

1 **Severe preeclampsia is not associated with significant DNA methylation**
2 **changes but cell proportion changes in the cord blood**
3 **- caution on the importance of confounding adjustment**

4
5 Wenting Liu^{1†}, Xiaotong Yang^{1†}, Zhixin Mao¹, Yuheng Du¹, Cameron Lassiter², Fadhl M.

6 AlAkwa³, Paula A Benny², Lana X Garmire^{1*}

7 1. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann
8 Arbor, MI

9 2. University of Hawaii Cancer Center, Epidemiology, Honolulu, HI

10 3. Department of Neurology, University of Michigan, Ann Arbor, MI

11 † These authors contributed equally to the work

12 * corresponding author

13

14 **Abstract**

15 Epigenome-wide DNA methylation analysis (EWAS) is an important approach to identify
16 biomarkers for early disease detection and prognosis prediction, yet its results could be
17 confounded by other factors such as cell-type heterogeneity and patient characteristics. In this
18 study, we address the importance of confounding adjustment by examining DNA methylation
19 patterns in cord blood exposed to severe preeclampsia (PE), a prevalent and potentially fatal
20 pregnancy complication. Without such adjustment, a misleading global hypomethylation pattern
21 is obtained. However, after adjusting cell type proportions and patient clinical characteristics,

22 most of the so-called significant CpG methylation changes associated with severe PE disappear.
23 Rather, the major effect of PE on cord blood is through the proportion changes in different cell
24 types. These results are validated using a previously published cord blood DNA methylation
25 dataset, where global hypomethylation pattern was also wrongfully obtained without confounding
26 adjustment. Additionally, several cell types significantly change as gestation progress (eg.
27 granulocyte, nRBC, CD4T, and B cells), further confirming the importance of cell type
28 adjustment in EWAS study of cord blood tissues. Our study urges the community to perform
29 confounding adjustments in EWAS studies, based on cell type heterogeneity and other patient
30 characteristics.

31

32 **Introduction**

33 DNA methylation is a type of epigenetic modification that plays a crucial role in regulating gene
34 expression and maintaining genome stability. It involves the addition of a methyl group to the
35 cytosine base of DNA, typically at the CpG dinucleotide sites¹. Methylation at these sites can
36 affect gene expression by altering the accessibility of DNA to transcription factors and other
37 proteins. DNA methylation profile can be modified by various factors, including aging, diseases,
38 and environmental changes^{2,3}. Previous studies on DNA methylation have contributed to
39 biomarker identifications for risk prediction, early detection, and prognosis tracking of various
40 diseases⁴⁻⁶. Differential methylation analysis, or epigenome-wide association (EWAS) study, is a
41 key computational process to identify disease-associated DNA methylation markers⁷. However,
42 rigorous statistical and bioinformatics approaches remain a central issue to draw unbiased
43 conclusions in these studies. Recently, some researchers have started to raise the awareness of
44 confounding adjustment for DNA methylation results, including clinical characteristics and
45 heterogeneity of cell types within sample tissues⁸. Failure to properly adjust for these confounders
46 may lead to biased and inaccurate results.

47

48 In this study, we alert the community to the importance of confounder adjustment, using the case
49 study of the DNA methylation change in cord blood samples from babies born of severe
50 preeclampsia (PE). PE is one of the leading causes of maternal and prenatal morbidities and
51 mortalities, affecting 2-8% of pregnancies globally and around 3.1% in the US^{9,10}. PE is
52 characterized by new-onset hypertension with proteinuria or one/more adverse conditions after 20
53 weeks of gestation¹¹. It can lead to severe outcomes including renal failure, seizure, multiorgan
54 dysfunction, and stroke in mothers; as well as intrauterine growth restriction (IUGR) and
55 premature delivery of the fetuses. Based on blood pressure, clinical findings, and degree of
56 proteinuria, PE can also be classified into severe PE or mild PE. Severe PE poses a greater risk to
57 maternal and fetal health and may involve different pathways than mild PE of similar onset
58 time¹². It is imperative to investigate the molecular mechanisms of severe preeclampsia.

59

60 Many previous epigenetic-wide studies aim to find biomarkers of PE using complex tissues, such
61 as placenta tissues^{13,14}, maternal gestational blood¹⁵, and fetal chorioamniotic membrane¹⁶.
62 Multiple previous studies looked into the effect of PE on cord blood¹⁷⁻²⁰. Yet most of these earlier
63 studies didn't adjust for potential confounding factors using either cell-type heterogeneity¹⁷⁻¹⁹ or
64 clinical variables, such as gestational age another significant confounder for pregnancy diseases
65 including PE²⁰. Tissues like placentas and cord blood consist of many diverse cell types, each
66 with a distinct epigenetic profile as defined^{21,22}. Thus the varying cell types in each sample can
67 affect the overall DNA methylation profile at the bulk level. It is therefore essential to account for
68 such heterogeneity, to improve the accuracy and sensitivity and avoid wrongful conclusions of
69 EWAS biomarker detection. Particularly, to ensure that any differences in DNA methylation are
70 due to confounding factors, the analysis needs to be adjusted for cell proportions. Moreover, if the
71 essential clinical data such as gestational ages are available (as they should be), then the analysis
72 needs to be adjusted for the important clinical variables as well. We show the practice of doing

73 such cell type and clinical confounding adjustment using the case study of cord blood DNA
74 methylation analysis. We warn the community of potential significant harm otherwise.

75

76 **Materials and Methods**

77 **Study cohort**

78 The umbilical cord whole blood DNA samples were obtained from Hawaii Biorepository (HiBR).
79 The HiBR collected placenta, maternal, and cord blood samples from deliveries at Kapiolani
80 Women and Children's Hospital from 2006 to 2013. It is one of the largest research tissue
81 repositories in the Pacific region, containing specimens from more than 9250 mother-child pairs
82 at the time of sample collection. Umbilical cord samples were collected immediately after
83 delivery. Severe PE was characterized by OBGYNs at Kapiolani Medical Center as sustained
84 pregnancy-induced hypertension(systolic/diastolic blood pressure $\geq 140/90$) with urine protein
85 and/or organ dysfunction. We originally collected 63 samples. The demographic and clinical
86 information of the patients was collected and analyzed to identify any potential confounding
87 effects. Data usage was approved by IRB #CHS23976.

88

89 **Sample preparation**

90 Umbilical cord blood samples were collected in the operating room immediately after delivery.
91 To prepare for DNA extraction, we first added three volumes of RBC Lysis Solution to one
92 volume of clotted blood, which is then vortexed and incubated on a shaker for 15 minutes at room
93 temperature. The sample was then centrifuged to pellet white blood cells and clot particulates,
94 and the supernatant is carefully poured into a waste bucket. The pellet was resuspended in an
95 additional volume of RBC Lysis Solution and incubated again for 15 minutes. After another
96 centrifugation step, the supernatant was carefully removed, leaving behind 200 μ L of residual
97 liquid. The pellet was then vigorously resuspended in the residual liquid before being combined

98 with a master mix of Cell Lysis and Proteinase K Solution. The mixture was vortexed and
99 incubated at 55°C until homogenous, with intermittent vortexing to facilitate digestion. Once
100 homogenous, the samples were subjected to DNA purification on the Autopure Machine
101 following the manufacturer's instructions.

102

103 **DNA extraction and methylation profiling**

104 DNA was extracted from prepared cord blood samples using AllPrep DNA/RNA/Protein Mini
105 Kit (Qiagen, USA) according to the manufacturer's instructions by HiBR. We obtained pre-
106 extracted genomic DNA of whole cord blood samples from the HiBR and conducted DNA
107 Illumina EPIC Beadchip assays through the University of Hawaii Cancer Center Genomics Core.

108

109 **DNA methylation data pre-processing and quality control**

110 We used the R package “ChAMP” for data pre-processing (**Supple Fig. 1**). We first filtered
111 probes using the following criteria sequentially: (1) removing probes with a detection p-value
112 above 0.01 (7,941 probes); (2) removing probes with a bead count <3 in at least 5% of samples
113 (27,731 probes); (3) removing probes with no matched CpG sites (2,673 probes); (4) removing
114 probes that align to multiple locations (8,248 probes). During the quality control step, we
115 removed 1 control sample with a distinct beta density distribution (**Supple Fig. 2A, 2B**). We
116 normalized the remaining samples using BMIQ methods²³, and corrected for batch effects using
117 the ComBat algorithm embedded in “ChAMP”. We used the singular vector decomposition
118 (SVD) heatmap to verify the effectiveness of batch removal (**Supple Fig. 2C, 2D**). The
119 preprocessed data matrix contains 62 samples and 819,325 probes. We converted the original
120 methylation intensity (beta) to M-values using “beta2m” function from “lumi” package to reduce
121 heteroskedasticity²⁴, where M-values are defined as the log₂ ratio of the beta value of each probe.

122

123 **Cell-type deconvolution in umbilical cord whole blood (CB)**

124 Bulk-level DNA in umbilical cord whole blood (CB) includes at least 7 most common blood cell
125 types: neutrophils, B cells, CD4T, CD8T, monocytes, natural killer cells (NK), and nucleated red
126 blood cells (nRBC). Each sample may have different compositions of the cell types above, thus
127 needing deconvolution. For this, we combined two methylation reference matrices of the cord
128 blood sample for better cell-proportion estimation from the whole blood samples. We extracted a
129 reference panel of B cell, CD4T, CD8T, monocyte, natural killer cell, and granulocyte from Lin
130 et al.²⁵ given the reported high quality of the reference samples. We also added the DNA
131 methylation profiles of nucleated red blood cells (nRBC) from Bakulski et al²⁶ to be part of the
132 whole blood DNA methylation reference. We used the “combineArrays” function in the “minfi”
133 R package to combine two data sets and rescaled them using the “BMIQ” method in the
134 “wateRmelon” R package. We used pairwise t-tests with Bonferroni adjustment (threshold of 1E-
135 8) to identify significant CpGs among the 7 cell types in each pair as cell-specific markers,
136 similar to others²⁵. This process yielded 151,794 cell-type specific CpG biomarkers for the DNA
137 methylation reference matrix for deconvolution. We uploaded the new whole blood CpG
138 reference dataset in the EpiDISH package.

139

140 We used principal component analysis (PCA) to check the quality of these new sets of markers.
141 We applied the selected markers above to estimate the cell type proportions in each sample, using
142 the reference-based cell-type deconvolution algorithm Constrained Projection (CP) from the
143 “EpiDISH” package²⁷. We used the resulting cell type proportions to adjust for confounding
144 effects in the epigenome-wide association analysis.

145

146 **Clinical confounders and source of variance analysis**

147 We retrieved a total of 6 commonly reported clinical variables in cord blood EWAS studies from
148 the biobank, including maternal age, ethnicity, parity, BMI, delivery gestational age, and smoking

149 status²⁰. We imputed 3 samples (including 1 severe preeclampsia and 2 controls) with missing
150 BMI using mean values of each sample group. We performed the source of variance (SOV)
151 analysis on the 6 clinical variables and previously estimated sample cell proportion to identify
152 important confounding variables that need to be adjusted, as done before^{28,29}. SOV analysis
153 calculates F-statistics that can be used to identify and quantify the contribution of different factors
154 to the total variance in a dataset. We adjusted the variables with F-statistics larger than 1 (the
155 error value) as confounding variables.

156

157 **CpG-level epigenetic-wide association analysis (EWAS)**

158 We calculated the differentially methylated probes (DMP) between severe PE cases and controls
159 using moderated t-test with Benjamini-Hochberg (BH) adjustment (threshold of 0.05). We
160 included study participants' gestational age (GA), BMI, parity status, ethnicity, and methylation-
161 derived cell compositions in the linear model to remove the confounding effects, and compared
162 the result with that without confounding adjustment using “limma” package³⁰. We defined
163 hypermethylated CpGs as significant CpGs with positive log₂-transformed fold change (logFC)
164 and hypomethylated CpGs as significant CpGs with negative logFC, respectively. We used
165 volcano plots to illustrate the global DNA methylation changes between the cases and controls.

166

167 **Gene-level EWAS**

168 We further examined the methylation signal difference between severe PE and controls at gene
169 and pathway levels. We annotated the CpGs³¹, selected those located on the promoter region and
170 aggregated the methylation signals of all CpGs within a gene promoter by taking the geometric
171 mean. Then we compared the aggregated methylation signals between severe PE cases and
172 control groups, using the moderated t-tests with BH adjustment. Additionally, we looked for
173 pathways associated with promotor region methylation differences using the R package
174 pathifier³². The pathifier algorithm calculates a pathway deregulation score (PDS) for each

175 sample and each pathway. We compared the pathway PDS scores in case and control samples by
176 moderated t-tests with Benjamini-Hochberg FDR adjustment (threshold p-values 0.05).

177

178 **Software Usage and Code Availability**

179 All analysis was done using R 4.1.2³³. Specially, we used “ChAMP” (version 2.24.0) for data
180 preparation³⁴, “limma” for differential analysis³⁰, “EpiDISH” (version 2.10.0) for cell-type
181 deconvolution²⁷, and “IlluminaHumanMethylationEPICanno.ilm10b4.hg19” for data
182 annotation³¹. All codes are available at https://github.com/lanagarmire/CB_DNAm_PE.

183

184 **Results**

185 **Overview of Study Design and Cohort Characteristics**

186 The overview of the study design is illustrated in **Fig. 1**. We obtained whole cord blood samples
187 from 24 severe PE cases and 39 controls collected at Hawaii Biorepository (HiBR) from January
188 2006 and June 2013. The maternal demographic and clinical characteristics are shown in **Table 1**.
189 There is no significant difference ($P > 0.05$) in maternal age, parity, BMI, ethnicity, and smoking
190 status. Noticeably, the severe PE cases have earlier delivery gestational age compared to healthy
191 controls ($P = 3.66E-06$), as the clinical management of severe PE usually often demands early
192 delivery to avoid severe maternal morbidities. We obtained pre-extracted genomic DNA of whole
193 cord blood samples from the HiBR and conducted DNA Illumina EPIC Beadchip assays through
194 the University of Hawaii Cancer Center Genomics Core. We conducted bioinformatics pre-
195 processing of the DNA methylation following standard steps in R package “ChAMP” (**Supple**
196 **Fig. 1**). Briefly, we first filtered out probes of bad quality, then examined the methylation level
197 distribution of each sample and removed 1 control sample with abnormal distribution (**Supple**
198 **Fig. 2A, 2B**). Lastly, we normalized the data and removed the batch effect between arrays
199 (**Supple Fig. 2C, 2D**). The remaining preprocessed data matrix contains 62 samples and 819,325

200 probes (see **Methods**).

201

202 **Cell-type deconvolution in cord blood samples**

203 Cord blood is a complex tissue with various cell types, including granulocyte, B cell, CD4T,
204 CD8T, monocyte, natural killer, and nucleated red blood cells (nRBC), each with a unique DNA
205 methylation profile. Cell type heterogeneity in the cord blood could significantly confound the
206 phenotypes of interest, affecting the differential methylation analysis results. Previous studies
207 showed that cell type heterogeneity contributes to much more variation in DNA methylation in
208 various tissues, compared to ethnicity, sex, age, and even phenotypes of interest^{8,35,36}. If left
209 unadjusted, such variation from cell type difference will result in biased and even misleading
210 results in the differential methylation analysis associated with the phenotype. Therefore adjusting
211 for cell type heterogeneity is an essential step in epigenetic-wide association studies (EWAS) of
212 the bulk-level data.

213 To adjust for cell-type heterogeneity, we first estimated the cell-type proportions in the cord
214 blood samples by performing cell-type deconvolution. Unfortunately, not all cord blood cell-type
215 specific references contain nucleated red blood cells (nRBC), whose amount decreases as
216 gestation progresses and can be non-trivial in severe PE where most deliveries are preterm³⁷⁻³⁹.

217 To address this limitation, we created a new whole cord blood DNA methylation reference by
218 merging two existing references: the EPIC array reference panel by Lin et al.²⁵ and the 450K
219 array reference by Bakulski et al.²⁶. Lin's work demonstrated a superior quality of cell type-
220 specific markers, as evidenced by better cluster separation in PCA analysis, compared to
221 Bakulski's data (**Fig. 2A, 2B**). However, Lin's reference lacked the crucial nRBC cell type.

222 Therefore, we combined the methylation levels of B cell, CD4T, CD8T, monocyte, natural killer
223 cell, and granulocyte in Lin's reference and nRBC from Bakulski's reference. We processed the
224 CpG markers similar to the original studies (see **Methods**). As a result, we identified 151,794
225 differentially methylated CpGs as high-quality markers for the whole cord blood. The PCA result

226 of the combined reference panel confirmed better separation of cell types than those from both
227 original studies (**Fig. 2C**).

228

229 **Association between cord blood cell type and severe PE**

230 We next aimed to determine the significance of cell type heterogeneity as a potential confounder
231 that should be adjusted for in the study. To learn the association between severe PE and cord
232 blood cell composition, we first performed cell-type deconvolution on cord blood samples using
233 the new combined reference and Houseman's constrained projection(CP) deconvolution
234 algorithm⁴⁰. Estimated cell-type proportions show large variations across samples, especially for
235 granulocyte, nRBC, and CD4T cells (**Fig. 3A**). Granulocyte proportions are significantly lower
236 ($\beta=-0.10$, $P=3.44e-7$) in the severe PE group compared to the control group, while B cell
237 ($\beta=0.0085$, $P=0.011$), nRBC cell ($\beta=0.062$, $P=1.67e-5$) and CD8T cell ($\beta=0.014$, $P=0.0084$)
238 proportions are significantly higher in cases (**Fig. 3B**).

239

240 However, the apparent differences in cell proportions in cases vs controls could be very well due
241 to other reasons (eg. gestational age) rather than severe PE. To confirm this speculation, we
242 conducted the source of variance (SOV) analysis of cell type compositions on the clinical
243 variables and ranked them by F-statistics. A variable with F-statistics bigger than 1 (the error
244 term) is considered a significant contributor to cell proportion variations. As shown in **Fig. 3C**, in
245 addition to severe PE, gestational age, maternal BMI, and ethnicity also contribute significantly to
246 cell type heterogeneity. Gestational age and maternal BMI rank higher than severe PE. With such
247 caution, we adjusted confounding by linearly regressing cell type proportions over severe PE and
248 other covariate factors, including gestational age, maternal age, ethnicity, BMI, and smoking
249 status. We plot the cell proportions in severe PE vs. controls, post-adjustment by other clinical
250 variables (**Fig. 3D**). The previously observed differences in granulocytes, B cells, and nRBCs
251 now all disappear. CD8T proportions, however, continue to be significantly higher in severe PE

252 cases ($\beta=0.02$, $P=0.0068$). Interestingly, monocyte proportions, which were not significantly
253 different before adjustment between cases and controls (**Fig. 3B**), now become significantly
254 associated with severe PE ($\beta=0.016$, $p=0.028$) after confounding adjustment. The detailed linear
255 regression results of cell proportion on clinical data can be found in **Supple Table 1**. In all, these
256 results show that cell type proportions vary among newborns and it is important to adjust for
257 potential confounding before interrogating the association with severe PE.

258

259 **Confounding adjustment drastically affects EWAS results**

260 Considering the current state of most EWAS studies which often overlook the adjustment of cell
261 types and other clinical covariates (such as gestational age) within their samples, we further
262 investigated the impact of these factors on the differential methylation analysis. We conducted the
263 source of variance (SOV) analysis of the DNA methylation matrix on cell type proportion and
264 clinical variables. Strikingly, all cell-type composition variables show the strongest and most
265 dominant explanation power of variation in the methylation data (**Fig. 4A**), ranking even higher
266 than the severe PE condition. After the cell proportions, severe PE case/control, gestational age,
267 maternal age, parity, and ethnicity also have larger F-statistics than the error term ($F\text{-statistics}=1$),
268 in descending order. Therefore, in the downstream analysis of differentially methylated CpGs, we
269 adjust these variables for confounding effects.

270

271 As a comparison, we first conducted differential methylation analysis on severe PE without
272 adjusting for any clinical confounders or cell-type proportions. The analysis reveals a global
273 hypomethylation pattern (**Fig. 4B**). We identified 229,730 differentially methylated CpGs with
274 adjusted p-values less than 0.05. Among these CpGs, 184,102 exhibited hypomethylation, while
275 45,628 displayed hypermethylation. However, when we redid the differential methylation
276 analysis after adjusting for cell type heterogeneity and patient characteristics, all the CpGs
277 differentially methylated above are no longer significant, except for a single CpG site labeled as

278 cg20135196 on Chromosome 6, with FDR=0.014 (**Fig. 4C**). Interestingly, this locus is rarely
279 reported before. A total of 11 genes (**Supple Table 8**) are identified as the nearest genes to this
280 locus on both sides. ZNF 184 is the closest gene by distance. It is a zinc finger protein involved in
281 transcription regulation. A series of genes of nucleosome components are located on the opposite
282 side of ZNF 184.

283

284 Additionally, we extended the differential methylation analysis to the gene level. We aggregated
285 the CpGs located on gene promoters as the representation of promoter-level methylation (see
286 Methods). Before adjusting for confounders, we detected 4,767 differentially methylated genes.
287 However, upon adjusting for both clinical variables and cell types, none of the genes exhibited
288 statistical significance. Similarly, we conducted a differential methylation analysis at the pathway
289 level, employing the Pathifier algorithm (see **Methods**). Before the confounding adjustment, we
290 detected 200 significant pathways. Nevertheless, after accounting for confounders, none of the
291 pathways remained significant. In conclusion, the most concerning finding is that the observed
292 DNA methylation variation among the whole cord blood samples is primarily associated with cell
293 type heterogeneity rather than severe PE.

294

295 To confirm the significant influence of confounders on EWAS associated with severe PE, we re-
296 analyzed the Illumina 450k DNA methylation data from Ching et al, which were obtained from
297 different samples¹⁶. Using the original analysis pipeline that did not consider clinical
298 confounding, we reproduced the differential methylation results earlier, which reported 68,458
299 significant CpGs (**Supple Fig. 3A**). Subsequently, we estimated the cell type proportions using
300 Houseman's CP algorithm and the new combined cord blood reference reported in this study. We
301 then conducted the SOV analysis by considering cell proportions and clinical variables. We
302 identified cell type proportions, gestational age, maternal age and PE as significant confounders,
303 as they have F-statistics > 1 (**Supple Fig. 3B**). Consistent with the observation on the EPIC array

304 cord blood dataset earlier, the dataset by Ching et al. yields no significant CpG (**Supple Fig. 3C**)
305 once adjusted for these confounders. In summary, using two cord blood datasets, we
306 demonstrated that confounding effects from clinical variables and cell type heterogeneity are
307 common challenges that need to be addressed by EWAS analysis in cord blood.

308

309 **Association between cord blood cell type and gestational age**

310 Our earlier analysis shows that estimated cell proportions in cord blood are correlated with
311 gestational age (**Fig. 3C**). We thus conducted a more in-depth analysis. The most noticeable
312 correlation comes from granulocytes, whose proportions increase from around 25% in week 32 to
313 over 50% in week 40, with $P=0.00015$ (**Fig. 5A**). The proportions of monocytes also significantly
314 increase as the gestation progresses, after adjusting for other variables ($p=0.0051$, **Fig. 5B**). On
315 the contrary, nRBC, natural killer, and B cell significantly decrease along the gestation ($p =$
316 $6.62e-6$; $p=0.0054$; $p=0.0087$, **Fig. 5C-E**). These trends of changes are maintained when the
317 samples are stratified into case and control groups (**Supple Fig. 4**), without significant interaction
318 between each cell proportion and case/control labels when regressing them on gestational ages (y-
319 values).

320

321 Furthermore, we validated the trends of cell type proportion through gestational age using another
322 public cord blood peripheral blood mononuclear cell (PBMC) Illumina HumanMethylation450
323 BeadChip methylation dataset (GEO accession ID: [GSE110828](#)), which comprises 20 PE cases
324 and 90 non-PE controls⁴¹. Both the case and control groups include large percentages of preterm
325 samples, with the delivery gestational ages ranging from 26.14 to 38.14 in cases and 23.00 to
326 41.29 in controls. We deconvoluted the PBMC cell types with Houseman's CP method using the
327 new reference data we created here without the granulocytes due to their absence in PBMC. We
328 performed linear regression of gestational age over the cell type, severe PE and their interaction
329 terms similar to those in **Fig. 5** We observe the same increasing trend for monocytes, and the

330 same decreasing trend for nRBC cell, natural killer cell and B cell (**Supple Fig. 5**). Again, none
331 of the interaction terms between severe PE and gestational age turned out to be significant,
332 suggesting that their effects on cell proportion changes are independent.

333

334 **Discussion**

335 In this study we showed and validated the importance of confounding adjustment in bulk DNA
336 methylation analysis on the EWAS result, using two datasets on cord blood samples exposed to
337 severe PE. As the most significant confounder among clinical variables, gestational age affects
338 various blood cell type proportions (eg. granulocyte, nRBC, CD4T and B cells). We showed that
339 many CpGs, genes and pathways will be artificially and wrongfully detected if we do not adjust
340 for the profound confounding due to cell type composition and other clinical variables such as
341 gestational age. Despite the lack of CpG changes to severe PE, many cell types' proportion
342 changes do differ between severe PE vs. controls. Thus, we conclude the major effect of PE on
343 cord blood is not through CpG methylation changes, but the proportion changes in different cell
344 types.

345

346 Among all confounders in EWAS analysis, cell type heterogeneity is one of the most common
347 and important confounders faced by researchers. Over the years, various cell type estimation
348 methods for bulk-level epigenetic data have been developed^{35,40}, which made the assessment of
349 cell type confounding effects possible. In 2013, Liu et al. first reported a large reduction in
350 differentially methylated probes related to Rheumatoid Arthritis after adjusting for cell type
351 composition in whole blood⁴². Some later studies confirmed the effect of cell type heterogeneity
352 on EWAS research in other tissues such as breast tissue, saliva, and placenta tissue^{8,43,44}. Unlike
353 many other diseases, pregnancy complications in the form of severe PE are unique in terms of
354 confounding. Many fetuses have to be delivered preterm, leading to the case group with smaller

355 GA compared to the healthy full-term controls. Therefore, gestational age and PE are two
356 conditions which go hand-in-hand, and gestational age is a major confounder for studying severe
357 PE. More importantly, the cell proportions in severe PE case groups are also affected by the
358 disease itself, making them critical confounders in EWAS studies. Although several previous
359 studies aimed to identify PE-related epigenetic biomarkers using cord blood samples¹⁷⁻²⁰, the
360 importance of adjusting for cell type heterogeneity was mostly (3 out of 4) overlooked. This
361 could have led to biased results or even false biomarkers. This could also be one of the reasons
362 for the inconsistency in these previous studies.

363

364 Furthermore, we noticed consistent associations between cell type compositions in cord blood
365 and gestational age, both in severe PE cases and controls. Granulocyte proportion showed the
366 strongest positive association with gestational age, agreeing with the previous findings that
367 granulocyte in the fetus increases drastically in the last trimester of pregnancy⁴¹. The estimated
368 proportion of nRBC in our study decreases as gestational age increases, also consistent with
369 previous findings³⁷⁻³⁹. Additionally, elevated nRBC was also found in preterm infants and infants
370 with lower birth weight⁴⁵, providing additional supporting evidence to our finding. We also
371 observed a significant increase in monocytes and significant decreases in B cells and NK cells as
372 gestational age increases. Taken together, these findings probe into the dynamic nature of cell
373 type composition in cord blood during gestation. Future EWAS studies utilizing cord blood
374 samples collected at different gestational ages should carefully adjust for cell type heterogeneity.

375

376 **Conclusion**

377 In summary, we found the lack of evidence for significant CpG methylation changes in EWAS
378 analysis in association with severe PE, after adjusting for cell type heterogeneity and clinical
379 variables such as gestational age. Wrongful and artificial CpG methylation biomarkers would

380 have been obtained without such adjustment. Instead of altering CpG methylations, severe PE is
381 associated with significant changes in several cell proportions in the cord blood. We also showed
382 that many cell types' proportions change drastically during pregnancy.

383

384 **Author Contributions**

385 LXG conceived this project and supervised the study. WL and XY contributed equally to data
386 analysis, result generation, and manuscript writing. ZM and YD assisted in the data processing.
387 CL, FMA, and PAB contributed to sample collection, coordination and the experimental design
388 of the DNA methylation. All authors have read, revised and approved the manuscript.

389 **Funding**

390 This research was supported by grants by NIH/NIGMS, R01 LM012373 and R01 LM012907
391 awarded by NLM, and R01 HD084633 awarded by NICHD to L.X. Garmire, and T32GM141746
392 to X.T Yang

393

394 **Acknowledgment**

395 We thank the Genomics Shared Resources of the University of Hawaii Cancer Center for
396 performing the methylation assays.

397

398 **Competing Interests**

399 The authors declare no conflict of interest.

400

401 **Materials & Correspondence**

402 Correspondence to Lana X Garmire

403

404 **Reference**

- 405 1. Moore LD, Le T, Fan G. DNA Methylation and Its Basic Function.
406 *Neuropsychopharmacol.* 2013;38(1):23-38. doi:10.1038/npp.2012.112
- 407 2. Martin EM, Fry RC. Environmental Influences on the Epigenome: Exposure-
408 Associated DNA Methylation in Human Populations. *Annu Rev Public Health.*
409 2018;39(1):309-333. doi:10.1146/annurev-publhealth-040617-014629
- 410 3. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock
411 theory of ageing. *Nat Rev Genet.* 2018;19(6):371-384. doi:10.1038/s41576-018-
412 0004-3
- 413 4. Locke WJ, Guanzon D, Ma C, et al. DNA Methylation Cancer Biomarkers:
414 Translation to the Clinic. *Front Genet.* 2019;10:1150. doi:10.3389/fgene.2019.01150
- 415 5. Klutstein M, Nejman D, Greenfield R, Cedar H. DNA Methylation in Cancer and
416 Aging. *Cancer Research.* 2016;76(12):3446-3450. doi:10.1158/0008-5472.CAN-15-
417 3278
- 418 6. Dor Y, Cedar H. Principles of DNA methylation and their implications for biology
419 and medicine. *The Lancet.* 2018;392(10149):777-786. doi:10.1016/S0140-
420 6736(18)31268-6
- 421 7. Campagna MP, Xavier A, Lechner-Scott J, et al. Epigenome-wide association
422 studies: current knowledge, strategies and recommendations. *Clin Epigenet.*
423 2021;13(1):214. doi:10.1186/s13148-021-01200-8
- 424 8. Qi L, Teschendorff AE. Cell-type heterogeneity: Why we should adjust for it in
425 epigenome and biomarker studies. *Clin Epigenet.* 2022;14(1):31.
426 doi:10.1186/s13148-022-01253-3
- 427 9. Rana S, Lemoine E, Granger JP, Karumanchi SA. Preeclampsia: Pathophysiology,
428 Challenges, and Perspectives. *Circ Res.* 2019;124(7):1094-1112.
429 doi:10.1161/CIRCRESAHA.118.313276
- 430 10. Gestational Hypertension and Preeclampsia: ACOG Practice Bulletin, Number 222.
431 *Obstetrics & Gynecology.* 2020;135(6):e237-e260.
432 doi:10.1097/AOG.0000000000003891
- 433 11. Magee LA, Smith GN, Bloch C, et al. Guideline No. 426: Hypertensive Disorders of
434 Pregnancy: Diagnosis, Prediction, Prevention, and Management. *Journal of*
435 *Obstetrics and Gynaecology Canada.* 2022;44(5):547-571.e1.
436 doi:10.1016/j.jogc.2022.03.002
- 437 12. Roberts JM, Rich-Edwards JW, McElrath TF, Garmire L, Myatt L, for the Global
438 Pregnancy Collaboration. Subtypes of Preeclampsia: Recognition and Determining
439 Clinical Usefulness. *Hypertension.* 2021;77(5):1430-1441.
440 doi:10.1161/HYPERTENSIONAHA.120.14781
- 441 13. Mayne BT, Leemaqz SY, Smith AK, Breen J, Roberts CT, Bianco-Miotto T.
442 Accelerated placental aging in early onset preeclampsia pregnancies identified by
443 DNA methylation. *Epigenomics.* 2017;9(3):279-289. doi:10.2217/epi-2016-0103
- 444 14. Li Y, Cui S, Shi W, et al. Differential placental methylation in preeclampsia, preterm
445 and term pregnancies. *Placenta.* 2020;93:56-63. doi:10.1016/j.placenta.2020.02.009
- 446 15. Mousa AA, Archer KJ, Cappello R, et al. DNA Methylation is Altered in Maternal
447 Blood Vessels of Women With Preeclampsia. *Reprod Sci.* 2012;19(12):1332-1342.
448 doi:10.1177/1933719112450336
- 449 16. Ching T, Song MA, Tiirikainen M, et al. Genome-wide hypermethylation coupled
450 with promoter hypomethylation in the chorioamniotic membranes of early onset pre-

- 451 eclampsia. *MHR: Basic science of reproductive medicine*. 2014;20(9):885-904.
452 doi:10.1093/molehr/gau046
- 453 17. Ching T, Ha J, Song MA, et al. Genome-scale hypomethylation in the cord blood
454 DNAs associated with early onset preeclampsia. *Clin Epigenet*. 2015;7(1):21.
455 doi:10.1186/s13148-015-0052-x
- 456 18. He J, Zhang A, Fang M, et al. Methylation levels at IGF2 and GNAS DMRs in
457 infants born to preeclamptic pregnancies. *BMC Genomics*. 2013;14(1):472.
458 doi:10.1186/1471-2164-14-472
- 459 19. Gao Q, Fan X, Xu T, et al. Promoter methylation changes and vascular dysfunction in
460 pre-eclamptic umbilical vein. *Clin Epigenet*. 2019;11(1):84. doi:10.1186/s13148-019-
461 0685-2
- 462 20. Kazmi N, Sharp GC, Reese SE, et al. Hypertensive Disorders of Pregnancy and DNA
463 Methylation in Newborns: Findings From the Pregnancy and Childhood Epigenetics
464 Consortium. *Hypertension*. 2019;74(2):375-383.
465 doi:10.1161/HYPERTENSIONAHA.119.12634
- 466 21. Yuan V, Hui D, Yin Y, Peñaherrera MS, Beristain AG, Robinson WP. Cell-specific
467 characterization of the placental methylome. *BMC Genomics*. 2021;22(1):6.
468 doi:10.1186/s12864-020-07186-6
- 469 22. Gervin K, Salas LA, Bakulski KM, et al. Systematic evaluation and validation of
470 reference and library selection methods for deconvolution of cord blood DNA
471 methylation data. *Clin Epigenet*. 2019;11(1):125. doi:10.1186/s13148-019-0717-y
- 472 23. Teschendorff AE, Marabita F, Lechner M, et al. A beta-mixture quantile
473 normalization method for correcting probe design bias in Illumina Infinium 450 k
474 DNA methylation data. *Bioinformatics*. 2013;29(2):189-196.
475 doi:10.1093/bioinformatics/bts680
- 476 24. Du P, Zhang X, Huang CC, et al. Comparison of Beta-value and M-value methods
477 for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*.
478 2010;11(1):587. doi:10.1186/1471-2105-11-587
- 479 25. Lin X, Tan JYL, Teh AL, et al. Cell type-specific DNA methylation in neonatal cord
480 tissue and cord blood: a 850K-reference panel and comparison of cell types.
481 *Epigenetics*. 2018;13(9):941-958. doi:10.1080/15592294.2018.1522929
- 482 26. Bakulski KM, Feinberg JI, Andrews SV, et al. DNA methylation of cord blood cell
483 types: Applications for mixed cell birth studies. *Epigenetics*. 2016;11(5):354-362.
484 doi:10.1080/15592294.2016.1161875
- 485 27. Andrew E. Teschendorff <A. Teschendorff@Ucl.Ac.Uk> SCZC. EpiDISH.
486 Published online 2017. doi:10.18129/B9.BIOC.EPIDISH
- 487 28. He B, Liu Y, Maurya MR, et al. The maternal blood lipidome is indicative of the
488 pathogenesis of severe preeclampsia. *Journal of Lipid Research*. 2021;62:100118.
489 doi:10.1016/j.jlr.2021.100118
- 490 29. Chen Y, He B, Liu Y, et al. Maternal plasma lipids are involved in the pathogenesis
491 of preterm birth. *GigaScience*. 2022;11:giac004. doi:10.1093/gigascience/giac004
- 492 30. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses
493 for RNA-sequencing and microarray studies. *Nucleic Acids Research*.
494 2015;43(7):e47-e47. doi:10.1093/nar/gkv007
- 495 31. Kasper Daniel Hansen [Cre A. IlluminaHumanMethylationEPICanno.ilm10b4.hg19.
496 Published online 2017.

- 497 doi:10.18129/B9.BIOC.ILLUMINAHUMANMETHYLATIONEPICANNO.ILM10
498 B4.HG19
- 499 32. Drier Y, Sheffer M, Domany E. Pathway-based personalized analysis of cancer. *Proc*
500 *Natl Acad Sci USA*. 2013;110(16):6388-6393. doi:10.1073/pnas.1219651110
- 501 33. R Core Team. R: A language and environment for statistical computing. Published
502 online 2021. <https://www.R-project.org/>
- 503 34. Tian Y, Morris TJ, Webster AP, et al. ChAMP: updated methylation analysis pipeline
504 for Illumina BeadChips. Valencia A, ed. *Bioinformatics*. 2017;33(24):3982-3984.
505 doi:10.1093/bioinformatics/btx513
- 506 35. Zheng SC, Breeze CE, Beck S, Teschendorff AE. Identification of differentially
507 methylated cell types in epigenome-wide association studies. *Nat Methods*.
508 2018;15(12):1059-1066. doi:10.1038/s41592-018-0213-x
- 509 36. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-
510 wide association studies. *Genome Biol*. 2014;15(2):R31. doi:10.1186/gb-2014-15-2-
511 r31
- 512 37. Hebbar S, Misha M, Rai L. Significance of Maternal and Cord Blood Nucleated Red
513 Blood Cell Count in Pregnancies Complicated by Preeclampsia. *Journal of*
514 *Pregnancy*. 2014;2014:1-7. doi:10.1155/2014/496416
- 515 38. Perrone S. Nucleated red blood cell count in term and preterm newborns: reference
516 values at birth. *Archives of Disease in Childhood - Fetal and Neonatal Edition*.
517 2005;90(2):F174-F175. doi:10.1136/adc.2004.051326
- 518 39. Hermansen MC. Nucleated red blood cells in the fetus and newborn. *Arch Dis Child*
519 *Fetal Neonatal Ed*. 2001;84(3):F211-F215. doi:10.1136/fn.84.3.F211
- 520 40. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as
521 surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012;13(1):86.
522 doi:10.1186/1471-2105-13-86
- 523 41. Braid SM, Okrah K, Shetty A, Corrada Bravo H. DNA Methylation Patterns in Cord
524 Blood of Neonates Across Gestational Age: Association With Cell-Type Proportions.
525 *Nursing Research*. 2017;66(2):115-122. doi:10.1097/NNR.0000000000000210
- 526 42. Liu Y, Aryee MJ, Padyukov L, et al. Epigenome-wide association data implicate
527 DNA methylation as an intermediary of genetic risk in Rheumatoid Arthritis. *Nat*
528 *Biotechnol*. 2013;31(2):142-147. doi:10.1038/nbt.2487
- 529 43. Middleton LYM, Dou J, Fisher J, et al. Saliva cell type DNA methylation reference
530 panel for epidemiological studies in children. *Epigenetics*. 2022;17(2):161-177.
531 doi:10.1080/15592294.2021.1890874
- 532 44. Campbell KA, Colacino JA, Puttabatappa M, et al. Placental cell type deconvolution
533 reveals that cell proportions drive preeclampsia gene expression differences.
534 *Commun Biol*. 2023;6(1):264. doi:10.1038/s42003-023-04623-6
- 535 45. Wang X, Cho HY, Campbell MR, et al. Epigenome-wide association study of
536 bronchopulmonary dysplasia in preterm infants: results from the discovery-BPD
537 program. *Clin Epigenet*. 2022;14(1):57. doi:10.1186/s13148-022-01272-0
- 538 46. Kashima K, Kawai T, Nishimura R, et al. Identification of epigenetic memory
539 candidates associated with gestational age at birth through analysis of methylome and
540 transcriptional data. *Sci Rep*. 2021;11(1):3381. doi:10.1038/s41598-021-83016-3
541
542

543

544

545

546

547

548

549

550

551

552

553

554 **Table 1: Patient Characteristics**

| <i>Variables</i> | <i>PE Cases (n = 24) mean (sd)</i> | <i>Controls (n = 38) mean (sd)</i> | <i>P-value</i> |
|--------------------------------|--|--|----------------|
| <i>Maternal Age (Years)</i> | 28.75 (5.88) | 27.24 (6.35) | 0.34 |
| <i>Parity</i> | 1.54 (1.41) | 1.57 (1.78) | 0.93 |
| <i>BMI</i> | 32.24 (9.38) | 27.75 (9.20) | 0.07 |
| <i>Smoker (n)</i> | 6 | 8 | 0.96 |
| <i>Gestational Age (Weeks)</i> | 35.58 (2.90) | 39.16 (0.92) | 3.66E-06 |
| <i>Ethnicity (n)</i> | | | 0.60 |
| <i>Asian</i> | 12 | 21 | - |
| <i>Caucasian</i> | 3 | 7 | - |
| <i>Pacific Islander</i> | 9 | 10 | - |

* Numeric variables are compared with *t*-test;

* Categorical variables are compared with the Chi-square test.

555

556

557 **Fig. 1: Study Overview and experiment design.** The entire data analysis procedure is outlined
558 in this workflow, which incorporates methods that account for clinical confounding and cell type
559 confounding.

560

561 **Fig. 2: Construction of cell type reference matrix.** Principal Component Analysis (PCA) plots
562 of DNA methylation reference matrix from (A) Lin et al.²⁵, (B) Bakulski et al.²⁶, and (C) our
563 merged reference, colored by cell type. Our reference combined nucleated red blood cells
564 (nRBCs) from Bakulski et al and 6 cell types (B cell, CD4T, CD8T, monocyte, natural killer cell,
565 and granulocyte) from Lin et al.²⁵. Our reference shows better separation between cell types,
566 compared to the references of Lin et al.²⁵. and Bakulski et al.²⁶

567

568 **Fig. 3: Cell types in samples.** (A) Heatmap displaying estimated cell-type proportions among 62
569 samples (including 24 PE cases and 38 controls), the colors indicate the relative proportions of
570 cell types, with red indicating a higher proportion and blue indicating a lower proportion. (B)
571 Side-by-side boxplots displaying cell-type proportions in PE cases vs. controls before adjusting
572 for clinical variables. An asterisk (*) is used to indicate a significant difference by using Multiple
573 Linear Regression (MLR) between the case and control groups (p-value < 0.05), while "ns" is
574 used to indicate a non-significant difference. (C) The Source of Variance (SOV) analysis of cell
575 type composition from patient characteristics. Confounding factors were identified by considering
576 variables with an F-mean value greater than 1. (D) Side-by-side boxplots displaying cell-type
577 proportions in PE cases vs. controls after adjusting for confounders identified in (C). An asterisk
578 (*) is used to indicate a significant difference by using Multiple Linear Regression (MLR)
579 between the case and control groups (p-value < 0.05), while "ns" indicates a non-significant
580 difference.

581

582 **Fig. 4: PE is not associated with significant changes in cord blood DNA methylation, after**
583 **confounding adjustment.** (A) The Source of Variance (SOV) analysis was conducted on both
584 clinical variables and cell types. Confounding factors were identified by considering variables
585 with an F-mean value greater than 1. (B) The volcano plot of the differential methylation analysis
586 results without confounding adjustment. The x-axis represents log fold change between severe PE
587 and controls; the y-axis is negative log-transformed p-values after BH adjustment. The red dots
588 are differentially methylated probes (DMP) associated with severe PE after BH adjustment,
589 whereas the black dots represent non-significant probes. (C) The volcano plot after adjusting for
590 all confounding factors.

591

592 **Fig. 5: Cell type proportion changes with gestational age.** Scatter plots labeled (A) to (E)
593 depict the changes in the proportions of each cell type in cord blood along with gestational age.
594 The reported p-value measures the relationship between GA and each cell type, with a threshold
595 of p-value < 0.05.

596

597

598 **Supplementary Figures**

599 **Supplementary Figure 1: Data Processing Workflow.** The complete data pre-processing
600 procedures consisted of filtration, imputation of missing values, quality control checks,
601 normalization, batch correction, singular value decomposition analysis, and conversion of beta
602 values to M-values.

603 **Supplementary Figure 2: Data Quality Control.** (A) and (B) Density plots for before and after
604 the removal of one control sample with a distinct beta density distribution. (C) and (D) Plots of
605 the singular value decomposition analysis are presented before and after the removal of batch
606 effects.

607

608 **Supplementary Figure 3: Validation of the impact of confounding adjustment using Ching**
609 **et al.'s 450k cord blood methylation data¹⁷.** (A) The volcano plot of differential methylation
610 results without confounder adjustment as done by Ching et al. using their data. The red dots are
611 differentially methylated probes (DMP) associated with EOPE after BH adjustment, whereas the
612 black dots represent non-significant probes. (B) The Source of Variance (SOV) analysis was
613 conducted on both clinical variables and cell types. Confounding factors were identified by
614 considering variables with an F-mean value greater than 1. (C) The volcano plot after adjusting
615 for all confounding factors.

616

617 **Supplementary Figure 4: Cell type changes with gestational age by sample group.** Scatter
618 plots labeled (A) to (G) depict the changes in the proportions of each cell type along with
619 gestational age. The green line in each plot represents the PE case group, while the red line
620 represents the control group. The reported p-value measures the interaction between GA and the
621 sample group of trends between the PE case group and the control group, with a threshold of p-

622 value < 0.05. A non-significant p-value indicates the trends between cell proportion and GA are
623 consistent in the case and control groups.

624

625 **Supplementary Figure 5: Cell proportion changes with gestational age are consistent in**
626 **different datasets.** The scatter plots compare the changes in cell-type proportions with
627 gestational age in two datasets. Plots (A - F) display the comparisons within PE case samples for
628 both studies, while plots (G) through (L) display the comparisons within control samples for both
629 studies. The purple line in each plot represents the cell proportion in our whole cord blood
630 samples, while the orange line represents the cell proportion in PBMC cord blood samples from
631 another study. The p-values of the interaction term between GA and datasets are reported in each
632 plot. A non-significant p-value suggests the trend between cell proportion and GA is consistent in
633 the two datasets.

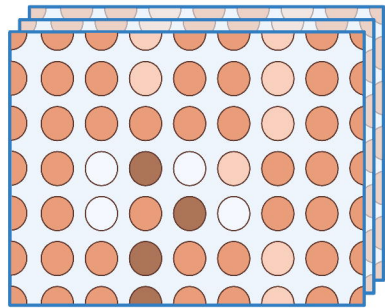
634

635 **Supplementary Tables**

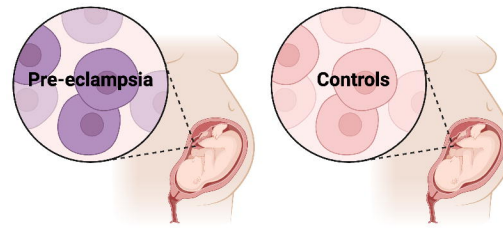
636 **Supplementary Table 1: Linear regression of each cell type on clinical variables**

637

638 **Supplementary Table 2: The closest functional genes to the significant CpG site, cg20135196**

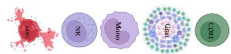


Processed Methylation Data

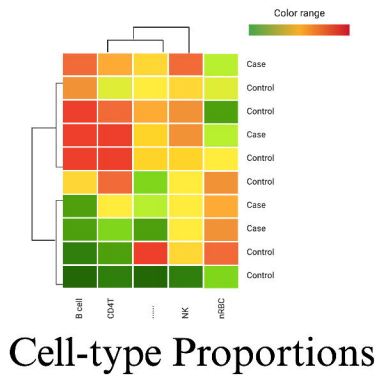


Cord Blood Samples

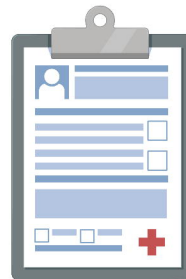
Cell Type Deconvolution



Cell-type specific DNA methylation Reference



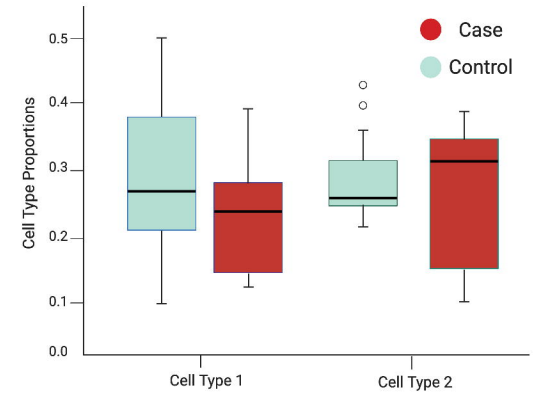
Cell-type Proportions



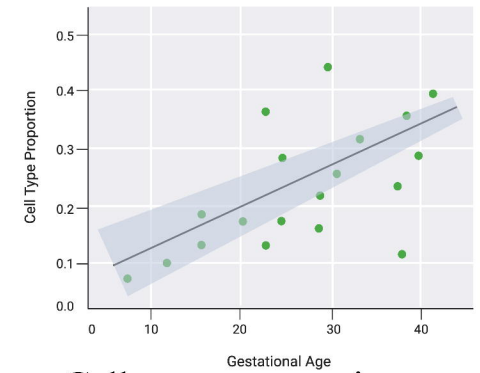
Clinical Information

Multivariate linear regression

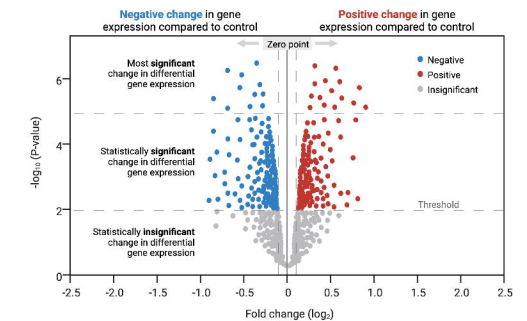
EWAS



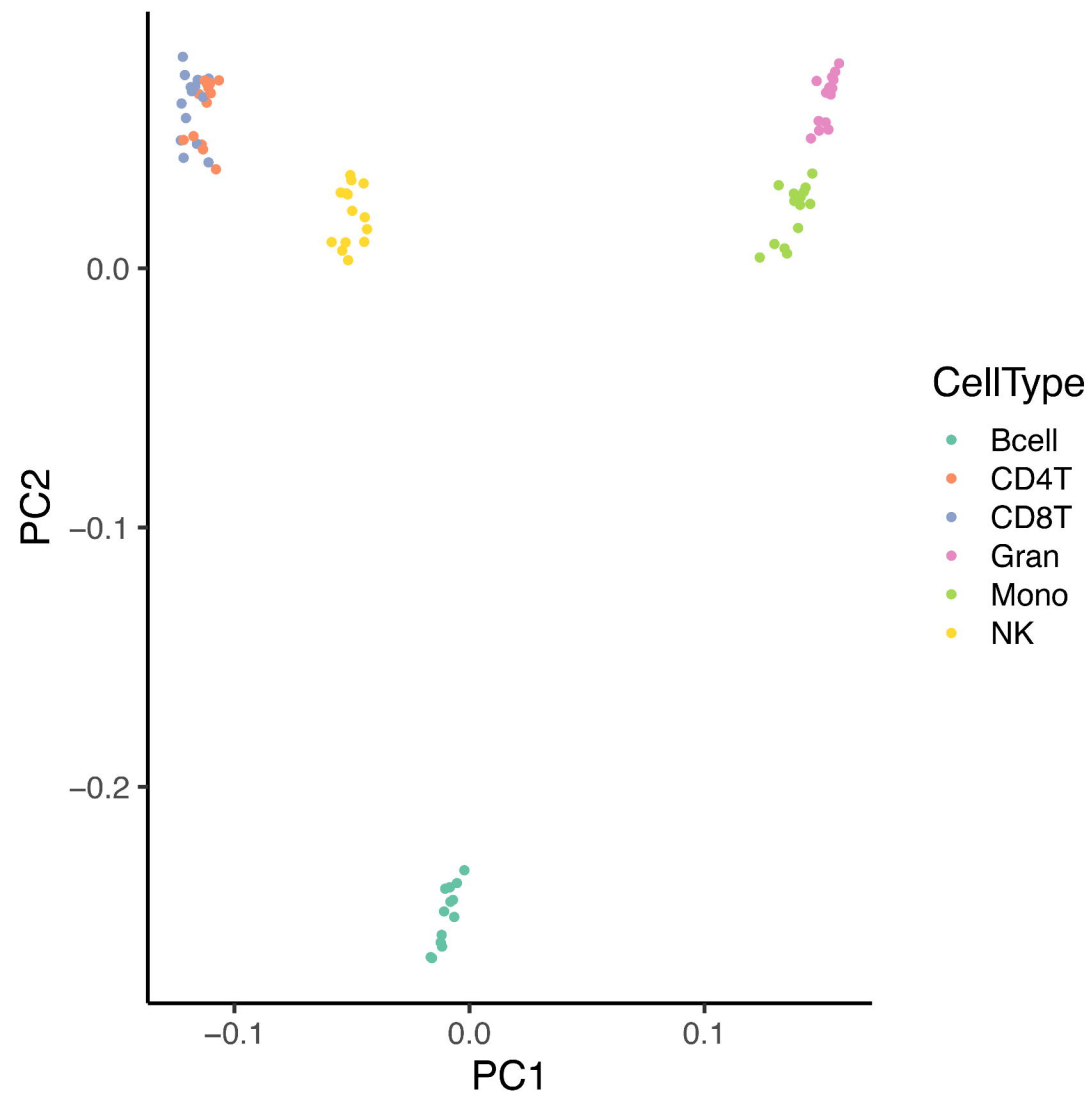
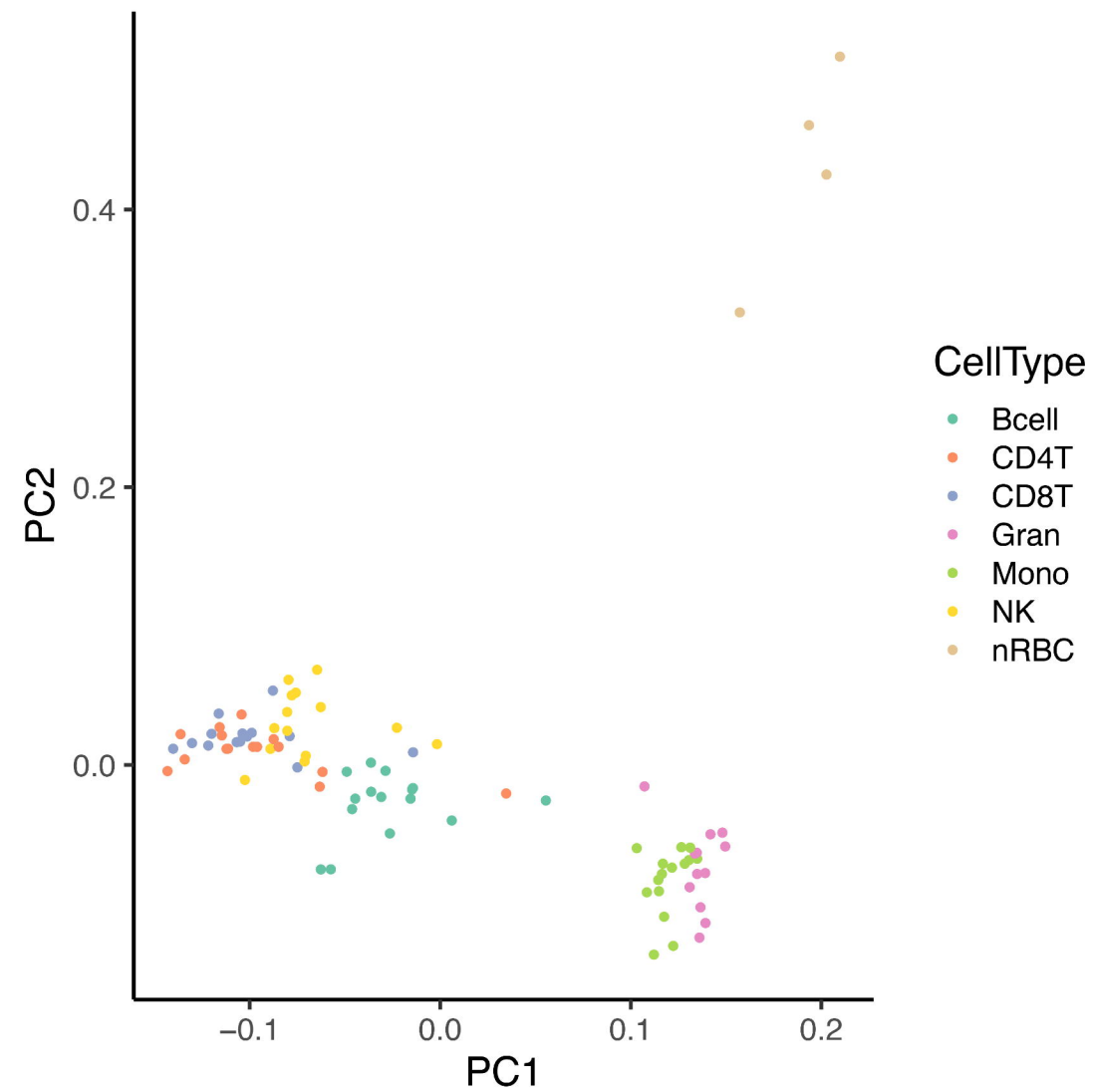
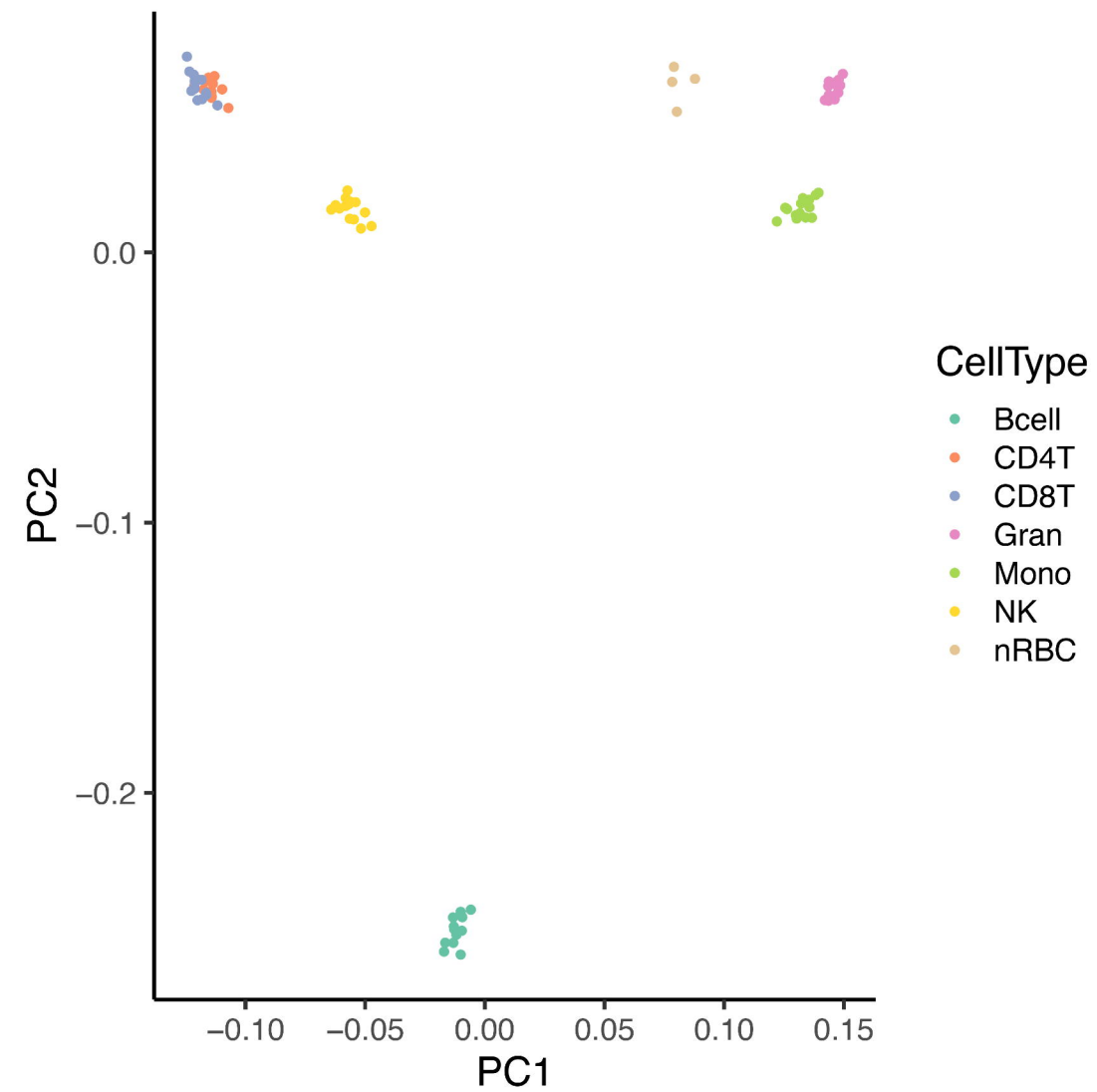
Cell type proportions vs Preeclampsia

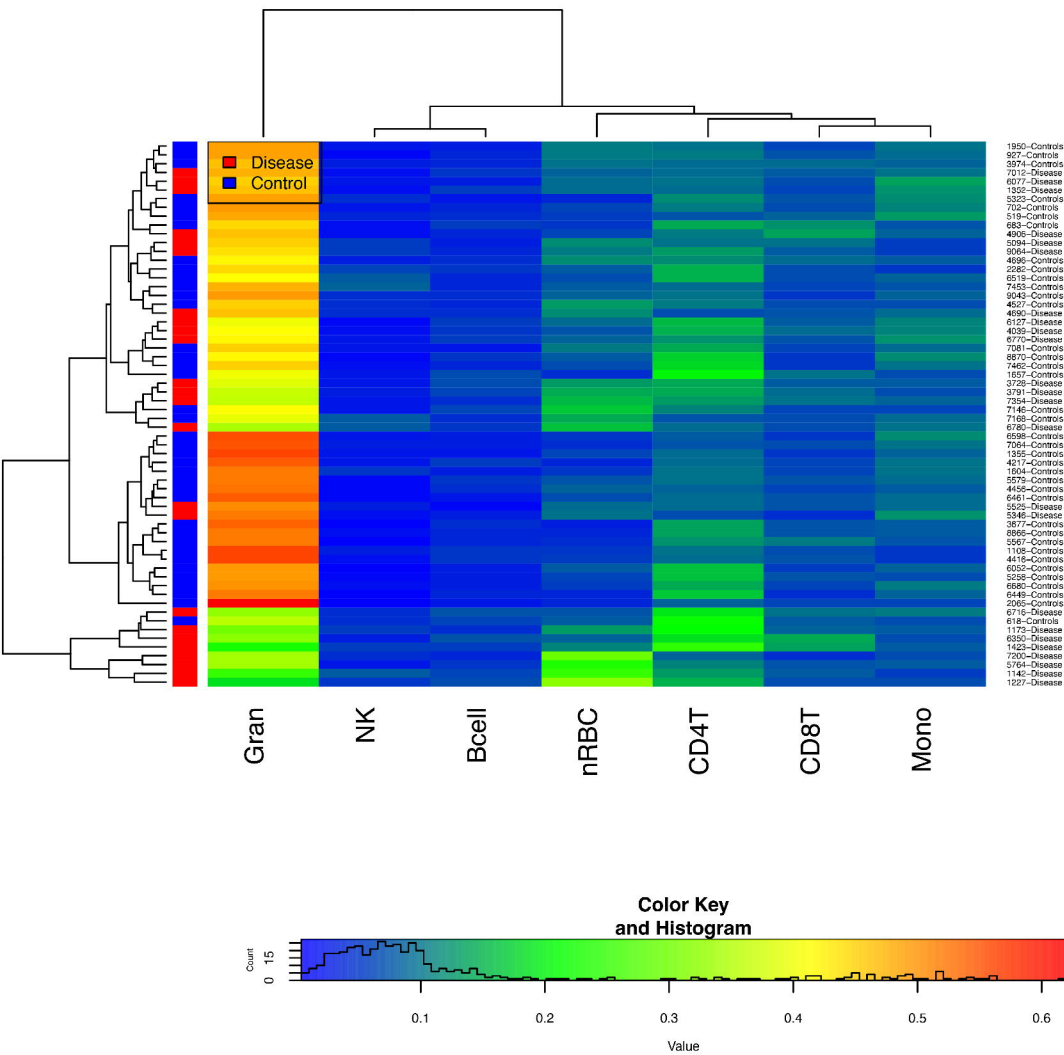
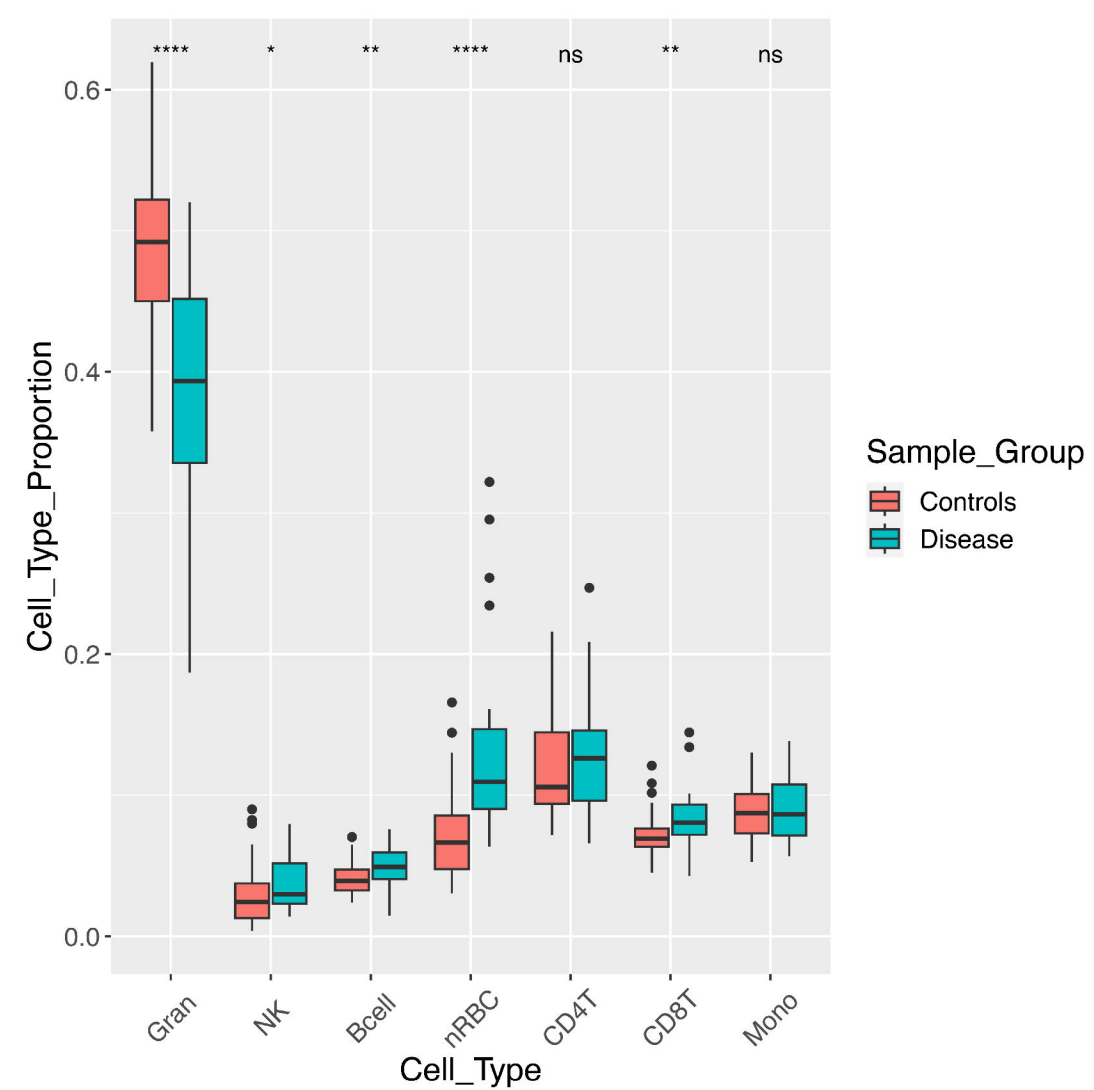
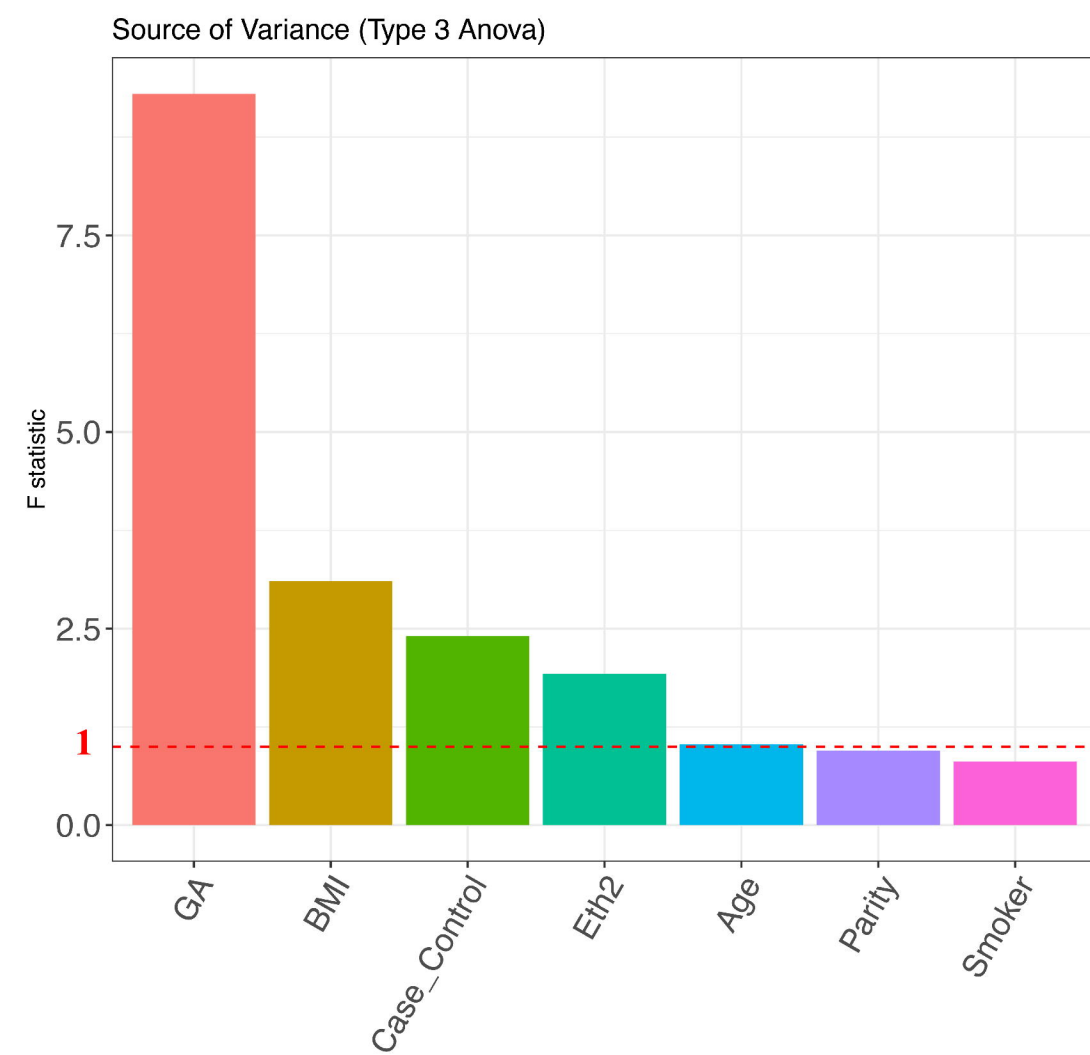
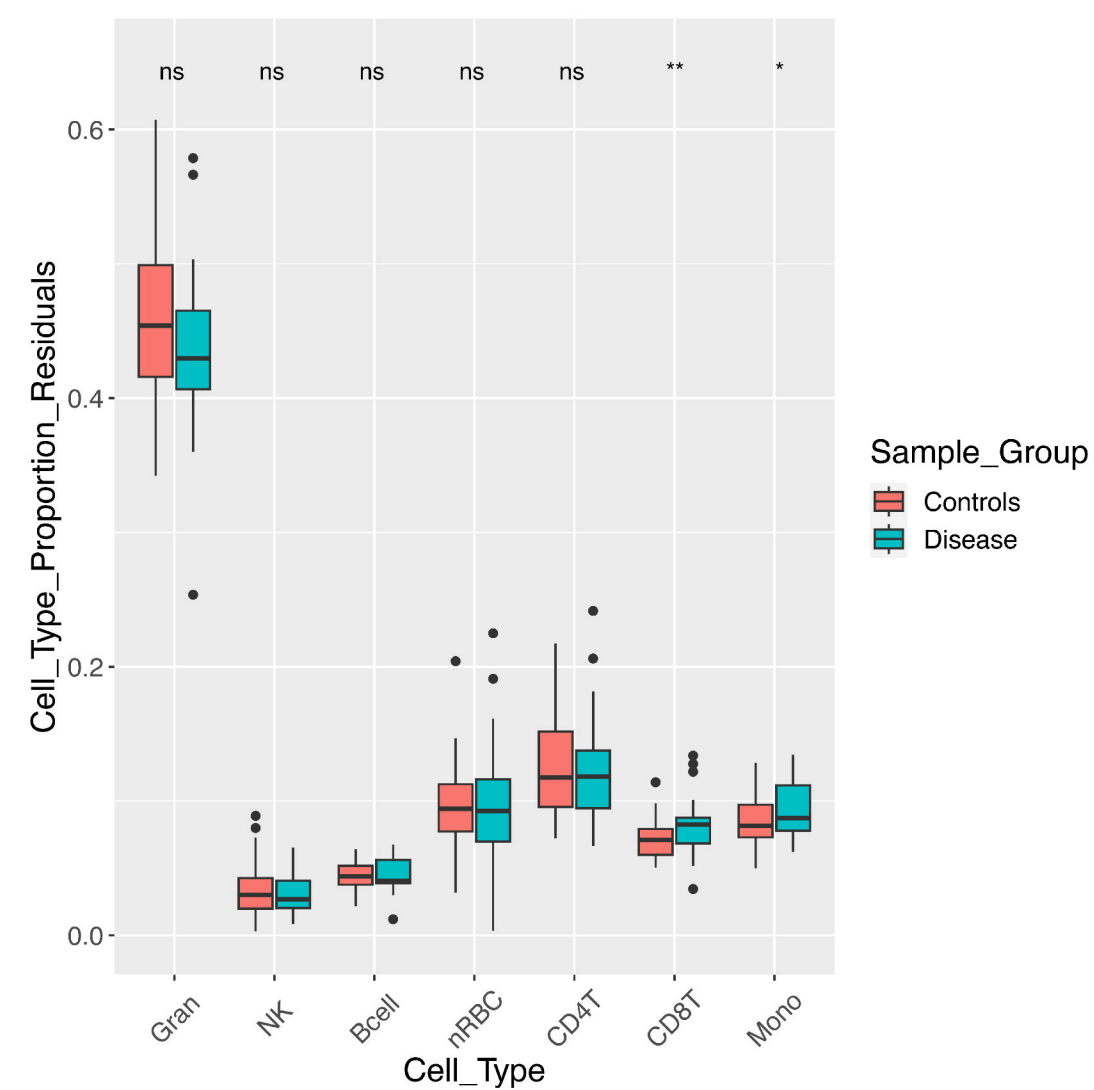


Cell type proportions vs Gestational Age



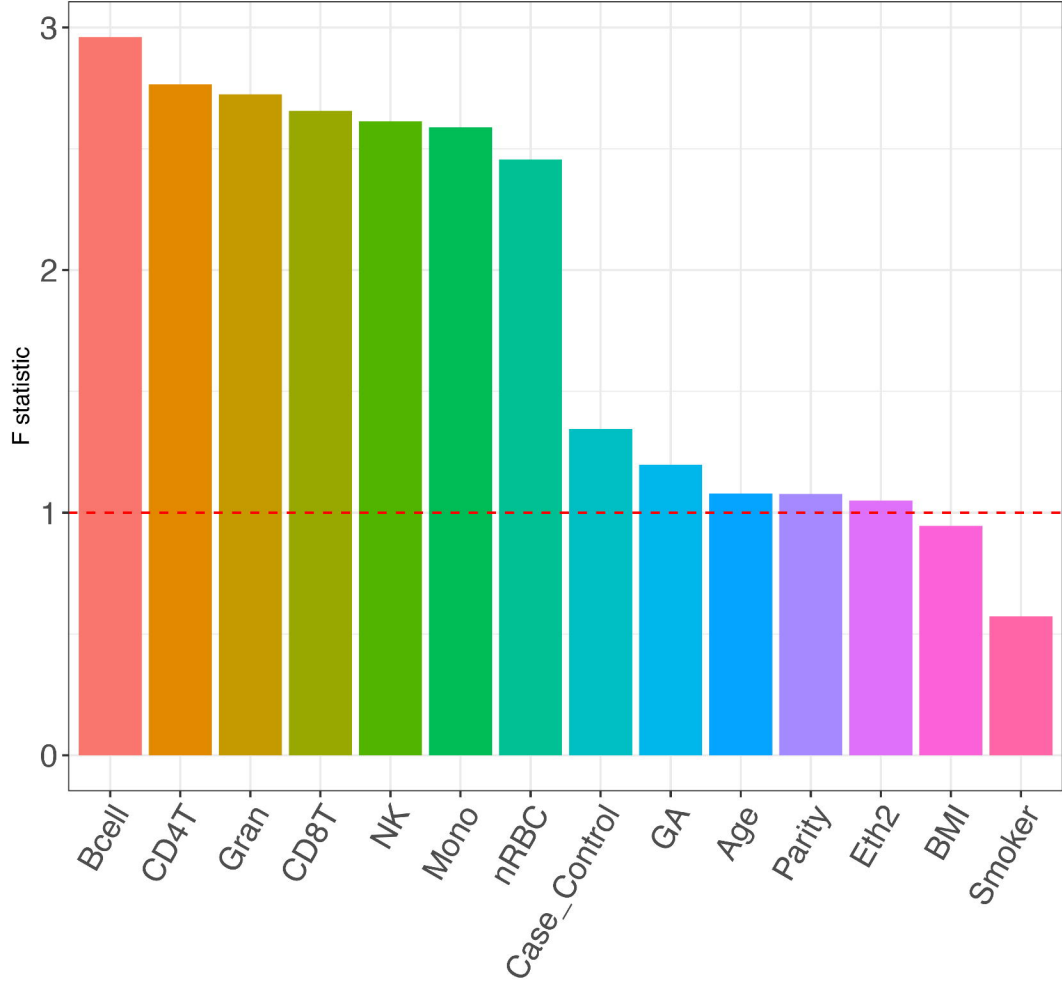
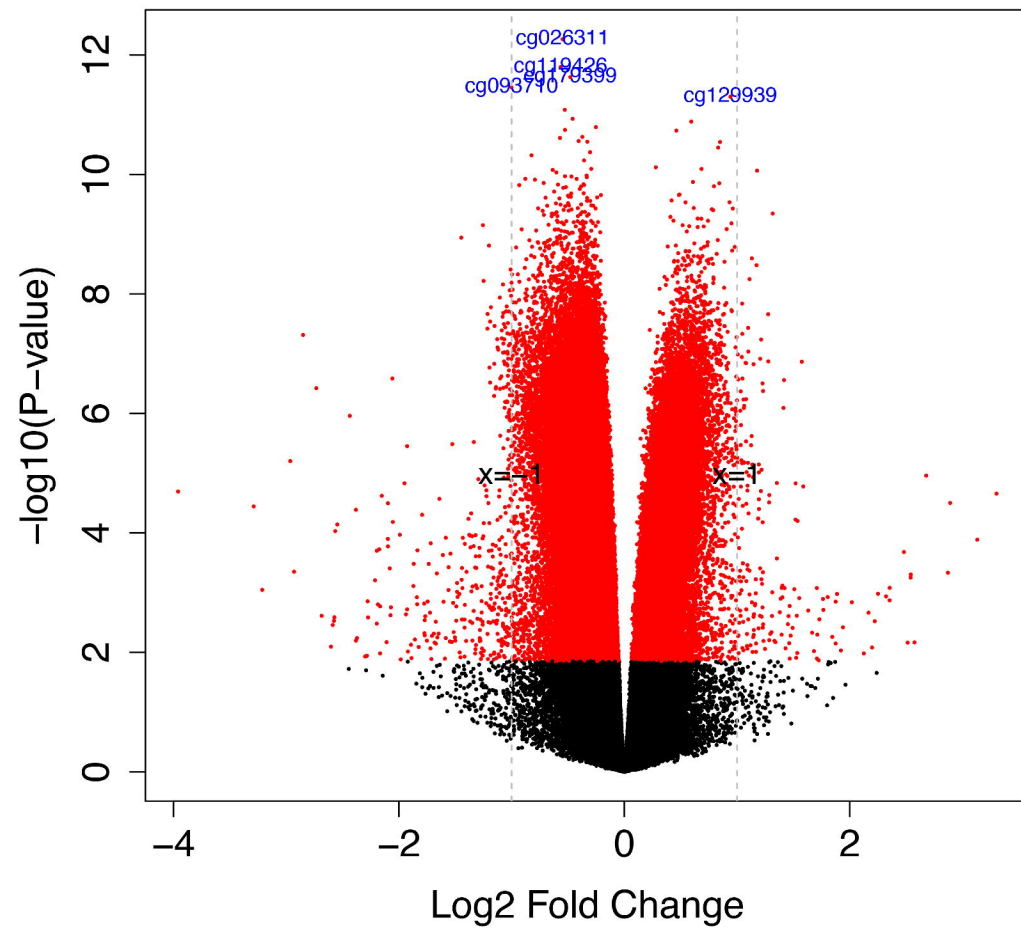
Differential Methylation

A**B****C**

A**B****C****D**

A

Source of Variance (Type 3 Anova)

**B****C**