Analysis plan: Epigenome-wide Association Study of Aggressive Behavior

Contact
Jenny van Dongen (j.van.dongen@vu.nl), Dorret Boomsma (di.boomsma@vu.nl), Meike Bartels (m.bartels@vu.nl)

Background and goal of this project

As part of the ACTION project on aggression (http://www.action-euproject.eu/), we previously conducted an epigenome-wide association study (EWAS) of aggressive behavior in adults based on whole blood Illumina 450k methylation data collected by the Netherlands Twin Register (NTR). This study found suggestive evidence for associations, with p-values just below the genome-wide significance threshold. We now propose to perform an epigenome-wide association study (EWAS) meta-analysis of aggressive behavior. We would like to ask each cohort to perform the EWAS analysis on their cohort and to provide the results (summary statistics) to us for the meta-analysis.

In addition to the EWAS approach, we aim to test the relationship between aggressive behavior and the epigenetic clock (DNA methylation age, DNAmAge). We hypothesize that because aggressive behavior is associated with adverse life conditions in general, a higher level of aggressive behavior is associated with accelerated epigenetic ageing.

Because participating cohorts include samples with childhood and adult aggression measures and vary with respect to the moment of DNA collection from just after birth till adulthood, we will be able to examine the association between methylation and aggression across the lifespan.

This analysis plan aims to coordinate the analysis to be performed by the analyst of each cohort, and to harmonize the output file format (summary statistics). For your convenience we also include example scripts.

Timeline

We would like to ask you to send us your EWAS results by March 14. Depending on the meta-analysis results, we plan to perform secondary analyses for top hits, including analyses to test if DNA methylation level is associated with gene expression level (RNA-seq and/or Affymetrix U219 array). If you have gene expression data and would like to contribute to the secondary analysis stage, or if you have any ideas for interesting follow-up analysis, please let us know.

Help

Please let us know if there is anything we can do to help you, or if you have any comments or suggestions, by e-mailing us (*please cc both contact persons in your email*):
Jenny van Dongen (j.van.dongen@vu.nl)
Meike Bartels (m.bartels@vu.nl)

Multiple measures of aggression or DNA methylation?

If you have data on aggressive behavior available from multiple measurement instruments or collected at different ages, or if you have DNA methylation data available from samples collected at multiple ages or from multiple tissues, please let us know so we can discuss the optimal analysis strategy.

Analysis Plan – Summary

Methylation data	Illumina 450k array, beta-values, after QC and normalization
Tissue	We assume that you extracted DNA from peripheral whole blood
Phenotype	Aggressive behavior score, on a continuous scale (higher score should correspond to a higher level of aggressive behavior)
EWAS analysis	For each methylation site, we test if DNA methylation level is associated with aggressive behavior score. Methylation level is the outcome, aggression is predictor. This test is repeated for all genome-wide methylation sites (that survived cohort-specific methylation data QC).
Covariates - EWAS model 1	Age, sex, white blood cell counts, technical and cohort-specific covariates
Covariates - EWAS model 2	Covariates model 1 + BMI and smoking status
EWAS results files	Aggression_Cohort_Model1_YYYYMMDD_InitialsAnalyst.csv: [5 columns] Aggression_Cohort_Model2_YYYYMMDD_InitialsAnalyst.csv: [8 columns] Requested information: cgid, N, Beta, SE, Pval
DNAmAge results file	Aggression_CohortName_DNAmAge _YYYYMMDD_InitialsAnalyst.csv Requested information: Measure, N, Beta, SE, Pval [5 columns]
Cohort information file	Aggression_Cohort_CohortInformation_YYYYMMDD_InitialsAnalyst.xls

Analysis Plan – Details

1. Before the analysis: input data for the EWAS and Quality Control

DNA methylation

We will use genome-wide DNA methylation data (Illumina 450k array) for the current EWAS. Before running the EWAS, please apply quality control (QC) and normalization. Let us know if you if we can help you with this. Please perform the analysis on the methylation beta-values (which range from 0 to 1).

DNA methylation quality control

Please let us know which QC-filters you applied to probes and samples and which normalization method you applied to the methylation data you used as input for this EWAS analysis, by completing the excel sheet (Aggression_CohortName_CohortInformation_YYYYMMDD_InitialsAnalyst.xls).

We assume that you have removed bad quality samples, and have set probes to NA within a sample (prior to the analysis) based on:

* detection p-value (we advise to set probes with p > 0.01 to NA)
* low bead count on array (we advise to set probes with bead count < 3 to NA)
* raw intensity value of exactly zero (probe not present on array)
* and removed probes with low success rate (we advise to remove probes if success rate < 0.95)

Apart from probes that fail QC in your cohort, please analyze **ALL** genome-wide CpGs (if possible).

Note: We will filter out cross-reactive probes, and probes that harbor a SNP prior to the meta-analysis.

Phenotype: Aggressive behavior

We aim to perform EWAS analyses on a quantitative measure of aggressive behavior (higher score should correspond to a higher level of aggressive behavior). Please let us know what measure you used, by completing the excel sheet (Aggression_CohortName_CohortInformation_YYYYMMDD_InitialsAnalyst.xls).

2. Analysis

2.1 - EWAS

Model:

For all methylation sites, please run 2 models:

1. Methylation ~ Aggression + Age + Sex + WBC percentages + Technical + Cohort Specific
2. Methylation ~ Aggression + Age + Sex + WBC percentages + Technical + Cohort Specific + BMI + Smoking

Note: If your cohort includes related individuals (e.g. family members, twins), please apply a statistical approach that takes the clustering of data into account (e.g. gee or linear mixed models).

Covariates

Age= Age when DNA sample was collected

WBC= White blood cell percentage in the same blood sample from which DNA was extracted. If you did not measure white blood percentages in the same sample as used for the DNA methylation measurement, please estimate WBC percentages using a prediction method (e.g. Houseman's referencebased method). For computational reasons, please do not include multiple WBC that are highly correlated with each other or that show very little variation between people in your cohort. For example, in the NTR, we use the following WBC as covariates: monocyte percentage, eosinophil percentage, neutrophil percentage.

Technical covariates + Cohort Specific covariates Please correct for technical (batch) covariates and other cohort-specific covariates as you deem necessary. For example, at the NTR, we include 450k array row and either sample plate or (carefully chosen) principal components from the methylation data. If your cohort includes multiple population ancestries, please take this into account as you think is most appropriate for your cohort (possible strategies include exclusion of small groups of 'ethnic outliers', running the analysis separately by ethnicity, inclusion of a covariate denoting ethnicity, or inclusion of principle components based on genotype data).

BMI= Body mass index

Smoking= Smoking status at the moment of blood sampling, 3 levels: 0=never smoked, 1=former smoker, 2=current smoker.

Software and example R-script

Feel free to use your own analysis pipeline or preferred software to run the models outlined above. Example R-code with instructions is provided in:

Example LM script_012016.r [suited for cohorts that include unrelated subjects]

Example gee script_012016.r [suited for cohorts that include related subjects (family members)]

2. Analysis

2.2 – DNAm age acceleration and aggression

Estimating DNAm age

DNAm age acceleration can be easily calculated with the convenient online tool from Steve Horvath: (http://labs.genetics.ucla.edu/horvath/dnamage/). Please follow the instructions on the website.

In short:

* Upload raw (un-normalized) methylation beta-values (you may preselect probes or include all genome-wide data)
* Remove the bad quality samples you also removed from your EWAS
* Include at least the probes listed in "datMiniAnnotiation.csv " (downloadable from the website)
* Upload a sample annotation file, with the column names "SampleID", "Age", "Female", "Tissue"
* Column "Female" should contain 1 for females, and 0 for males.
* Crucial: Methylation file and sample annotation file must be in the same order
* For very large cohorts, it is recommended to upload your data in batches of up to around 1000 samples (and to apply preselection of probes)
* Max file size that can be uploaded is 800 Mb
* Crucial: select the options **Normalize Data** and **Advanced Analysis of Blood**
* Crucial: **do NOT use** the option Fast Imputation
* Hint: First test if your input files are in the correct format by running the calculator without the option normalize data and without the option advanced analysis of blood, which should give you the output very quickly. If you do not receive an email with results within one hour, the format of your input files is not correct.

Variables

Please collect the following variables from the DNAm age calculator output file:

* AgeAccelerationResidual
* AHOAdjCellCounts
* AAHAAdjCellCounts

Model

Please run 3 models:

1. AgeAccelerationResidual ~ Aggression + Sex + BMI + Smoking + Cohort Specific
2. AHOAdjCellCounts ~ Aggression + Sex + BMI + Smoking + Cohort Specific
3. AAHAAdjCellCounts ~ Aggression + Sex + BMI + Smoking + Cohort Specific

Note: If your cohort includes related individuals (e.g. family members, twins), please apply a statistical approach that takes the clustering of data into account (e.g. gee or linear mixed models).

3. Please send us the following files

Cohort information file

File name:

Aggression_CohortName_CohortInformation_YYYYMMDD_InitialsAnalyst.xls

example: Aggression_NTR_ CohortInformation _20160119_JVD.xls.

Please complete the 3 sheets of this excel file.

EWAS output

File format: .csv (comma-delimited file)

File name:

Aggression_CohortName_Model1_YYYYMMDD_InitialsAnalyst.csv [5 columns]

Aggression_CohortName_Model2_YYYYMMDD_InitialsAnalyst.csv [8 columns]

example: Aggression_NTR_Model1_20160119_JVD.csv.

Requested columns (include the headers in your file) **model 1**:

cgid	Illumina probe ID
N	Sample size
Beta_Agg	Effect size (for aggression)
SE_Agg	Standard error of beta (for aggression)
Pval_Agg	P-value (for aggression)

** We would appreciate it if you could adhere precisely to the column naming scheme as indicated ("Pval_Agg", not: "pval_agg" or "P")*

Requested columns (include the headers in your file) **model 2**:

cgid Illumina probe ID

N Sample size

Beta_Agg Effect size (for aggression)

SE_Agg Standard error of beta (for aggression)

Pval_Agg P-value (for aggression)

Beta_Smoking Effect size (for smoking)

SE_Smoking Standard error of beta (for smoking)

Pval_Smoking P-value (for smoking)

* We will use the summary statistics of the variable smoking for QC purposes.

** We would appreciate it if you could adhere precisely to the column naming scheme as indicated ("Pval_Agg", not: "pval_agg" or "P")*

DNAmAge output

File format: .csv (comma-delimited file)

File name:

Aggression_CohortName_DNAmAge_YYYYMMDD_InitialsAnalyst.csv [5 columns]

example: Aggression_NTR_ DNAmAge_20160119_JVD.csv.

Number of rows:4.

Row 1=header

Row 2-4= the 3 DNAmAge acceleration measures

Requested columns (include the headers in your file):

Measure	"AgeAccelerationResidual" or **"**AHOAdjCellCounts" or "AAHAAdjCellCounts"
N	Sample size
Beta_Agg	Effect size (for aggression)
SE_Agg	Standard error of beta (for aggression)
Pval_Agg	P-value (for aggression)

** We would appreciate it if you could adhere precisely to the column naming scheme as indicated ("Pval_Agg", not: "pval_agg" or "P") and use exactly the same names for the DNAmAge acceleration measures to avoid confusion.*

File upload

4 files Aggression_CohortName_CohortInformation_YYYYMMDD_InitialsAnalyst.xls

Aggression_CohortName_Model1_YYYYMMDD_InitialsAnalyst.csv

Aggression_CohortName_Model2_YYYYMMDD_InitialsAnalyst.csv

Aggression_CohortName_DNAmAge_YYYYMMDD_InitialsAnalyst.

SFTP server address: lisa.surfsara.nl

Username: agg_ewas

Please upload your files to the folder /DATA_UPLOAD

Password: Iwant2UpLoadAGG.

Thank you for your contribution to this project!

We realize that filling out excel sheets is not your favorite job. But please note that all the information we are collecting will be used to make sure that the analyses of all cohorts are harmonized before we do the meta-analysis. The success and progress of the meta-analysis crucially depends on the availability of all information, so we would really appreciate it if everyone could return all completed files together with their EWAS results. To put all this information together and run the analysis, we think that 6 weeks should be an appropriate timeframe (assuming that your methylation data is ready to analyze), however, if this is too soon for you, just let us know. Your effort in carefully performing the analysis and completing all information is more important to us than the deadline.

Thank you very much in advance for your cooperation!