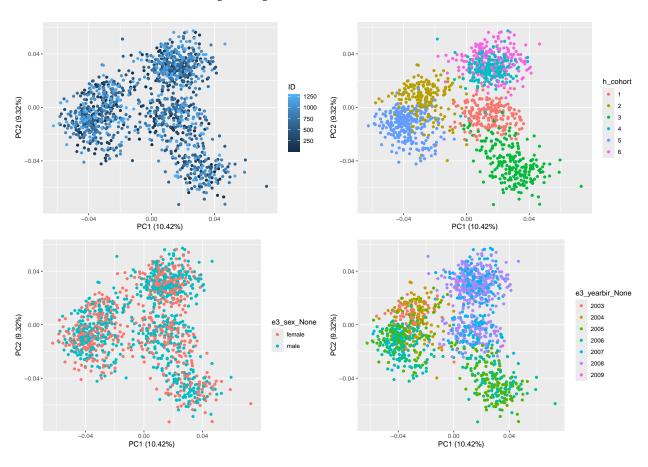
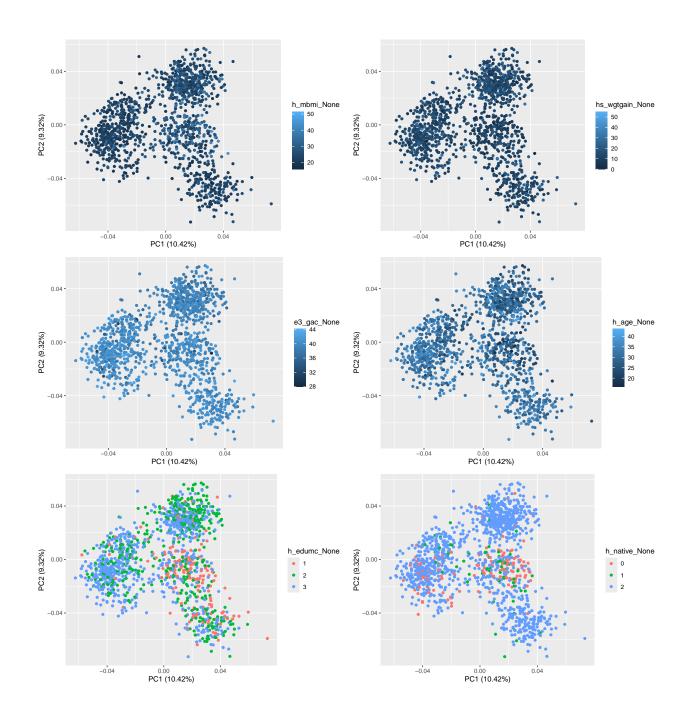
## Rapport

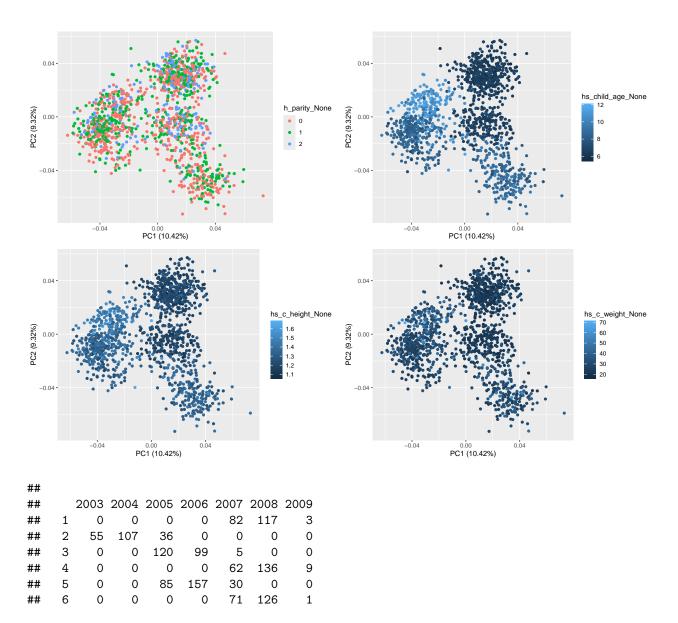
Groupe 5

2024-06-05

## Détection des covariantes principales influant le modèle







Un classement des bases de données d'expositions a été fait en se basant sur deux critères :

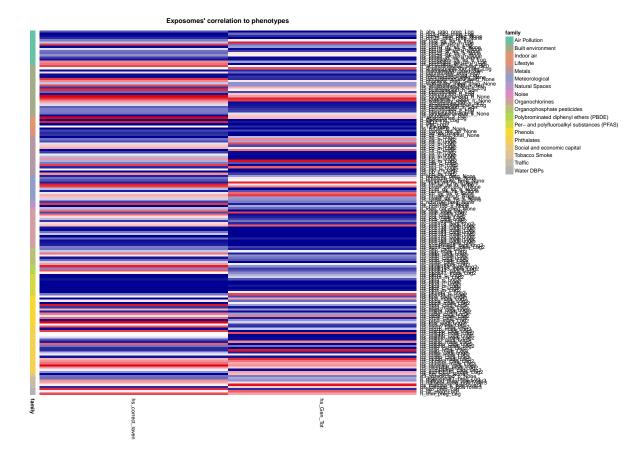
- Les types de données : quantitative (numériques) ou qualitative (facteurs)
- Les données liées à la grossesse (prénatales) les données postnatales.

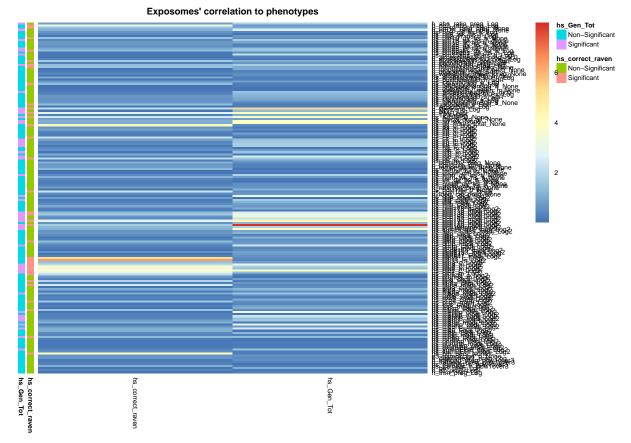
Ceci a été fait pour pouvoir effectuer une analyse de composantes principales sur les différentes sous bases de données obtenues et identifier les covariantes influant les différentes expositions.

Dans l'analyse des composantes principales on peut trouver des fortes corrélations liées au la variable h\_cohort qui représente le pays auquel les données ont été collectées et au e3\_yearbir\_None qui représente l'année de naissance. Pour la suite du projet, nous ne prendrons en compte que la variable h\_cohort parce qu'ils sont liés l'un à l'autre.

## Matrice de "corrélations" (p\_values) entre les expositions et les différents phénotypes

## New names:





Les facteurs de confusions étant finalement déterminés, et pour mieux identifier les relations entre les différentes expositions et les phénotypes, une matrice contenant les p-values de la régression linéaire généralisé (glm) entre chaque exposome et chaque phénotype a été calculé. Ceci a été effectué sur les deux sous bases de données (numérique pendant la grossesse et numérique postnatale) . Grâce à ceci, on a pu décider le phénotype d'intérêt sur lequel on allait travailler (Développement neuronal chez l'enfant h\_Gen\_Tot) et les familles d'expositions considérées selon les valeurs de p values les plus petites présentes dans la matrice.

Les familles d'exposition qui vont être considérés sont : Organochlorines, Air pollution, Phthalates, Lifestyle and PFAS