DNA Methylation Score Development for In utero Exposure to Paternal Smoking Using a Supervised Machine Learning Approach

Authors : Cristy Stagnar, Nina Hubig, Diana Ivankovic

Abstract : The epigenome is a compelling candidate for mediating long-term responses to environmental effects modifying disease risk. The main goal of this research is to develop a machine learning-based DNA methylation score, which will be valuable in delineating the unique contribution of paternal epigenetic modifications to the germline impacting childhood health outcomes. It will also be a useful tool in validating self-reports of nonsmoking and in adjusting epigenome-wide DNA methylation association studies for this early-life exposure. Using secondary data from two population-based methylation profiling studies, our DNA methylation score is based on CpG DNA methylation measurements from cord blood gathered from children whose fathers smoked pre- and peri-conceptually. Each child’s mother and father fell into one of three class labels in the accompanying questionnaires -never smoker, former smoker, or current smoker. By applying different machine learning algorithms to the accessible resource for integrated epigenomic studies (ARIES) sub-study of the Avon longitudinal study of parents and children (ALSPAC) data set, which we used for training and testing of our model, the best-performing algorithm for classifying the father smoker and mother never smoker was selected based on Cohen’s κ. Error in the model was identified and optimized. The final DNA methylation score was further tested and validated in an independent data set. This resulted in a linear combination of methylation values of selected probes via a logistic link function that accurately classified each group and contributed the most towards classification. The result is a unique, robust DNA methylation score which combines information on DNA methylation and early life exposure of offspring to paternal smoking during pregnancy and which may be used to examine the paternal contribution to offspring health outcomes.

Keywords : epigenome, health outcomes, paternal preconception environmental exposures, supervised machine learning
Conference Title : ICSGGE 2021 : International Conference on Statistical Genetics and Genetic Epidemiology
Conference Location : Venice, Italy
Conference Dates : August 12-13, 2021

Open Science Index, Biomedical and Biological Engineering Vol:15, No:08, 2021 publications.waset.org/abstracts/139782.pdf