How accurately can we predict someone's age from a DNA sample?

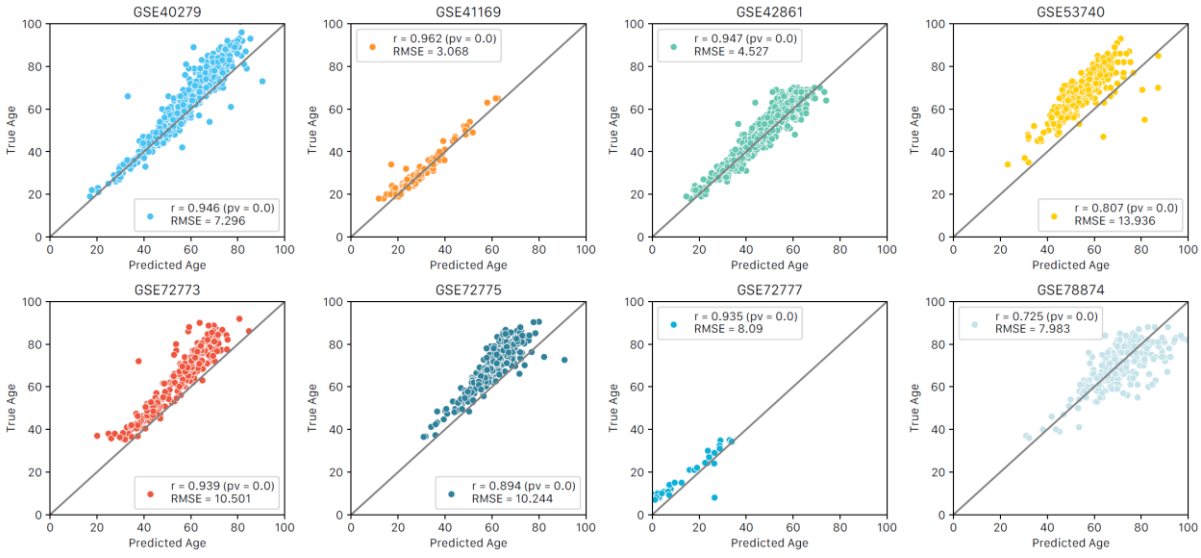
Expert: Riccardo Marioni (University of Edinburgh – Centre for Genomic and Experimental Medicine)

Whereas our genetic code – a string of 6 billion letters from a four letter alphabet (A, T, C and G nucleotides) – remains unchanged from conception to death, there are several chemical additions to DNA that can help to turn genes on and off. One of the best-studied changes is called DNA methylation, which involves the addition of a methyl group to the C nucleotide, typically when it is paired next to a G. We refer to this as CpG methylation. DNA methylation is a reversible process and it takes place at different levels in different tissues and cells.

Strikingly, there are very strong patterns between DNA methylation and chronological age. Several studies have shown that penalised regression models (typically elastic net) can produce weighted linear combinations of CpGs that correlate very highly (r~0.96) with chronological age in independent test datasets (Horvath 2013, Hannum 2013). However, the median absolute error for these predictors remains relatively large. An example of some ongoing age prediction work is shown in the **Figure** below.

Here, we will use publicly available blood-based DNA methylation datasets ranging from 46 to 689 participants and with ~25,000 CpGs per cohort. We will attempt to answer several questions:

1. Is there evidence for non-linear associations between CpGs and chronological age?
2. Which method yields the most accurate linear predictor in a test dataset?
    a. Can this be improved upon by the incorporation of quadratic and cubic terms?
3. Can we optimise pre-selection methods (e.g., PCA, univariate associated CpGs) to give subsets of CpGs for more advanced modelling approaches (tree-based models etc) that outperform simple linear predictors?
4. Do the predictors work equally across different age ranges and sexes?



Suggested Reading

Horvath & Raj, Nature Reviews Genetics, 2018, 19; 371-384

Bell et al, Genome Biology, 2019, 20; 249

Zhang et al, Genome Medicine, 2019, 11; 54

Datasets

A few datasets that could be considered for this project:

GSE40279: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40279

GSE72775: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72775

GSE78874: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE78874

GSE41169: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41169

GSE42861: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42861

GSE53740: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53740

GSE72773: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72773

GSE72777: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72777