includes both patients and healthy controls, please include a covariate called "**AffectionStatus**" coded as a binary indicator variable, where Controls = 0 and Patients = 1. The final file, saved as “**CorticalCovariates.csv**”, should have the following column at a minimum: subjID, Age, Sex, ICV, and AffectionStatus. Additional columns for dummy covariates (i.e., a covariate to control for different acquisitions sites, if applicable) are optional. Note that we will have five outputs for the whole sample, the female individuals only, the male individuals only, the case individuals only and the control individuals only.

**Notes**

* Please make sure that missing data has been recoded as NA in these files!! Self-coded missing data should be transformed to NA through Data== -9 <- NA, where Data should be replaced with the name of your data file in question and -9 should be replaced with your self-coded missing data value.
* Sex must be specified as follows: (Males=1, Females=2), and "FID" and "IID" should be named exactly the same in all files.
* If population stratification is of reasonable concern, the self-reported ethnicity information should be included in the covariates file, i.e., as dummy variables. As an alternative, the first 4 MDS of genetic data can also be included.

**Output files that need to be sent to us are as follows:**

For case-control studies, six RData files (i.e., corresponding to outputs from the full sample with and without controlling for ICV, female sample, male sample, case only and controls only) for EWAS results of cortical measures should be sent to us.

* An example of these six files:

A screenshot of a phone

Description automatically generated

For population-based cohorts, four RData files, (i.e., corresponding to outputs from the full sample with and without controlling for ICV, female sample, and male sample) for EWAS results of cortical measures should be sent to us.

An example of these four files:

A screenshot of a phone

Description automatically generated

All RData files should include beta, SE, and P-values for EWAS, cohort name, age information of your cohort, and a list of invariant probes.

**Other information required**: information about sample size and covariates included in the above analyses.

**Section 1: Methylation data QC**

After performing individual-level methylation QC as described in the document "ENIGMA-Epigenetics\_Updated\_DNA Methylation QC\_final.docx", additional QC steps for individual probes will be conducted using the script "QC\_and\_reformat\_methylation\_data.R".

**Section 2: EWAS analysis**

First, we need to obtain summary statistics for brain MRI phenotypes. If you do not already have QCed FreeSurfer outputs at your sites, please refer to <https://github.com/ENIGMA-git/ENIGMA-FreeSurfer-protocol> to obtain your QCed brain phenotypes.

* To generate CSV files from original FreeSurfer outputs, use the script "get\_CT\_and\_SA\_from\_FreeSurfer.sh".
* To get summary measures of MRI phenotypes (such as average frontal thickness), please run "prep\_cortical\_measures.R".

Following these preparations, proceed with the EWAS analysis using "perform\_EWAS.R".

After performing the analyses, please contact [Xinyang.1.yu@kcl.ac.uk](mailto:Xinyang.1.yu@kcl.ac.uk) obtain a OneDrive link for uploading your results, or provide a link from which we can download your results if that method is preferable.

If you encounter any issues while running the scripts, please don’t hesitate to reach out to us at [Xinyang.1.yu@kcl.ac.uk](mailto:Xinyang.1.yu@kcl.ac.uk) and [sylvane.desrivieres@kcl.ac.uk](mailto:sylvane.desrivieres@kcl.ac.uk).

Additionally, after performing the analyses, please fill out this form (<https://forms.gle/7sZSSVj8gWft1WZL7>) to provide information about your cohort for publication purposes, including your author and funding details. You are encouraged to complete this form at your earliest convenience. We will send further reminders when we need this information before our submission.

Thank you very much for your contribution!