QC EWAS

QUALITY CONTROL FLOW FOR EPIGENOME-WIDE ASSOCIATION STUDIES

Ainhoa Sánchez Ramón, AU726681

MSc in Bioinformatics

June 2024

Advisor: Anna Starnawska
Jakob Grove

AARHUS UNIVERSITET

QC EWAS

Quality control flow for Epigenome-Wide Association Studies

AINHOA

MSc in Bioinformatics

Bioinformatics Research Center
Faculty of Natural Sciences
Aarhus University

June 2024

CONTENTS

1 ABSTRACT 1

2 BACKGROUND 3
- 2.1 DNA methylation 3
- 2.2 450K Infinium HumanMethylation450 BeadChip design 5
- 2.3 Quality Control Analysis 6
- 2.4 iPSYCH 12

3 METHODS 13
- 3.1 MINERvA cohort 13
- 3.2 Reading the methylation files and quality filtering 15
- 3.3 Normalization and Correction 18
- 3.4 Ancestry and relatedness 20
- 3.5 Predictors 21
- 3.6 EWAS 22

4 RESULTS 25

5 DISCUSSION 43

6 CONCLUSIONS 49

BIBLIOGRAPHY 51

LIST OF FIGURES

Figure 1 Schematic Representation of DNA Methylation [74] 3

Figure 2 Illumina Infinium HumanMethylation450 Assay. 7

Figure 3 Example of Density Distribution of β Values [88] 10

Figure 4 Schematic Workflow of the Procedure Followed for Quality Control (QC) 14

Figure 5 Sex Prediction Plot 26

Figure 6 Bad Samples Plot 27

Figure 7 SVD Heatmap of Raw Filtered Beta Values 29

Figure 8 Intensity Distributions of Infinium I and II Probe Types Before and After Dasen Normalization 30

Figure 9 SVD Heatmap of Dasen Normalized Beta Values 31

Figure 10 SVD Heatmap of Beta Values Corrected with 5 Control PCs 32

Figure 11 SVD Heatmap of Beta Values Corrected with 7 Control PCs 32

Figure 12 SVD Heatmap of Beta Values Corrected with 8 Control PCs 33

Figure 13 SVD Heatmap of Betas After ComBat Array Correction 33

Figure 14 Boxplot of Predicted Smoking Score Stratified by Maternal Smoking Status During Pregnancy 35

Figure 15 Relationship Between Actual and Predicted Gestational Week 36

Figure 16 QQ Plot of the Association Between Maternal Smoking During Pregnancy in the Register and Cord Blood DNAm 37

Figure 17 QQ Plot of the Association Between Predicted Maternal Smoking During Pregnancy and Cord Blood DNAm 37

Figure 18 Manhattan Plot of Genome-wide Results for Methylation Association with Smoking in the Register 38

Figure 19 Manhattan Plot of Genome-wide Results for Methylation Association with Predicted Smoking Score 39

Figure 20 QQ Plot of the Association Between Gestational Weeks in the Register and Cord Blood DNAm 40

Figure 21 QQ Plot of the Association Between Predicted Gestational Weeks and Cord Blood DNAm 40
Figure 22 Manhattan Plot of Genome-Wide Results for Methylation Association with Gestational Weeks in the Register 41
Figure 23 Manhattan Plot of Genome-wide Results for Methylation Association with Predicted Gestational Weeks 42

ACRONYMS AND ABBREVIATIONS

DNMTs	DNA methyltransferases
SAM	S-adenosylmethionine
5mC	5-methylcytosine
CGIs	CpG islands
TSSs	transcription start sites
iPSYCH	Integrative Psychiatric Research
ICC	intraclass correlation coefficient
DMSs	differentially methylated CpG sites
DPCRR	Danish Psychiatric Central Research Register
DNPR	Danish National Patient Register
ICD	World Health Organization International Classification of Disease
IDAT	Intensity Data
ASD	Autism spectrum disorder
MBR	Danish Medical Birth Register
ASD	autism spectrum disorder
nRBCs	nucleated red blood cells
QC	quality control

1 ABSTRACT

Epigenetics is a essential field of study in understanding the complexities of gene regulation and expression, particularly in relation to environmental influences and their long-term effects on health and disease. The need for research investigating DNA methylation (DNAm) in clinical studies has greatly increased, leading to the evolution of new analytic methods to improve accuracy and reproducibility of the interpretation of results from these studies. Among the advancements in this field, the Illumina Infinium BeadChips have revolutionized large-scale epigenome-wide association studies (EWAS) in human populations. In most studies, the main objective of using DNAm analysis is to detect differences in methylation at CpG sites between phenotypic groups.

Despite its robustness this technology has limitations. Batch effects and confounding variables can introduce significant noise, potentially masking true biological signals. For this reason, rigorous quality control (QC) procedures are essential to ensure the reliability and validity of the data before any meaningful analysis can be performed. This thesis focuses on the application of comprehensive QC measures to the MINERvA dataset, which is inside the iPSYCH cohort. The iPSYCH cohort is a large and robust dataset that offers a unique opportunity to explore epigenetic mechanisms underlying complex traits in the Danish population.

We applied an extensive preprocessing into MINERvA to address various potential issues, including sample quality, probe reliability, and genetic ancestry. Initial quality assessments ensured that only samples meeting stringent criteria were included, thereby minimizing the risk of bias from low-quality data. Probe filtering was applied to retain only those probes that provided reliable measurements of DNAm. Ancestry filtering was employed to reduce confounding effects arising from population stratification, ensuring that observed associations were not artifacts of genetic background.

Although there are no standard methods for analyzing EWAS data, the application of established packages and methodologies already used in previous research made us sure of the dataset's integrity. Comprehensive data evaluation, including the use of statistical tests and visualizations, provided initial insights and confirmed that the data will be ready for subsequent analysis. The main goal of this is to amplify the statistical power of the next analyses of MINERvA, reduce the risk of false positives, and support robust and reliable findings. Ensuring future reliable epigenetic studies of the MINERvA cohort. And increasing the potential for further research and clinical outcomes. By maintaining a high standard of data quality this research contributes to the field of epigenetics and its application in understanding and finding complex human diseases.