Title: DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models

Authors: Anastasia Aliferi, David Ballard, Matteo D. Gallidabino, Helen Thurtle, Leon Barron, Denise Syndercombe Court

PII: S1872-4973(18)30371-5
DOI: https://doi.org/10.1016/j.fsigen.2018.09.003
Reference: FSIGEN 1964

To appear in: *Forensic Science International: Genetics*

Received date: 29-6-2018
Revised date: 23-8-2018
Accepted date: 6-9-2018

Please cite this article as: Aliferi A, Ballard D, Gallidabino MD, Thurtle H, Barron L, Syndercombe Court D, DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models, *Forensic Science International: Genetics* (2018), https://doi.org/10.1016/j.fsigen.2018.09.003

DNA methylation-based age prediction using massively parallel sequencing data and multiple machine learning models

Anastasia Aliferi[a], David Ballard[a], Matteo D. Gallidabino[a,b], Helen Thurtle[a], Leon Barron[a], Denise Syndercombe Court[a]

[a]*King's Forensics, Department of Analytical, Environmental and Forensic Sciences, Faculty of Life Sciences and Medicine, King's College London, 150 Stamford Street, London SE1 9NH, United Kingdom*

[b]*Centre for Forensic Science, Department of Applied Sciences, Faculty of Health and Life Sciences, Northumbria University Newcastle, Ellison Building, NE1 8ST Newcastle Upon Tyne, United Kingdom.*

Highlights

- 110 whole blood samples were analysed for 12 age-correlated CpGs using the Illumina MiSeq
- 17 different statistical models were developed using R and Trajan Neural Network Simulation
- A support vector machine model was selected with a MAE of 4.1 years in the blind test set
- The accuracy of both DNA methylation quantification and age prediction was retained down to 2ng of DNA input in the PCR stage
- Saliva samples (n=34) were accurately predicted with an error of less than 4 years for 50% of the samples and less than 7 years for 70%.
- No variation in DNA methylation was detected for sperm samples

Abstract

The field of DNA intelligence focuses on retrieving information from DNA evidence that can help narrow down large groups of suspects or define target groups of interest. With recent breakthroughs on the estimation of geographical ancestry and physical appearance, the estimation of chronological age comes to complete this circle of information. Recent studies have identified methylation sites in the human genome that correlate strongly with age and can be used for the development of age-estimation algorithms. In this study, 110 whole blood samples from individuals aged 11-93 years were analysed using a DNA methylation quantification assay based on bisulphite conversion and massively parallel sequencing (Illumina MiSeq) of 12 CpG sites. Using this data, 17 different statistical modelling approaches were compared based on root mean square error (RMSE) and a Support Vector Machine