

Analysis of maternal glucose, insulin and OGTT

Elmar.W.Tobi, Caitlin Howe, Diana Juvinao-Quintero, Sylvain Sebert, Marie-France Hivert

For questions and joining the meta-EWAS mail: Elmar.tobi@wur.nl & MHIVERT@PARTNERS.ORG

Introduction

Gestational diabetes mellitus (GDM) occurs in up to 18% of pregnancies and is a risk factor for additional pregnancy complications [1] and metabolic disease in the offspring in adulthood. The relationship between maternal glucose metabolism and the risk of pregnancy complications [2], birth weight [3] and the development of obesity and diabetes in adulthood [4, 5] has been found to be linear. This latter relationship is attributable to the adverse prenatal environment, more so than familial or genetic factors [6] and a Mendelian Randomization (MR) study suggested that this dose-dependent effect is causal [7].

This imprint of the prenatal environment on adult health is hypothesized to be mediated through alterations in epigenetic marks [8]. The best studied epigenetic mark is DNA methylation (DNAm), a modification of CpG dinucleotides associated with gene expression potential [9]. Candidate gene studies uncovered a linear [10], and an MR study a possible causal [11], effect of maternal glucose on child DNAm.

Objective

We will undertake multiple EWAs to identify candidate “maternal glucose responsive (MGR)” loci. As (fasted) serum glucose, our main exposure variable, is just one facet that represents maternal glucose metabolism we will also undertake a genome-scale screen of maternal (fasted) insulin and oral glucose tolerance test (OGTT) (when available) [12] measurements. These candidate MGR loci identified in these analyses will be subjected to causal modelling and mediation analysis to understand the role of potentially modifiable co-variates (e.g. life-style, breastfeeding duration etc) and will like-wise be tested for associations with adult disease.

The models in our genome-wide screen will be minimal, e.g. adjusting for only a few covariates, based on Directed Acyclic Diagrams (DAG) (Supplemental 1, Supplemental Figures 1 and 2).

If you wish to join the meta-EWAS

Please fill in the supplemental excel file which asks cohort details and detailed contact information and read the analysis plan. Send the supplemental excel file and any questions you have when reading the analysis plan BEFORE embarking on the actual EWAS analyses to Elmar.tobi@wur.nl and mhivert@partners.org

Overall objective

- To investigate the associations of maternal glucose metabolism during pregnancy with offspring DNAm.

Specific objectives

- To investigate DNA methylation profiles in newborn blood and blood collected in later childhood in relation to **maternal fasting glucose** levels (plasma or whole blood values) during pregnancy.

If available in the cohort:

- To investigate DNA methylation profiles in newborn blood and blood collected in childhood in relation to the **maternal glucose response to OGTT** during pregnancy.
- To investigate DNA methylation profiles in newborn blood and blood collected in childhood in relation to **maternal fasted serum insulin** levels during pregnancy.

Exclusions and special considerations

- Exclude multiple births (i.e. singleton only analysis) and multiple siblings from the same family (include 1 child per family in case of multiple siblings from the same family).
- Exclude preterm offspring born at <36weeks (e.g. <252 days into gestation)
- Exclude offspring of mothers with pre-existing type 1 or type 2 diabetes before pregnancy. If your cohort has a large fraction of such cases, please let us know, as these may be interesting for potential downstream analyses.
- For pregnancies complicated by GDM: **INCLUDE** offspring in which the mother started GDM treatment after the glucose/insulin measurement, using adjustments defined later in this document. **EXCLUDE** offspring from mothers in whom treatment for GDM was started before the main glucose/insulin measurement moment used in the cohort.
- **High GDM risk cohorts** should clearly note the high-risk nature of their cohort in their cohort description in the readme file and when supplying the test statistics to the meta-analysis centre.

Exposures

- **Maternal fasted glucose** during pregnancy.
 - **the glucose values should be in mmol/L!!**
 - if **only** non-fasted serum glucose measurements are available contact the meta-analysis center
 - if a mixture of fasted and non-fasted measurements is available: **adjust** in the statistical models for non-fasted state (with fasted as reference category, e.g “0” and non-fasted as “1”)
 - If the available measurements are a **mixture** of serum and whole blood measurements contact the meta-analysis center
 - if multiple measurements in pregnancy are available: take the **earliest** measurement
 - if only an oral glucose tolerance test is available take the serum glucose **measurement at 0 minutes**
- **maternal glucose response to OGTT** during pregnancy.
 - Calculate the area under the curve using the trapezoid method (like detailed in Matthew *et al.* BMJ 1990, appendix II) **for 0 min, 60 min and 120 min measurement moments only (!) in mmol/L (!!!).**

t_i is time of measurement at moment i

y_i is the serum glucose measurement at moment i

$$AUC = \frac{1}{2} \sum_{i=0}^{n-1} (t_{i+1} - t_i) (y_i + y_{i+1})$$

Example OGTT AUC calculation

Time	0 min	60 min	120 min
Glucose	5 mmol/L	8 mmol/L	6 mmol/L

$$AUC = 0.5 * ((60 - 0) * (5+8) + (120 - 60) * (8+6))$$

[**AUC** = 0.5 * ((time=60 minutes) – (time=0 minutes)) * (glucose level at time=0 minutes + glucose level at time=60 minutes) + ((time=120 minutes) – (time=60 minutes)) * (glucose level at time=60 minutes) + (glucose level at time=120 minutes))]

- **Log2 transformed maternal fasted insulin** levels during pregnancy.
 - **Make sure the insulin values are in pmol/L (e.g. picomol/L) before the log2 transformation (so NOT log10)**, if not recalculate to pmol/L (the molecular weight of insulin is 5734 grams/mol)
 - if ***only*** non-fasted measurements are available contact the meta-analysis center
 - if a mixture of fasted and non-fasted measurements is available: ***adjust*** in the statistical models for non-fasted state (with fasted as reference category)
 - if multiple measurements are available: take the ***earliest*** measurement
 - if only an oral glucose tolerance test is available take the serum insulin at 0 minutes of the OGTT

Outcome: newborn/childhood DNA methylation

Illumina Infinium 450k / 850k methylation data in (**newborn**) **offspring blood** as available in your cohort. The **autosomal** “Beta values” will be used as the outcome. Please normalize using your preferred normalization package and indicate which that was in the readme file. Use preferred study QC settings for probe filtering. If in any doubt, please include rather than exclude probes at this stage. We are not excluding outlier values in the beta values (so no “3IQR filtering”).

Repeated measurement for DNA methylation

Some cohorts will have multiple DNA methylation measurements for the same child. Please do the following for this meta-EWAS:

- In the case of a measurement at birth (cord blood) and at *one* later childhood/adolescence age:
Run separate analyses for the cord blood measurement at birth and a separate analysis for the measurement at a later childhood age (e.g. two EWAS)
- In the case of a measurement at birth (cord blood) and *multiple* measurements during childhood and adolescence:
Run separate analyses for the cord blood measurement at birth and an analysis for the youngest childhood measurement.
- In the case of *multiple* measurements during childhood/adolescence:
Run an analysis for the youngest measurement.

We will want to do follow-up longitudinal analyses of significant findings from the cross-sectional analyses. Please fill in details about the availability of repeated measurement in your cohort description

Covariates

- **Sex:** sex of the child. Binary encoding (e.g. 0 /1 (female/male))
- **Maternal Age:** continuous in years (e.g. 34 and numeric)
- **Gestational Age at glucose/insulin measurement:** continuous in days (e.g. 110 and numeric)
- **Gestational Age at birth:** continuous in days (e.g. 110 and numeric)
- **Parity:** Binary encoding with first child coded as 0 and higher parities coded as 1 (e.g. 0 or 1 and numeric).
- **Batch** covariates: Adjustment for batch effects should be done by including the most important covariate(s) (for example, position on micro-array and bisulfite plate) from each individual cohort. Alternatively, a batch correction method such as ComBat is fine. Please indicate clearly in the cohort description how the data is processed and which batch effects are corrected for.
- **Estimated cell types:** Please include in relevant models (see below). These cell proportions should be included additively in the model. Cell proportions are estimated using `estimateCellCounts()` in the *minfi* R package. For newborn cord blood methylation, please select “cordBlood” in the `estimateCellCounts()` function, and include the **7 cell types** imputed: nRBC, CD8T, CD4T, NK, Bcell, Mono, Gran. For the childhood/adolescence analysis use `estimateCellCounts()` in the *minfi* R package with reference `FlowSorted.Blood.450k` (450k, e.k.a the “Reinius reference set”) or `FlowSorted.Blood.EPIC` (850k), and include the **6 cell types**: CD8T, CD4T, NK, Bcell, Mono, Gran.
- **Ancestry:** *optional covariate*. Please analyse major ethnic groups separately (e.g. European, Latino, African, Asian ancestry). If necessary, a cohort may include PCs from a GWAS within a specific ancestry if available. Please define clearly in the cohort description how ancestry was coded/handled.
- **Age:** *for the childhood/adolescence analysis*, the age at blood sampling. Continuous in years (e.g. 12.4y and numeric)
- **Selection_factors:** *optional covariate*. Please include if relevant for your study, for example if your sample contains cases and controls of some condition, please include the case/control variable and clearly denote what was done in your readme file.

- **Fasted:** optional covariate for glucose and insulin analyses. When a subset of glycaemic measurements has been done in the cohort on a non-fasted pregnant woman adjust for this difference. Binary encoding (e.g. 0 or 1 and numeric) with the fasted group as reference (e.g. 0).
- **GDM_treatment:** optional covariate for glucose and insulin analyses. In the case where a treatment for GDM has been initiated after the glucose/insulin analyses adjust for this effect. Factorial variable with the non GDM group as reference (e.g. 0), GDM cases with treatment via diet coded as the next level (e.g. 1) and GDM cases treated with pharmacological agents (insulin etc) with and without diet changes as the highest factorial level (e.g. 2).
- Please do not include further covariates. If you feel strongly that you need to include additional covariates, please contact us.

Models and analyses

- Robust linear regression modelling (rlm() option in R) for each CpG site individually.
- Important differences between models are highlighted

1) Cord Blood Models

a. Crude model for cord blood (+ add if necessary)

- i. Beta-value ~ glycemic measurement + gestational age at glycemic measurement + batch effects (+ *fasted* + *Ancestry* + *GDM_treatment* + *selection_factor*)

b. Crude model for cord blood with cell type adjustment

Beta-value ~ glycemic measurement + gestational age at glycemic measurement + batch effects + **sex** + **Cellular heterogeneity**
(+ *fasted* + *Ancestry* + *GDM_treatment* + *selection_factor*)

c. Extended model for cord blood:

- i. Beta-value ~ glycemic measurement + gestational age at glycemic measurement + sex + gestational age at birth + parity + maternal age + batch (+ *fasted* + *Ancestry* + *GDM_treatment* + *selection_factor*)

d. Extended model for cord blood with cell type adjustment:

- i. Beta-value ~ glycemic measurement + gestational age at glycemic measurement + gestational age at birth + parity + maternal age + batch + **sex** + **Cellular heterogeneity** (+ *fasted* + *Ancestry* + *GDM_treatment* + *selection_factor*)

2) Non-Cord Blood Models (differences **highlighted**) for childhood/adolescence analysis

a. Crude model for non-cord blood

- i. Beta-value ~ glycemic measurement + gestational age at glycemic measurement + **age at blood sampling** + batch effects (+ *fasted* + *Ancestry* + *GDM_treatment* + *selection_factor*)

b. Crude model for non-cord blood with cell type adjustment

- i. Beta-value ~ glycemic measurement + gestational age at glycemic measurement + **sex** + batch effects + **age at blood sampling** + **Cellular heterogeneity** (+ *fasted* + *Ancestry* + *GDM_treatment* + *selection_factor*)

c. Extended model for non-cord blood:

- i. Beta-value ~ glycemic measurement + gestational age at glycemic measurement + **sex** + gestational age at birth + parity + maternal age during pregnancy + **age at blood sampling** + batch (+ *fasted* + *Ancestry* + *GDM_treatment* + *selection_factor*)

d. Extended model for non-cord blood with cell type adjustment:

- i. Beta-value ~ glycemic measurement + gestational age at glycemic measurement + **sex** + gestational age at birth + parity + maternal age + batch + **age at blood sampling** + **Cellular heterogeneity** (+ *fasted* + *Ancestry* + *GDM_treatment* + *selection_factor*)

- Please **do not** adjust your P-values for multiple testing.

How to deal with missing variables

- Exclude samples with missing observations for any of the covariates (complete case analysis). Also base the descriptive table on the set of samples without missings (a descriptives table is made in the EWAS R script).

Time plan

- | | |
|-------------------------------------|--|
| • July 25 th , 2019 | Final call for joining the EWAS |
| • August 5 th , 2019 | Results upload to meta-analysis center |
| • September 25 th , 2019 | Meta-analyses completed; start downstream analyses |

Data file format

- Tab delimited file; one row per probe. First row is a header with labels as defined in the table below.
- For file naming, please use the following convention:
 - [COHORT]_EWAS_[phenotype]_[Moment]_[Datestamp].txt.gz
 - Cohort = name of cohort in capital letters, no special symbols
 - Phenotype = “glucose”, “insulin”, “OGT_Xgrams”
 - Moment is “birth” (e.g. cord-blood) or “later” (non-cordblood)
 - Datestamp is the date on which the file was prepared. Please use the following format “YYYY_MM_DD”.For example: HAVEN_EWAS_glucose_later_2019_02_20.txt.gz
- A table containing cohort name, phenotype, moment, total sample size, calculated lambda for each model (e.g “lambda_mA”, “lambda_mB”, “lambda_mC”, “lambda_mD” for model a till d: see Example Code). Please use naming convention [COHORT]_Lambdas_[phenotype]_[Moment]_[Datestamp].txt
- Data file contents requirements:
 - If you have missing data in your results file, please do not leave any cells blank. Missing data should be denoted by NA.
 - No quotes should be used around any data cells or headers.
 - Please provide all numeric data with at least 4 decimal places. P-values should be specified to at least 4 relevant digits.

README file

With your data, please upload a readme file with a short description containing the following paragraphs:

- Cohort description: short description of the cohort
- Data handling: the normalization and QC steps taken in your study
- Batches: the batch effects for which you adjusted
- Specific variables: any cohort specific variable for which you adjusted
- Ages: information about the ages of the offspring included
- File information: Any additional information about the uploaded files worth noting

Results format table

Column Header	Description	Format	Examples
Estimate_mA	Effect size model A	String	0.0237
SE_mA	Standard error model A	Numeric	0.0172
Zval_mA	Z value model A	Numeric	1.3802
P_mA	P value model B	Scientific E notation with at least 3 digits to the right of the decimal	3.244E-10
N_mA	N individuals analysed for model A	Numeric	650
Estimate_mB	Idem A for model B		
SE_mB	Idem A for model B		
Zval_mB	Idem A for model B		
P_mB	Idem A for model B		
N_mB	Idem A for model B		
Estimate_mC	Idem A for model C		
SE_mC	Idem A for model C		
Zval_mC	Idem A for model C		
P_mC	Idem A for model C		
N_mC	Idem A for model C		
Estimate_mD	Idem A for model D		
SE_mD	Idem A for model D		
Zval_mD	Idem A for model D		
P_mD	Idem A for model D		
N_mD	Idem A for model D		
probeID	Name of probe	Character	cg000000957

Data exchange

When finished, contact us, and we will send a link for result upload. Send an email to Elmar.tobi@wur.nl and Diana.Lizeth.Juvinao-Quintero@harvardpilgrim.org

Primary contacts for questions about the analysis:

Elmar Tobi (elmar.tobi@wur.nl)

References

1. Hedderston MM, Ferrara A, Sacks DA. Gestational diabetes mellitus and lesser degrees of pregnancy hyperglycemia: Association with increased risk of spontaneous preterm birth. *Obstet Gynecol.* 2003;102:850–6.
2. HAPO Study Cooperative Research Group, Metzger BE, Lowe LP, Dyer AR, Trimble ER, Chaovarindr U, et al. Hyperglycemia and Adverse Pregnancy Outcomes. *N Engl J Med.* 2008;358:1991–2002. doi:10.1056/NEJMoa0707943.
3. HAPO Study Cooperative Research Group. Hyperglycemia and Adverse Pregnancy Outcome (HAPO) Study: Associations With Neonatal Anthropometrics. *Diabetes.* 2009;58:453–9. doi:10.2337/db08-1112.
4. Hillier TA, Pedula KL, Schmidt MM, Mullen JA, Charles M-A, Pettitt DJ. Childhood obesity and metabolic imprinting: the ongoing effects of maternal hyperglycemia. *Diabetes Care.* 2007;30:2287–92. doi:10.2337/dc06-2361.
5. Franks PW, Looker HC, Kobes S, Touger L, Tataranni PA, Hanson RL, et al. Gestational glucose tolerance and risk of type 2 diabetes in young Pima Indian offspring. *Diabetes.* 2006;55:460–5. <http://www.ncbi.nlm.nih.gov/pubmed/16443781>. Accessed 25 Jan 2019.
6. Dabelea D, Hanson RL, Lindsay RS, Pettitt DJ, Imperatore G, Gabir MM, et al. Intrauterine exposure to diabetes conveys risks for type 2 diabetes and obesity: A Study of Discordant Sibships. *Diabetes.* 2000;49:2208–11.
7. Tyrrell J, Richmond RC, Palmer TM, Feenstra B, Rangarajan J, Metrustry S, et al. Genetic Evidence for Causal Relationships Between Maternal Obesity-Related Traits and Birth Weight. *Jama.* 2016;315:1129. doi:10.1001/jama.2016.1975.
8. Waterland RA, Michels KB. Epigenetic epidemiology of the developmental origins hypothesis. *Annu Rev Nutr.* 2007;27:363–88. doi:10.1146/annurev.nutr.27.061406.093705.
9. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet.* 2003;33 Suppl:245–54.
10. Houde A-A, Ruchat S-M, Allard C, Baillargeon J-P, St-Pierre J, Perron P, et al. LRP1B, BRD2 and CACNA1D: new candidate genes in fetal metabolic programming of newborns exposed to maternal hyperglycemia. *Epigenomics.* 2015;7:1111–22. doi:10.2217/epi.15.72.
11. Allard C, Desgagne V, Patenaude J, Lacroix M, Guillemette L, Battista MC, et al. Mendelian randomization supports causality between maternal hyperglycemia and epigenetic regulation of leptin gene in newborns. *Epigenetics.* 2015;10:342–51.
12. Matthews J, Campell M, Altman D, Royston P. Analysis of serial measurements in medical research. *BMJ.* 1990;300:680. doi:10.1136/bmj.300.6719.230.
13. Looman M, Schoenaker DAJM, Soedamah-Muthu SS, Geelen A, Feskens EJM, Mishra GD. Pre-pregnancy dietary carbohydrate quantity and quality, and risk of developing gestational diabetes: the Australian Longitudinal Study on Women’s Health. *Br J Nutr.* 2018;120:435–44. doi:10.1017/S0007114518001277.
14. Asemi Z, Hashemi T, Karamali M, Samimi M, Esmailzadeh A. Effects of vitamin D supplementation on glucose metabolism, lipid concentrations, inflammation, and oxidative stress in gestational diabetes: a double-blind randomized controlled clinical trial. *Am J Clin Nutr.*

2013;98:1425–32. doi:10.3945/ajcn.113.072785.

15. Maktabi M, Jamilian M, Amirani E, Chamani M, Asemi Z. The effects of magnesium and vitamin E co-supplementation on parameters of glucose homeostasis and lipid profiles in patients with gestational diabetes. *Lipids Health Dis.* 2018;17:163. doi:10.1186/s12944-018-0814-5.

16. Tabrizi R, Akbari M, Moosazadeh M, Lankarani KB, Heydari ST, Kolahdooz F, et al. The Effects of Selenium Supplementation on Glucose Metabolism and Lipid Profiles Among Patients with Metabolic Diseases: A Systematic Review and Meta-Analysis of Randomized Controlled Trials. *Horm Metab Res.* 2017;49:826–30. doi:10.1055/s-0043-119544.

17. Ming W-K, Ding W, Zhang CJP, Zhong L, Long Y, Li Z, et al. The effect of exercise during pregnancy on gestational diabetes mellitus in normal-weight women: a systematic review and meta-analysis. *BMC Pregnancy Childbirth.* 2018;18:440. doi:10.1186/s12884-018-2068-7.

18. Lisonkova S, Janssen PA, Sheps SB, Lee SK, Dahlgren L. The effect of maternal age on adverse birth outcomes: does parity matter? *J Obstet Gynaecol Can.* 2010;32:541–8. doi:10.1016/S1701-2163(16)34522-4.

19. Traversy G, Chaput J-P. Alcohol Consumption and Obesity: An Update. *Curr Obes Rep.* 2015;4:122–30. doi:10.1007/s13679-014-0129-4.

20. Feldman HS, Jones KL, Lindsay S, Slymen D, Klonoff-Cohen H, Kao K, et al. Prenatal alcohol exposure patterns and alcohol-related birth defects and growth deficiencies: a prospective study. *Alcohol Clin Exp Res.* 2012;36:670–6. doi:10.1111/j.1530-0277.2011.01664.x.

21. Cho Y, Shin S-Y, Won S, Relton CL, Davey Smith G, Shin M-J. Alcohol intake and cardiovascular risk factors: A Mendelian randomisation study. *Sci Rep.* 2015;5:18422. doi:10.1038/srep18422.

22. Hinkle SN, Albert PS, Mendola P, Sjaarda LA, Yeung E, Boghossian NS, et al. The association between parity and birthweight in a longitudinal consecutive pregnancy cohort. *Paediatr Perinat Epidemiol.* 2014;28:106–15. doi:10.1111/ppe.12099.

23. Lurà MP, Gorlanova O, Müller L, Proietti E, Vienneau D, Reppucci D, et al. Response of cord blood cells to environmental, hereditary and perinatal factors: A prospective birth cohort study. *PLoS One.* 2018;13:e0200236. doi:10.1371/journal.pone.0200236.

24. Herzog EM, Eggink AJ, van der Zee M, Lagendijk J, Willemsen SP, de Jonge R, et al. The impact of early- and late-onset preeclampsia on umbilical cord blood cell populations. *J Reprod Immunol.* 2016;116:81–5. doi:10.1016/j.jri.2016.05.002.

25. Textor J, van der Zander B, Gilthorpe MS, Liskiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package “dagitty”. *Int J Epidemiol.* 2016;45:1887–94. doi:10.1093/ije/dyw341.

Supplement 1. DAG analysis

We hypothesize that there is a direct effect of maternal glucose on child DNAm methylation in a dose-response manner, analogue to the associations between maternal glucose and birthweight[3] and pregnancy complications, like preterm birth and caesarian delivery [2]. For the DAG (Figure 1) we formalize the following relationships. Maternal glucose levels are influenced by maternal body mass index and multiple (micro)nutrients in the maternal diet, in particular carbohydrate quality [13], but also vitamin D [14], magnesium and vitamin E [15]. Diet may also influence glucose metabolism through maternal insulin sensitivity [16], which is also influenced by exercise during pregnancy [17]. Maternal age has been linked to SES, as women in lower SES groups generally give birth to their first child at an earlier age, while older maternal age has been linked to the increased risk of caesarian section and preterm birth [18]. Most of the mentioned lifestyle covariates are linked to socio-economic status and ethnicity, which are also associated with alcohol consumption. Alcohol is energy-rich, but evidence suggests that only heavy drinking has a measurable contribution to obesity [19]. Alcohol is a source of cellular stress and thereby inhibits prenatal growth [20]. There are conflicting reports about the influence of alcohol on serum glucose levels, but Mendelian randomization approaches so far did not provide evidence to support a causal relationship in women [21] and this relation is therefore not considered here. Glucose is a major driver of fetal growth as the main energy carrier, but it is logical that essential (micro)nutrients and fatty acids from the maternal diet are also directly linked to growth, as the child cannot synthesize them, and shortage of the basic building blocks may be argued to stagnate growth. Furthermore, parity is a major independent determinant of birthweight [22]. Sex will explain a large portion of the variation in DNAm and is also associated with birthweight. DNAm is influenced heavily by batch effects, which will be cohort-specific, but should be independent of biological covariates when randomization of samples across arrays has been done correctly. Also, cellular heterogeneity of cord blood leads to additional variation in DNAm for a large subset of 450k probes. For a subset of CpGs it might be argued that DNAm is merely a reflection of cellular heterogeneity (e.g. cellular heterogeneity is the DAG parent of DNAm) (Figure 1), while for another subset DNAm is part of the causal epigenetic mechanism enforcing cell identity (e.g. DNAm is the parent of cell heterogeneity) (Figure 2). Smoke exposure and pregnancy complications like pre-eclampsia and caesarian section are linked to variation in blood cellular heterogeneity [23, 24]. Therefore, it is important to consider cellular heterogeneity as a potential collider in the latter scenario (Figure 2).

We considered these scenarios in the DAG program dagitty [25]. For the scenario in which cellular heterogeneity influences DNAm (Figure 1) adjustment for both cellular heterogeneity and sex is required to ascertain a direct effect of glucose exposure on DNA methylation. In a scenario where DNAm in cord blood drives cellular heterogeneity (Figure 2) adjustment for cellular heterogeneity will make the ascertainment of a direct effect impossible according to the DAG. In this scenario no adjustment for sex and cellular heterogeneity is required to ascertain a direct effect of glucose on DNAm. In both scenarios adjustment for batch effects may help increase the study power by explaining nuisance variation. These leads to the following minimal models:

For the first scenario (Supplemental Figure 1) the model is:

Beta-value ~ glycemic measurement + gestational age at glycemic measurement + sex + batch effects + Cellular heterogeneity

For the second scenario (Supplemental Figure 2) the model is:

Beta-value \sim glycemic measurement + gestational age at glycemic measurement + batch effects

Multiple studies in the meta-analysis have measured glucose at differing gestational ages within a study. Moreover, maternal BMI and behavior (diet and exercise) will vary across gestation. Since most studies start measurements after the moment that the expecting mother knows she is pregnant the measurement timepoint will unlikely have effect on smoking and drinking behavior. DAG analysis shows that the moment of measurement does not influence the ascertainment of a direct effect. Inclusion may take away heterogeneity and may therefore improve study power. DAG analysis shows that the inclusion of this covariate does not influence the possibility to ascertain a direct effect and allows for sex adjustment in both cell heterogeneity scenarios (e.g. both Figure 1 and Figure 2). When we assess the effect of these additional covariates within our DAG framework a direct effect can be found between glucose and DNAm with such an extended model in the first scenario (Figure 1), now including adjustment for sex.

Beta-value \sim glycemic measurement + gestational age at glycemic measurement + sex + gestational age at birth + parity + maternal age + batch effects + Cellular heterogeneity

In the second scenario (Figure 2) where cellular heterogeneity is a collider adjustment is possible, but without adjustment for cellular heterogeneity.

Beta-value \sim glycemic measurement + gestational age at glycemic measurement + sex + gestational age at birth + parity + maternal age + batch

Figure 1. DAG of *in utero* situation

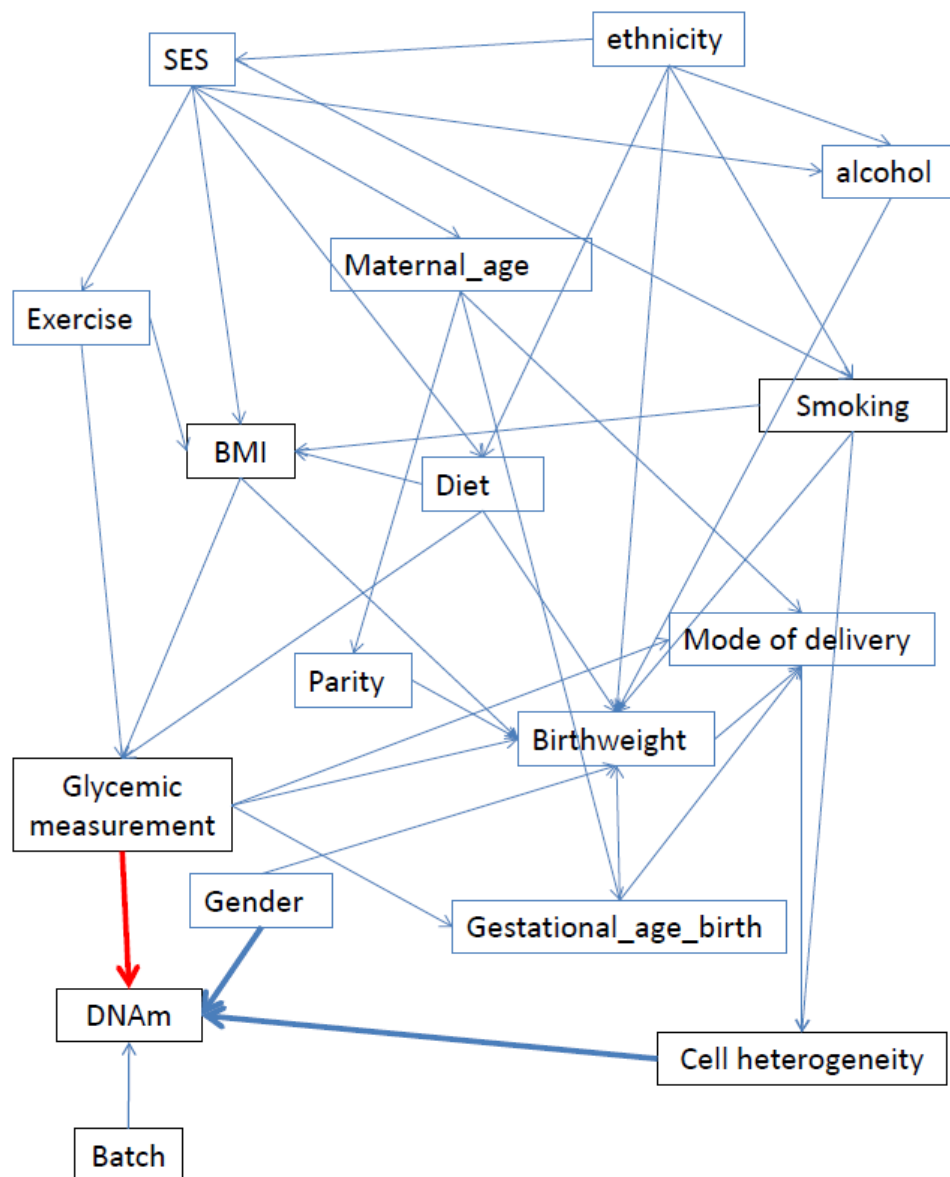
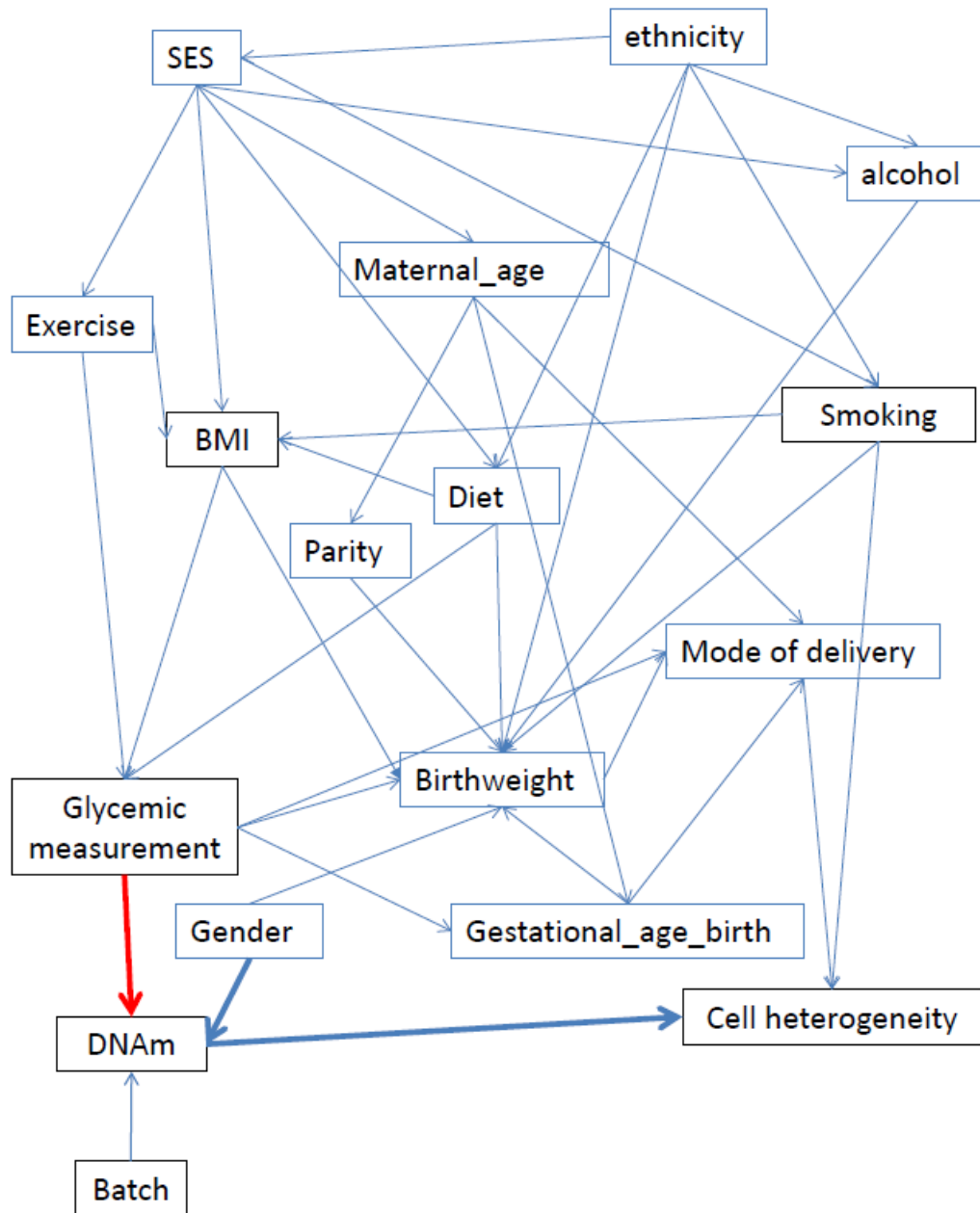


Figure 2. DAG of *in utero* situation, with cell heterogeneity of blood as collider



Supplement 2. R code for EWAS

Please email Elmar Tobi (elmar.tobi@wur.nl) for questions or clarifications.

```
#####
#####
#### Script for glyceimic varibels EWASses PREciSE #####
#####

#####
# author: E.W. Tobi elmar.tobi@wur.nl
# based on scripts used in the PACE consortium; compatible with both Windows AND Linux based operating systems
# by deleting or inclusion of one line of code! Please carefully read
#
#####

### the following code checks if the required packages are installed, if not it installs them
##### requires (local) admin rights on your computer/cluster to install packages! AND access to internet

packages <- c("future.apply", "MASS", "sandwich", "data.table", "lmtest", "R.utils")
if (length(setdiff(packages, rownames(installed.packages()))) > 0) {
  install.packages(setdiff(packages, rownames(installed.packages())), repos="http://cloud.r-project.org/")
}

#####
##### some handy functions for later on #####
#####
# Calculate lambda
Lambda <- function(x) { qchisq(median(x, na.rm=T), df = 1, lower.tail = F)/ qchisq(0.5, 1) }

# helper for descriptives
descriptives <- function(x){
  tmp <- c(mean(x, na.rm=T), sd(x, na.rm=T), length(x), sum(is.na(x)) )
  names(tmp) <- c("Mean", "SD", "N", "N_NA")
  return(tmp)
}

#####
### variables used for the naming of the result files and running rlm
#####
# Please give the name of the cohort without dots/spaces/tabs
Cohort <- "HAVEN"

# the date of analysis: e.g. like 2019_03_01 (march first 2019)
Datestamp <- format(Sys.Date(), format="%Y_%m_%d")

# age: fill in "birth" for cord blood or "later" for ages beyond birth1
Moment <- "later"

#####
### Which EWAS is this choose from: glucose, insulin, OGTT_XXgrams
### of note denote the number of grams of glucose load used in the cohort
### when refering to the OGTT (e.g OGTT_50grams, OGTT_75grams or OGTT_100grams)
#####

# example: here fasted serum glucose:
EWASType <- "glucose"
```

```
#####
# your directory containing both your beta matrix and phenotypes
#####
DataDir <- "C:/Users/Me/PRECISE_project/temp/data"

#####
# your desired directory for placing the outcomes/stats
#####
OutputDir <- file.path(DataDir, "EWAS_results")
# may also be a full path e.g. ("C:/Users/Me/PRECISE_project/temp/data/EWAS_results" in this example)

##### load the methylation and phenotype data #####
#####
#####
#####

# jump your R session to the
# directory where the data is....
setwd(DataDir)

#####
# read in your DNA methylation data (only autosomes), large files in .txt and comparable formats
# are best loaded via the data.table package fread() function!

# Example a RData object containing both DNAm data (named Betas.auto in example )and phenotype file (named pheno in
example)
load(file="Betas_5FN_Combat.RData")

### !!!!! required formats!
# DNA methylation is in a matrix: rows -> probes, columns -> samples
# phenotype is in a data.frame: rows samples, columns covariates

# Pass the Data matrix (beta values matrix) to object Betamatrix; in example "Betas.auto"
# default behaviour of R does not actually copy these sometimes large objects

##### rename DNA methylation data file to "Betamatrix" !
Betamatrix <- Betas.auto
# for RAM saving measures: delete the old named object
rm(Betas.auto)

# Identical to above: Pass the phenotype file (data.frame) to object Phenotypes; in this example "Pheno"
Phenotypes <- pheno
rm(pheno)

#####
## linking phenotype file to DNAm #####
#####

# name of variable in "Phenotypes" files linking to the array or sample name in Betamatrix object
# In example below the variable x in "Phenotypes" contains the microarray name (column names of Betamatrix object)
# e.g. "7668610030_R04C02" etc
array_to_phenotypefile_identifier = "x"

# Now we make sure that the ordering of both files is identical:
# since the Beta matrix can be quite large, it is efficient to re-order the phenotype and let the Betamatrix be
Phenotypes <- Phenotypes[match(colnames(Betamatrix), Phenotypes[,array_to_phenotypefile_identifier]),]
```

```
#####
# Doublechecks: Please mind any output that is generated
# Questions: 1. are the number of samples identical between phenotype file and Betamatrix
#           2. Are the phenotype and betamatrix files ordered in the same way?
#           3. Is the Betamatrix a matrix?
#           4. Is the Phenotype file a data.frame?
#####
if (ldim(Phenotypes)[1]==dim(Betamatrix)[2]) { # the correct number of samples present? Should be TRUE
  print("WRONG:number of samples is not identical between Phenotypes and Betamatrix: please fix!")
} else { print("OK:the Betamatrix and Phenotypecontain the same number of individuals, continue") }

# as a doublecheck: test if ordering of Phenotypes and Betamatrix is identical
if (!identical(as.character(Phenotypes[,array_to_phenotypefile_identifier]),as.character(colnames(Betamatrix))) ) {
  print("WRONG:ordering of files or mismatch in variable type (factor vs, character?); please check and fix")
} else { print("OK:the Betamatrix and Phenotype file contain the same individuals in the correct order, continue") }

if(!is.matrix(Betamatrix)) {
  print("WRONG: Beta matrix (DNAm values) is not a matrix ; please convert via: Betamatrix <- as.matrix(Betamatrix)")
} else { print("OK:Betamatrix is matrix, therefore OK, continue") }

if(!is.data.frame(Phenotypes)) {
  print("WRONG:Phenotypes (covariate file) is not a data.frame ; please convert via: Phenotypes <-
as.data.frame(Phenotypes)")
} else { print("OK: Phenotype file is a data.frame, OK, continue") }

#####
## Please read the Analysis Plan for inclusion/exclusion criteria
#####

# example: here we select only those individuals from European descent (europeans==TRUE)
index_Europeans <- Phenotypes$m_ethn==1 # ethnicity of child family; selecting only European descent

# trim the data: include only eligible individuals
Phenotypes <- Phenotypes[index_Europeans,]
# delete in the matrix all non-europeans
todelete <- colnames(Betamatrix)[!index_Europeans] # vector of samples to delete
Betamatrix <- Betamatrix[,!colnames(Betamatrix) %in% todelete] # removes the columns

# as a doublecheck: test if ordering of Phenotypes and Betamatrix is identical
if (!identical(as.character(Phenotypes[,array_to_phenotypefile_identifier]),as.character(colnames(Betamatrix))) ) {
  print("WRONG:ordering of files or mismatch in variable type (factor vs, character?); please check and fix")
} else { print("OK:the Betamatrix and Phenotype file contain the same individuals in the correct order, continue") }

##### gather the required covariates for the analysis #####
#####
#
#####
#
# Please read the Analysis Plan for details on which covariates to include for your cohort!
# AND how to code them! (only added in short here!)

ID <- Phenotypes[,array_to_phenotypefile_identifier] # array identifier; here used as a proxy for person identifier

# get the glycemic variable: log2 transform insulin, do NOTlog transform glucose
# please check if the variables are in S.I. measurements required for the meta-analysis
```

```

Y <- log2(Phenotypes[, "m_insulin"]) # the fasted glucose measurements, or other glycaemic measure

# additional (biological) covariates
# gestational age at glycaemic measurement
GA_measure <- Phenotypes[, "GAatmeasurement"] # gestational age at maternal measurement of glucose; in days

# Sex of child
Sex <- Phenotypes[, "Sex_child"] # Sex of child; binary; 0/1
Sex[Sex=="Male"] <- 1 # Sex of child; binary; 0/1
Sex[Sex=="Female"] <- 0 # Sex of child; binary; 0/1

# parity:
Parity <- Phenotypes[, "Parity"] # 0= nulliparous and 1 is multiparous; binary; 0/1
Parity[Parity>0] <- 1

# gestational age at birth;
GA_birth <- round(Phenotypes[, "GA_days"]/7, digits=0) # gestational age at birth, in weeks

# maternal age
Mat_age <- Phenotypes[, "mother_age"] # mother's age at pregnancy; in years; numeric

# for non-cord blood EWAS only!! Child's age at blood sampling
Child_age <- Phenotypes[, "child_age"] # child's age at blood sampling; in exact years (e.g. 12.3y); numeric

# fasted/non-fasted at the measurement; e.g. not include this variable for the AUC of the OGTT
Fasted <- Phenotypes[, "m_nuchter"] # glycaemic measure when fasted==0 and non-fasted==1; binary; 0/1
Fasted[Fasted=="YES"] <- 0
Fasted[Fasted=="NO"] <- 1

# optional: GDM_treated ; factorial, 0 (reference) = non-GDM, 1 =GDM diet treatment, 2= GDM pharmacological treatment
# GDM status:
GDM_treatment <- Phenotypes[, "GDM_diagnosis"]
GDM_treatment[GDM_treatment=="no_GDM"] <- 0 # 0= no-GDM
GDM_treatment[GDM_treatment=="GDM_diet_treatment"] <- 1 # 1=GDM with diet treatment
GDM_treatment[GDM_treatment=="GDM_pharmacological_treatment"] <- 2 # 2=GDM pharmacological intervention

# cellular heterogeneity: use Houseman imputed or measured major (cord) blood cell fractions
# for the adjusted model
C1 <- Phenotypes[, "CD8T"]
C2 <- Phenotypes[, "CD4T"]
C3 <- Phenotypes[, "NK"]
C4 <- Phenotypes[, "Bcell"]
C5 <- Phenotypes[, "Mono"]
C6 <- Phenotypes[, "Gran"]
# !!!! Do not forget the nucleated red blood cells in cord blood analyses!!!!
C7 <- Phenotypes[, "NRC"]

# specific cohort variables: like case-control status, delete if not appropriate for your cohort design!
S1 <- Phenotypes[, "Status"] # e.g. case-control status or other relevant covariate

#####
# technical covariates; cohort specific
# !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
#####

# strongly urge participants to investigate the influence of the row on a 450k/850k array measurements
# if the samples have been run on automatic Illumina hybridization/washing stations
# that are placed vertically; in that case there is a gradient in signal intensity from the

```

```

# top to the bottom of the array.
# The column (left or right ) has no influence in that case, so adjusting for "sentrix position" (12 levels in a factor) is
# less efficient
# the R01 till R06 (1-6) is the position effect that may need adjustment in your study.
T1 <- Phenotypes[, "Position"] # continuous (1-6) batch effect denoting the Row of "sentrix_position" R01C1 and R01C2 is
coded as 1 ; R02C01 & R02C02 is coded as 2

## some room for cohort specific Technical stuff;
T2 <- Phenotypes[, "batches"] # BS-plate or other relevant corrections

#####
#####
#####
### create a "Data" data.frame object from all variables noted above
### remove individuals with missing data
### and set all variables to either factor, character, numeric
#####
#####
Data <- data.frame(cbind(ID,Y,GA_measure,Sex,GA_birth,GDM_treatment,Parity,Mat_age,Child_age,Fasted,
                        C1,C2,C3,C4,C5,C6,C7,T1,T2,S1))

### phenotype data should be complete! Find rows with any NAs
index_complete_persons <- !apply(Data,1,anyNA)
#####

#####
### retain only individuals with complete information in the Phenotypes and Betamatrix objects

Data <- Data[index_complete_persons,]
Betamatrix <- Betamatrix[,match(Data$ID,colnames(Betamatrix))]

if( ! identical(as.character(Data$ID), colnames(Betamatrix)) ){
  print("Wrong: Data object containing covariates does not contain the same individuals as the Betamatrix, fix before
continuing")
} else { print("OK! continue") }

### set the variables in Data to the correct attribute type (e.g. character, factor, numeric, ordinal)
# e.g. as the Data object has most variables denoted as a factor we need to convert to as.character first
# to prevent the factor levels to be re-coded as new numerical values
Data$ID <- as.character(Data$ID) # "x" is a character (e.g. column headers in the Betamatrix are also a character)
Data$Y <- as.numeric(as.character(Data$Y)) # glycaemic variable should be numeric; first as.character to prevent
erroneous factor levels coded as new values
Data$GA_measure <- as.numeric(as.character(Data$GA_measure)) # Gestational age at glycaemic measurement; numeric
Data$Sex <- as.numeric(as.character(Data$Sex)) # Sex: 0/1 binary numeric variable
Data$Parity <- as.numeric(as.character(Data$Parity)) # Parity: 0/1 binary numeric variable
Data$GA_birth <- as.numeric(as.character(Data$GA_birth)) # gestational age at birth: numeric
Data$GDM_treatment <- as.factor(Data$GDM_treatment) # GDM treatment,0,1,2; factorial
Data$Mat_age <- as.numeric(as.character(Data$Mat_age)) # age at pregnancy in years; numeric
Data$Child_age <- as.numeric(as.character(Data$Child_age)) # age of child at measurement in years; numeric
Data$Fasted <- as.numeric(as.character(Data$Fasted)) # fasted state: 0=yes, 1=no, binary; numeric

# celltypes: as numeric please! CORD BLOOD: DO NOT FORGET THAT THERE ARE 7 instead of 6 celltypes! (did you not
forget?)
Data$C1 <- as.numeric(as.character(Data$C1))
Data$C2 <- as.numeric(as.character(Data$C2))
Data$C3 <- as.numeric(as.character(Data$C3))
Data$C4 <- as.numeric(as.character(Data$C4))
Data$C5 <- as.numeric(as.character(Data$C5))

```

```

Data$C6 <- as.numeric(as.character(Data$C6))
Data$C7 <- as.numeric(as.character(Data$C7))

# Technical (batch) variation: factor or numeric depending on variable
Data$T1 <- as.numeric(as.character(Data$T1)) # position effect 1-6 as numeric
Data$T2 <- as.factor(Data$T2) # Bisulfite plate, as factor

# cohort specific variables
Data$S1 <- as.numeric(as.character(Data$S1))

##### create a decriptives table of only the included pregnancies #####
#####
#####
#####
### create a decriptives table of only the included pregnancies! (NOT whole cohort!)
### save this as a descriptives file in the results folder
#####
#####

# we want to know:
# Number of children in cohort
# maternal age; mean (SD)
# maternal BMI; mean (SD)
# glycaemic variable; mean (SD)
# child age (SD)
# gestational age at birth (SD)
# Optional: GDM cases: N=no GDM, N=GDM diet N=GDM with pharmacological intervention
# N individuals for each different SES scales!
# get all the data together, only for those included in the EWAS!
allphenotypes <- merge(Data,Phenotypes, by.x="ID", by.y=array_to_phenotypefile_identifier, type="left")

### alterations required depending on which EWAS you run
MaternalGLYC <- descriptives(allphenotypes$m_glucose) # make descriptives of maternal glycaemic variable; here glucose
as example
MaternalBMI <- descriptives(allphenotypes$m_BMI) # make descriptives of maternal BMI

#### the same for each EWAS runned
Maternalage <- descriptives(allphenotypes$Mat_age) # make decriptives of maternal age
GAatbirth <- descriptives(allphenotypes$GA_birth) # make descriptives of GA at birth

### for older children:
agechildren <- descriptives(allphenotypes$Child_age)

##### SES scales (education levels as proxies)
# example for a 3 level scale
Nlow <- sum(allphenotypes$SES==1, na.rm=T)
Nmid <- sum(allphenotypes$SES==2, na.rm=T)
Nhigh <- sum(allphenotypes$SES==3, na.rm=T)
NNA <- sum(is.na(allphenotypes$SES)) # number of SES unknown
# Make a vector so that it fits togheter with the descriptives above
Nlow <- c(NA, NA, Nlow,NA )
Nmid <- c(NA, NA, Nmid,NA )
Nhigh <- c(NA, NA, Nhigh,NA )
NNA <- c(NA, NA, NA , NNA )

# optional variable: when GDM cases exist include the following counts
#NnoGDM <- sum(as.numeric(allphenotypes $GDM_treatment)==0, na.rm=T)
#NGDMdiet <- sum(as.numeric(allphenotypes$GDM_treatment) ==1, na.rm=T)
#NGDMdrugs <- sum(as.numeric(allphenotypes$GDM_treatment)==2, na.rm=T)
# Make a vector so that it fits togheter with the descriptives above
#NnoGDM <- c(NA, NA, Nlow,NA )

```

```

#NGDMdiet <- c(NA, NA, Nmid,NA )
#NGDMdrugs <- c(NA, NA, Nhigh,NA )

### make one table of these results, optionally add GDM diagnosis counts!
Descriptive_Table <- rbind(Maternalage,MaternalBMI,GAatbirth,agechildren, Nlow,Nmid,Nhigh,NNA )
rowDescriptives <- rownames(Descriptive_Table)
rownames(Descriptive_Table) <- NULL
Descriptive_Table <- cbind(rowDescriptives,Descriptive_Table)

# name for Descriptive Table file:
name_descriptive_table <- paste(paste(paste(paste(Cohort,"Descriptives",EWAStype,sep="_"),Moment,sep="_"),
Datestamp,sep="_"),".txt",sep=".")

# write away tab delimited descriptives table
write.table(Descriptive_Table, file=file.path(OutputDir,name_descriptive_table), col.names=T, row.names=F,
quote=F,sep="\t")

##### MODELS
#####
#####
#####
#### define the models: please edit cohort specific covariates and technical covariates as appropriate
#### please refer to the analysis plan!

#### models for cord blood! Please add optional variables as appropriate
#formA <- formula("B ~ Y + GA_measure + Fasted + T1 + T2 + S1")
#formB <- formula("B ~ Y + GA_measure + Fasted + Sex + T1 + T2 + S1 + C1 + C2 + C3 + C4 + C5 + C6")
#formC <- formula("B ~ Y + GA_measure + Fasted + Sex+ GA_birth + Parity + Mat_age + T1 + T2+ S1")
#formD <- formula("B ~ Y + GA_measure + Fasted +Sex+ GA_birth + Parity + Mat_age + C1 + C2 + C3 + C4 + C5 + C6 + T1 +
T2+ S1")

### models for non-cord blood! Please add optional variables as appropriate
formA <- formula("B ~ Y + GA_measure + Fasted + Child_age + T1 + T2 + S1 ")
formB <- formula("B ~ Y + GA_measure + Fasted + Child_age + Sex + T1 + T2 + S1 + C1 + C2 + C3 + C4 + C5 + C6")
formC <- formula("B ~ Y + GA_measure + Fasted + Child_age + Sex+ GA_birth + Parity + Mat_age + Child_age + T1 + T2+
S1")
formD <- formula("B ~ Y + GA_measure + Fasted + Child_age + Sex+ GA_birth + Parity + Mat_age + Child_age + C1 + C2 + C3
+ C4 + C5 + C6 + T1 + T2+ S1")

#####
#### Required packages #####

library(MASS) # rlm function for robust linear regression
library(sandwich) #Huber's estimation of the standard error
library(lmtest) # to use coeftest
library(R.utils) # for gzip
library(data.table)

# !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
#####
### all users please read carefully!
# !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
library(future.apply) # for future_lapply function; which is compatible with most back-ends (windows, linux en most
clusters)
# use function availableCores() to check how many cores you have on your machine

availableCores()

```

```
##### WINDOWS USERS BEWARE!! READ THIS THOROUGHLY!!
# You can do the EWAS serial, on one processor core, or parallel, dividing the work over multiple processor cores
# BUT WINDOWS will make a copy of your methylation and phenotype data for each processor core activated!!!
# IF you have enough RAM, dont worry, but if you are working on a small desktop/laptop activating
# to many processor cores might crash your R session!!

# what is your available RAM?
memory.limit()
# what are you currently using?
memory.size()
# now make an informed decision on the number of cores you wish to use (e.g. how much duplications of RAM you can
handle: beware running other programs at the same time also costs RAM)
# if none choose: 1 for "workers" below; if enough RAM is available choose more!

# windows users use this line of code below
## LINUX/UNIX USER DELETE THIS 1 LINE OF CODE BELOW:
plan(multiprocess, workers = 2)

# UNIX/LINUX system:
# to "parallelize" the running of the statistical tests over multiple processor/nodes on a LINUX system
# choose your number of cores: example 2, see below:

# WINDOWS USERS DELETE THIS ONE LINE OF CODE
plan(multicore, workers = 2)
```

```
##### Linux AND Windows: please continue below with the EWAS
#####
#####
#####
### formulating the EWAS function
```

```
### Load into memory:
glucoseEWAS <- function(methcol, meth_matrix, P) {
  if ( identical(as.character(P$ID), colnames(meth_matrix)) ) { # check if samples in meth_matrix and P (phenotypes) are
identical
```

```
  P$B <- meth_matrix[methcol,]
```

```
  cflm = tryCatch({
    mod = rlm(formA, data=P, maxit=200)
    cf = coeftest(mod, vcov=vcovHC(mod, type="HCO")) #HCO suitable in samples >N=100
    N <- length(mod$fitted.values)
    cf <- cf[2, c("Estimate", "Std. Error", "z value", "Pr(>|z|)")] #
    cf <- append(cf, N); names(cf)[4] <- "N"
    names(cf) <- c("Estimate_mA", "SE_mA", "Zval_mA", "P_mA", "N_mA")
    cf
  }, error=function(err){
    cf <- c(NA, NA, NA, NA, NA)
    names(cf) <- c("Estimate_mA", "SE_mA", "Zval_mA", "P_mA", "N_mA")
    cf
  })
```

```
  cflm2 = tryCatch({
    mod = rlm(formB, data=P, maxit=200)
    cf = coeftest(mod, vcov=vcovHC(mod, type="HCO")) #HCO
    N <- length(mod$fitted.values)
    cf <- cf[2, c("Estimate", "Std. Error", "z value", "Pr(>|z|)")]
    cf <- append(cf, N); names(cf)[4] <- "N"
    names(cf) <- c("Estimate_mB", "SE_mB", "Zval_mB", "P_mB", "N_mB")
    cf
  })
```



```

}, error=function(err){
  cf <- c(NaN, NaN, NaN, NaN, NaN)
  names(cf) <- c("Estimate_mB", "SE_mB", "Zval_mB", "P_mB", "N_mB")
  cf
})

cflm3 = tryCatch({
  mod = rlm(formC, data=P, maxit=200)
  cf = coeftest(mod, vcov=vcovHC(mod, type="HCO")) #HCO
  N <- length(mod$fitted.values)
  cf <- cf[2, c("Estimate", "Std. Error", "z value", "Pr(>|z|)")]
  cf <- append(cf, N); names(cf)[4] <- "N"
  names(cf) <- c("Estimate_mC", "SE_mC", "Zval_mC", "P_mC", "N_mC")
  cf
}, error=function(err){
  cf <- c(NaN, NaN, NaN, NaN, NaN)
  names(cf) <- c("Estimate_mC", "SE_mC", "Zval_mC", "P_mC", "N_mC")
  cf
})

cflm4 = tryCatch({
  mod = rlm(formD, data=P, maxit=200)
  cf = coeftest(mod, vcov=vcovHC(mod, type="HCO")) #HCO
  N <- length(mod$fitted.values)
  cf <- cf[2, c("Estimate", "Std. Error", "z value", "Pr(>|z|)")]
  cf <- append(cf, N); names(cf)[4] <- "N"
  names(cf) <- c("Estimate_mD", "SE_mD", "Zval_mD", "P_mD", "N_mD")
  cf
}, error=function(err){
  cf <- c(NaN, NaN, NaN, NaN, NaN)
  names(cf) <- c("Estimate_mD", "SE_mD", "Zval_mD", "P_mD", "N_mD")
  cf
})
# stick it all together
outcomes <- c(cflm, cflm2, cflm3, cflm4)
return(outcomes)
} else { stop() } # if ordering of meth_matrix en phenotype file P is NOT identical quit and throw error out of function!
}

```

Actually run the EWAS

```
#####
#####
#####
```

some users encounter an error when running the future.apply function below for the first time.

The error states that the RAM (usually 500MiB) available for "globals" (the data to be duplicated across processor cores) is exceeded.

In this case expand the amount of RAM allowed for a global using the following code

example for 3Gb (3000Mb) of available RAM

options(future.globals.maxSize= 3000*1024^2)

test these CpGs: here all CpGs in the methylation file

```
CpGs <- rownames(Betamatrix)
```

do the EWAS using future_apply(), transpose result and make it a data.frame for easy handling.

```
results <- as.data.frame(t(future_sapply(CpGs, function(x) { glucoseEWAS(methcol = x, meth_matrix = Betamatrix, P=Data ) }
)))
```

```
results$probeID <- rownames(results) # add CpG identifiers to results as separate column
```

```
rownames(results) <- NULL # remove rownames (CpG identifiers) after storing them in separate column
```

create name for results file

```
name_results_file <- paste(paste(paste(paste(Cohort,"EWAS",EWAStype,sep="_"),Moment,sep="_"),
Datestamp,sep="_"),".txt",sep=".")
```

```
# write away tab delimited descriptives table
```

```
write.table(results, file=file.path(OutputDir,name_results_file), col.names=T, row.names=F, quote=F, sep="\t")
```

```
# gzip results file:
```

```
gzip( file=file.path(OutputDir,name_results_file) )
```

```
##### write lambdas of all analyses to separate file:
```

```
## Save model number, sample size & lambda to text file:
```

```
TotalSampleSize <- dim(Data)[1]
```

```
lambdas <- apply(results[,c("P_mA","P_mB","P_mC","P_mD")],2,function(x){ Lambda(x) } )
```

```
tab <- c(Cohort,EWAStype,Moment, TotalSampleSize, lambdas)
```

```
names(tab) <- c("Cohort","EWAS","Moment","Nmax","lambda_mA","lambda_mB","lambda_mC","lambda_mD")
```

```
tab <- t(as.data.frame(tab))
```

```
rownames(tab) <- NULL
```

```
### !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
```

```
### look at your lambdas: are they ~OK? If not please check your data/phenotypes
```

```
tab
```

```
# create name for results file
```

```
name_tab <- paste(paste(paste(paste(Cohort,"Lambdas",EWAStype,sep="_"),Moment,sep="_"),
```

```
Datestamp,sep="_"),".txt",sep=".")
```

```
# write away lambdas
```

```
write.table(tab,file=file.path(OutputDir,name_tab),row.names=F, quote=F, sep="\t")
```

```
##### Thank you for your effort!
```