More info on LIFECYCLE online:
lifecycle-project.eu

LifeCycle and ATHLETE projects - SOP:

How to create an ExpressionSet with methylation data for analysis through DataSHIELD

Authors:
Sofia Aguilar
Janine Felix
Martine Vrijheid
Juan Ramon González
Mariona Bustamante

Version: v1

Date: 16-04-2021

Appendix 1 of LifeCycle report D5.4 - Report on the DNA methylation loci that mediate the relationships of early-life stressors with asthma and chronic obstructive respiratory disease

LIFECYCLE Project
Coordination and leadership:

Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands

Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

Appendix 1

LifeCycle and ATHLETE projects - SOP:
How to create an ExpressionSet with methylation data for analysis through DataSHIELD

Developed by: Sofia Aguilar, Janine Felix, Martine Vrijheid, Juan Ramon González, Mariona Bustamante

Version: v1
Date: 2021.04.16

Workflow of EWAS through DataSHIELD in LifeCycle and ATHLETE projects

In LifeCycle and ATHLETE, differently from PACE consortium, EWAS will not be run by the cohorts but by the leading team through DataSHIELD. DataSHIELD is an infrastructure and series of R packages that enable remote and non-disclosive analysis of sensitive research data (https://www.datashield.ac.uk/).

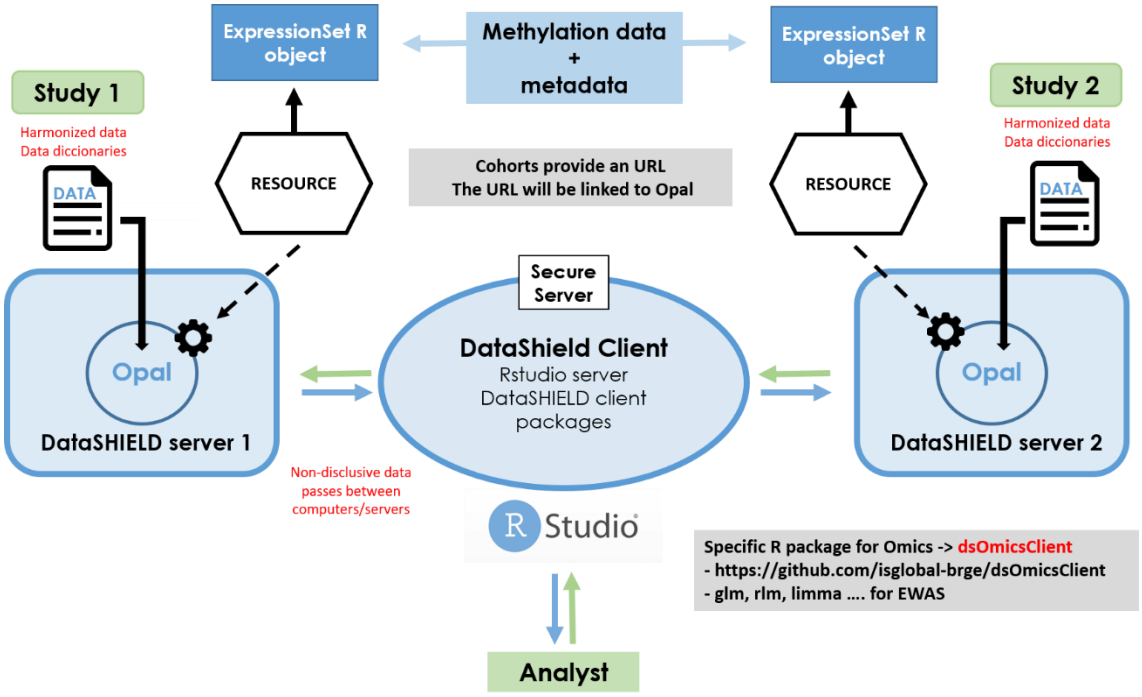

Figure 1. Representation of epigenome-wide association analysis (EWAS) through DataSHIELD.

The workflow will include tasks done by the cohorts and others done by the leaders.
To be done by the cohorts:

1) **Prepare DNA methylation and associated metadata**
   Cohorts will have to prepare the DNA methylation and associated metadata in a specific format named ExpressionSet and link it as a resource to an Opal server. This ExpressionSet has only to be created once. See below for more details.
2) **Sign a Data Access Agreement**

Cohorts and the leading teams will sign a Data Access Agreement (DAA) to allow access the data needed for the analysis, both the ExpressionSet and the other data already harmonized in LifeCycle/ATHLETE.

To be done by the leading team:

1) **Merge datasets and calculate descriptive**
   First, the leading team will merge the methylation data with other variables (exposures, phenotypic traits, covariates), and after selecting complete cases, will calculate descriptive by cohort. This will be done through DataShield and the dsBaseClient package/tool (1,2).
2) **EWAS**
   Then, the leading team will run EWAS for each of the models in each cohort. This will be done through DataShield and the dsOmicsClient package/tool (3). Summarized results will be saved locally at the leading team servers for the next steps.
2 **Meta-analyses**
   After that, the leading team will combine the results from each cohort through meta-analyse, using their preferred method/tool.
3 **Sensitivity analyses**
   Several sensitivity analyses are possible: by array, by ancestry, by region, leave-one-out, etc.
4 **Biological interpretation of findings**
   Finally, the leading team will compare the results across models and top CpGs will be annotated and several types of functional enrichment analyses will be conducted.
5 **Manuscript writing**
   The leading team will prepare a first draft of the manuscript which will be sent to all co-authors.

LifeCycle and ATHLETE recommend being at least two co-leading teams in each project:
1) To take advantage of the expertize of each team.
2) To decrease errors during the analyses.
3) To allow leadership by all partners in their topics of interest.

We recommend splitting the work as follows:
1) **Analysis plan:** to be designed by all co-leaders
2) **EWAS:** each co-leader runs the EWAS of ½ of the cohorts with around 10-20% of the cohorts being run by both teams.
3) **Meta-analyses:** to be run in parallel by the two co-leaders.
4) **Sensitivity analysis and downstream functional enrichment analysis:** to be split among the co-leaders.
5) **Manuscript writing:** to be split among the co-leaders.

What is an ExpressionSet?

An ExpressionSet is an R object designed to combine several different sources of information into a single convenient structure. It contains:

- A matrix with the DNA methylation data (CpGs in rows and samples in columns).
- A dataframe, specifically an AnnotadedDataFrame, with metadata about samples (samples in rows and variables in columns).

NOTE! Column names of the methylation matrix have to be the same row names of the metadata.

- An annotation data frame describing the probes/CpGs included in the ExpressionSet.

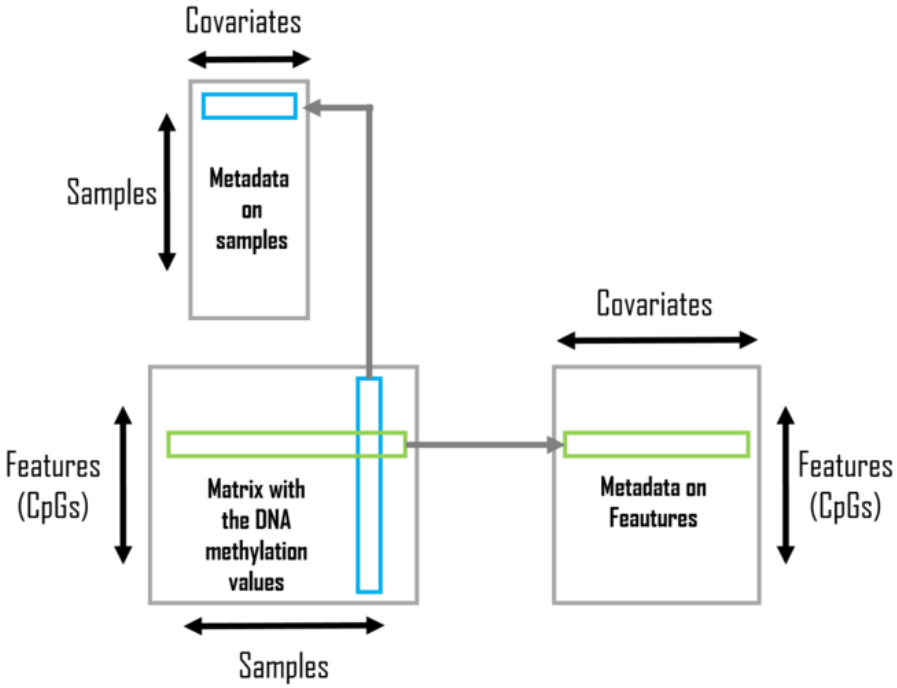

Figure 2. Representation of the structure an ExpressionSet

How many ExpressionSets have to be created?

The cohorts will have to create an ExpressionSet for each:

- **Subject type:**
  - Child
  - Mother
  - Father
- **Tissue:**
  - Blood
  - Placenta
  - Saliva
  - Buccal epithelial cells (BEC)
  - Nasal epithelial cells (NEC)
- **Age:**
  - For mothers during pregnancy, mean weeks of pregnancy of the visit: 12 weeks (12w), 20w, 32w, etc.
  - For children, at birth: 0 years (0y)
  - For children, mean age of the postnatal visit: 2y, 4y, 7y, etc.
  - For parents, during postnatal visits: Adult
- **Array:**
  - 450K
  - EPIC

ExpressionSets will have to be named as:
Cohort_Methyl_Subject_Tissue_Age_Array_Date.Rdata
(ie. INMA_Methyl_Child_Blood_0y_450K_20210215.Rdata).

The ExpressionSets only have to be created once, as the same ExpressionSets will be used in different projects of LifeCycle and ATHLETE (after signing the proper Data Access Agreements). Only in the case where we decide to apply a cross-cohort harmonize pipeline for the quality control and normalization of the methylation data, the cohorts will have to create new ExpresisonSets.
Besides the ExpressionSet, the cohorts will have to prepare a document named **Cohort_Methyl_Subject_Array_Blood_Age_Methods_Date.doc** (ie. INMA_Methyl_Child_Blood_0y_450K_Methods_20210513.doc, attached as an example), giving methodological details about the DNA methylation data. It should include the following information:

- Sample collection and DNA extraction (age, blood fractioning, placental region, extraction method, etc.)
- DNA methylation acquisition (array, laboratory, randomization, etc.)
- QC and normalization and batch control (sample QC, probes QC, normalization and batch correction, etc.).
- Additional information of the metadata in the ExpressionSet. Please give us details on which variables should be used to adjust the models for (batch_methyl, cohort_methyl, sel_methyl, gwas_pcs, etc), and their definition.

Which type of methylation data has to be included in the ExpressionSet?
DNA methylation assessed with the Illumina Infinium 450K or EPIC arrays and expressed as beta values (0 = completely un-methylated, 1 = completely methylated).
Each cohort can use their preferred normalization method and quality control pipeline. At this point, include rather than exclude probes. Exclusion of probes will be done by leaders of each project.

Which metadata has to be included in the ExpressionSet?
The metadata file of the ExpressionSet should include the following variables, with the names and levels/units indicated:

- **ID1 (id_methyl):** ID used in the methylation dataset. It is the row name of the metadata and has to be consistent with the column name of the methylation matrix.
- **ID2 (id):** ID used in the cohort (it might be or not the same ID as the methylation dataset). It will be used to link the DNA methylation data with the other data harmonized in LifeCycle and ATHLETE (EU Child Cohort Variable Catalogue).
- **Sex (sex_methyl):** Sex of the participant [1=Male, 2=Female].
- **Age (age_methyl):** Age of each participant when DNA methylation was assessed [continuous, years].
- **Ancestry (anc_methyl):** Ancestry of the participant. Major ethnic groups [1=European, 2=African or African American, 3=East Asian, 4=South Asian, 5=Native or admixed American, 6=Other]. If you cannot create a variable with the suggested levels, then add

the most appropriate variable for your cohort study and describe it in the Methods file. If possible, use genetic data to define ancestry. If genetic data is missing for a subset of subjects, then complement it with self-reported ancestry. If genetic data is not available or it has many missing values, use self-reported ancestry.

- **Gestational age (ga_methyl):** Gestational age of each participant [continuous, days]. Only for cord blood DNA methylation.
- **Birth weight (bw_methyl):** Weight of the participant at birth [continuous, grams]. Only for cord blood DNA methylation.
- **Child/adolescent body mass index (BMI) z-score (zbmi_methyl)**: Body mass index (BMI) z-score calculated at the age when blood DNA methylation was assessed, defined following the WHO standard curves for children (4,5) [continuous, z-score]. Only for samples from children or adolescents (from 2 to 18 years old). If you need help to calculate it, contact us.
- **Adult body mass index (BMI) (bmi_methyl):** Body mass index (BMI) at the age when blood DNA methylation was assessed. Only for samples from adult participants (>18 years old). For maternal samples obtained during pregnancy use maternal pre-pregnancy BMI or maternal early-pregnancy BMI (up to 16th week of gestation). Indicate it in the Methods file.
- **Cell type proportions:** Blood cell type proportions calculated from different reference panels. Note that depending on the type of biological samples and age, the reference panel and algorithm will be different. We are providing the code to calculate them, if not done yet in your study.

  A) Cord blood cell type proportions: Cord blood cell proportions should be estimated using the Gervin and Salas reference panel (6), the IDOL algorithm (7) for selection of 517 CpGs (for 450K and EPIC arrays) (8) and the constrained projection-quadratic programming algorithm by Houseman (6) for deconvolution of 7 main blood cell types.
    - Variable names: CD8T, CD4T, NK, Bcell, Mono, Gran, nRBC
    - Code: LC-ATH_Deconv_CordBlood_Gervin_Code_v1_20210513.R(attached to this SOP)

  B) Child and adult blood cell type proportions: Child and adult blood cell proportions should be estimated using two reference panels:

  B.1) The Reinius reference panel (9) with the pickCompProbes method (minfi) for CpG selection, and the Houseman algorithm (10) for deconvolution of 6 main blood cell types (450K array):
    - Variable names: CD4T_H, CD8T_H, NK_H, Bcell_H, Mono_H, Gran_H
    - Code: LC-ATH_Deconv_AdultBlood_Houseman_Code_v1_20210513.R (attached to this SOP)

  B.2) The Salas reference panel (8) with the IDOL algorithm (7) for CpG selection (450 CpGs for EPIC and 350 CpGs for 450K) and the Houseman algorithm (10) for deconvolution of 6 main blood:
    - Variable names: CD4T_S, CD8T_S, NK_S, Bcell_S, Mono_S, Neu_S
    - Code: LC-ATH_Deconv_AdultBlood_Salas_Code_v1_20210513.R (attached to this SOP)

NOTE: We would like you to include in the ExpressionSet cell type proportions calculated with both the Reinius and Salas reference panels naming them as we explained above. In this way, each leading group will be able to select one or the other for their EWAS.

- **Technical batch variable (batch_methyl):** Optional. Models will be adjusted for the batch variable suggested by the cohort, unless batch effect has already been corrected during the quality control through methods such as ComBat. Batch variable will be specific to each ExpressionSet created at each age.
- **Cohort specific variable (cohort_methyl):** Optional. If your cohort includes several centres or subcohorts, add this variable in the ExpressionSet.
- **Selection factor (sel_methyl)**: Optional. If your study population oversampled on a condition, then you should add this variable in the ExpressionSet (e.g., if your study population is from a case-control study of asthma, then asthma status should be included in the ExpresionSet).
- **GWAS PCs (gwas_pc1, gwas_pc2, etc.)**: Optional. Models will be run by major ancestry groups (e.g. European). However, if your cohort might still have population substratification within major ancestry groups, then models will be adjusted for GWAS PCs. Whether this is needed and, if so, the number of GWAS PCs to be included as covariates has to be determined by the cohort.

NOTES:

- ID1 (id_methyl). It is essential! We will use it to link methylation data with metadata.
- ID2 (id). It is essential! We will use it to link methylation data and metadata with other harmonized variables from LifeCycle and/or ATHLETE.
- It is really important that you name and level the variables as indicates in this SOP. Any deviation in the name or definition should be indicated in the Methods file.
- If your cohort has twins or siblings, you should keep just one of them in the ExpressionSet. Select them randomly. Removing this sample from the metadata will be enough to exclude this sample from the ExpressionSet.
- Samples with missing values in the covariates (ei. with missing values in birth weight, ancestry, etc.) do NOT have to be filtered out from the ExpressionSet. Each project will have a different set of covariates and will decide how to deal with missing values.

Which is the code to create an ExpressionSet with methylation data?

A tutorial and R code to create the ExpressionSets with methylation data are attached to this SOP. They also contain the code to link the ExpressionSet as a resource to your Opal server.

LC-ATH_ExprSet_MethBlood_Tutorial_v2_20210623.html
LC-ATH_Code_Methyl-ExpressionSet_v2_20210623.R

NOTE: If you have questions, please contact sofia.aguilar@isglobal.org.

References to Appendix

1. Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol. 2014;43(6):1929-44.
2. datashield/dsBaseClient: DataSHIELD client side base functions. Available at: https://github.com/datashield/dsBaseClient. (Accessed: 5th February 2021)

3. GitHub - isglobal-brge/dsOmicsClient: dsOmic client site base functions. Available at: https://github.com/isglobal-brge/dsOmicsClient. (Accessed: 15th October 2020)
4. WHO | Application tools. WHO (2018).
5. WHO | WHO Anthro Survey Analyser and other tools. WHO (2019).
6. Gervin K, Salas LA, Bakulski KM, van Zelm MC, Koestler DC, Wiencke JK, et al. Systematic evaluation and validation of reference and library selection methods for deconvolution of cord blood DNA methylation data. Clin Epigenetics. 2019;11(1):125.
7. Koestler DC, Jones MJ, Usset J, Christensen BC, Butler RA, Kobor MS, et al. Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). BMC Bioinformatics. 2016;17:120.
8. Salas LA, Koestler DC, Butler RA, Hansen HM, Wiencke JK, Kelsey KT, et al. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol. 2018;19(1):64.
9. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén SE, Greco D, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS One. 2012;7(7):e41361.
10. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86.