

Objective 1: Predicted vs Observed Methylation Concordance

Project Update

Haymanot Berehan

2026-07-23



Objectives

Key Objectives

- **Evaluate concordance between observed and predicted DNAm**
← current focus
- Test association between childhood adversity and DNAm
- Identify biological pathways and functions
- Explore mediation between adversity and selected health outcomes

Data Sources

Cohort

GTP

Predicted DNAm

Elastic net model

Observed DNAm

QC'ed (ssNoob + ComBat)

Array

EPIC platform

Data	Description	Role
Predicted methylation	Genetically-regulated estimates from elastic-net models trained on cis-SNPs (Athina, Mclean)	X (genetically-regulated estimate)
Observed methylation	ssNoob-normalised, ComBat batch-corrected beta values (Seyma, McLean)	Y (measured methylation)
Sample linker	To link genotype IDs >>> SID >>> Methylation Sentrix IDs.	ID harmonisation

Pipeline Overview

Load pred + obs

→

Linker harmonisation

→

Match samples

→

Match CpGs

→

Per-CpG correlation

→

Concordance summary

Key decisions

- Both predicted and observed matrices are CpG (rows) x sample (columns), matched on Sentrix ID

- Overlap restricted to CpGs and samples present in **both** datasets

Sample & CpG Matching

Before matching

Dataset	CpGs	Samples
Predicted methylation(Imputed from ARIES dataset)	50653	7206
Observed methylation	819708	503

After matching

Metric	N
Matched samples	462

Metric	N
Overlapping CpGs	42064

Analytic Approach

Per-CpG concordance across matched samples

Metric	Purpose	Unit
Pearson r	Linear association between predicted and observed methylation	Per CpG, across samples
Spearman rho	Monotonic (rank-based) association, robust to non-linearity	Per CpG, across samples

Visualisation

- Distribution of per-CpG correlation coefficients (histogram)

- Scatter plot (predicted vs observed)
- Bland-Altman plot on standardised (z-scored) values, since raw scales differ

Concordance Summary

Metric	Value
Pearson r - mean	0.236
Pearson r - median	0.194
Spearman rho - mean	0.223
Spearman rho - median	0.184

Tips

- Correlations in this space are typically **small-to-modest** ($r \approx 0.1$ – 0.3 per CpG is common in the literature).

- Predicted values capture **only the genetically-driven component** of methylation; observed values contain that signal **plus** age, cell-type composition, batch, and biological noise.

Distribution of Per-CpG Correlations

Reading the distribution

- Most CpGs are cluster **near zero**, genotype is not the dominant driver of methylation at most sites.
- The **right tail** (higher r values), these are CpGs where genetic control over methylation is strong enough to be detected even with observed data carrying additional non-genetic variance.

Observed vs Predicted (Scatter)

Tip

Predicted (x-axis) reflects genetically-regulated, residualized estimates and is not expected to align in absolute scale with observed beta values (y-axis).

Agreement (Bland-Altman, Standardised)

Tip

Values z-scored per CpG prior to plotting, since predicted and observed are not on comparable native scales - this reflects **relative** agreement rather than raw unit agreement.

Key Challenge: Scale Mismatch

Predicted and observed methylation are not on the same scale

Property	Predicted	Observed
Range	Unbounded, continuous	[0, 1]
Interpretation	Genetically-regulated (residualized) estimate	Raw beta value (proportion methylated)

Property	Predicted	Observed
Can be negative?	Yes	No

Progress Summary

Completed

Task	Status
Predicted DNAm data received & cleaned	Done
Observed DNAm data received (cleaned)	Done
Sample ID harmonisation	Done

Task	Status
Per-CpG correlation pipeline built	Done

Next Steps

Task	Status
Finalise CpG/sample matching QC	In progress
Replicate this pipeline on 2nd predicted dataset (Understanding Society)	In progress

Task	Status
Compare concordance across the two prediction sources	Pending
Draft Objective 1 results write-up	Pending

Thank You!!



Speaker notes