

Forensic prediction of sex, age, height, body mass index, hip-to-waist ratio, smoking status and lipid lowering drugs using epigenetic markers and plasma proteins

Mònica Ortega Llobet^a, Åsa Johansson^a, Ulf Gyllensten^a, Marie Allen^a, Stefan Enroth^{a,b,*}

^a Department of Immunology, Genetics, and Pathology, Biomedical Center, SciLifeLab Uppsala, Uppsala University, SE-75108 Uppsala, Sweden

^b Swedish Collegium for Advanced Study, Thunbergsvägen 2, SE-752 38 Uppsala, Sweden

ARTICLE INFO

Keywords:

Externally visible characteristics (EVC)
Prediction
Epigenetics
Proteomics
Phenotyping

ABSTRACT

The prediction of human characteristics from blood using molecular markers would be very helpful in forensic science. Such information can be particularly important in providing investigative leads in police casework from, for example, blood found at crime scenes in cases without a suspect. Here, we investigated the possibilities and limitations of predicting seven phenotypic traits (sex, age, height, body mass index [BMI], hip-to-waist [WTH] ratio, smoking status and lipid-lowering drug use) using either DNA methylation or plasma proteins separately or in combination. We developed a prediction pipeline starting with the prediction of sex followed by sex-specific, stepwise, individual age, sex-specific anthropometric traits and, finally, lifestyle-related traits. Our data revealed that age, sex and smoking status can be accurately predicted from DNA methylation alone, while the use of plasma proteins was highly accurate for prediction of the WTH ratio, and a combined analysis of the best predictions for BMI and lipid-lowering drug use. In unseen individuals, age was predicted with a standard error of 3.3 years for women and 6.5 years for men, while the accuracy in smoking prediction across both men and women was 0.86. In conclusion, we have developed a stepwise approach for the de-novo prediction of individual characteristics from plasma proteins and DNA methylation markers. These models are accurate and may provide valuable information and investigative leads in future forensic casework.

1. Introduction

The prediction of human traits and characteristics has potential for use in forensic science, particularly in regard to provide investigative leads in police casework. Such leads may help to reduce the number of possible perpetrators in cases without a suspect, and to reduce the number of individuals to be screened during DNA mass testing. The prediction of traits based on genetic information, so-called forensic DNA phenotyping, can be used to assign the sex of an individual and, furthermore, may allow the prediction of several externally visible characteristics and appearance such as hair, eye and skin colour, along with face and hair morphology [1,2]. In addition to these characteristics, predictions of others, such as age, height or weight, may contribute to the identification of an unknown individual. Previous research has demonstrated that epigenetic information, such as DNA methylation levels, is correlated with age [3]. It has, moreover, been shown that the

DNA methylation status can be influenced by lifestyle choices such as smoking [4]. Additional studies demonstrate that DNA methylation levels across multiple methylation sites can be used to build models that can predict the age of an individual [5,6]. One such model developed by Horvath was based on 353 DNA methylation markers that could predict the age of 50 % of studied individuals within 3.6 years of their actual age. A later model developed by Hannum et al. consisted of 71 DNA methylation markers and had a root mean square error (RMSE) of 4.9 years in the prediction of age [5]. Additional predictions from epigenetic markers in relation to lifestyle habits, such as alcohol consumption [7], have been suggested to be areas of future interest for forensic research [8], although it has been shown to be difficult to replicate specific results [9]. In addition to DNA methylation, mRNA and plasma proteins may also be informative for the prediction of various traits, such as age [5] and height [10]. For instance, circulating concentrations of plasma proteins are strongly influenced by genetic variation [11,12] and are

* Corresponding author at: Department of Immunology, Genetics, and Pathology, Biomedical Center, SciLifeLab Uppsala, Uppsala University, SE-75108 Uppsala, Sweden.

E-mail address: Stefan.enroth@igp.uu.se (S. Enroth).

<https://doi.org/10.1016/j.fsigen.2023.102871>

Received 21 November 2022; Received in revised form 4 April 2023; Accepted 6 April 2023

Available online 7 April 2023

1872-4973/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

also reflective of anthropometrics, lifestyle, and the use of medication [13]. Previous studies have indicated that these associations may be strong enough to also be used for the prediction of these traits. An age- and height-predicting model based on plasma proteins developed by Enroth et al. [10] consisted of 29 and 26 proteins that predicted 50 % of the studied individuals within 5.4 years of their actual age and 5.4 cm of their actual height, respectively.

A challenging factor in actual forensic casework analysis is that DNA or mRNA molecules degrade because of environmental exposures (e.g., high temperature, ultraviolet light and humidity) on the crime scene, which complicates any phenotype analysis and the interpretation of results. A recent study [14] by Walker et al. compared DNA-methylation levels from DBS characterized with the Illumina 450 K bead chip with coupled EDTA plasma in the same individuals and found very high correlations between the two sample types. They were also able to replicate associates with both age and smoking status from DNA-methylation in DBS suggesting that at least for those traits, DNA-methylation based prediction from dried blood could be possible. A study by Dugue et al. [15] however also investigated the measurability of DNA methylation levels from dried blood spots (DBS) and concluded that although several DNA methylation sites showed acceptable technical performance, others did not, which restricted the selection of DNA methylation markers for DBS analysis. Other studies have investigated the measurability and correlation of proteins in wet plasma compared to DBS [16,17]. It was found that although the detectability of proteins was very high, the observed concentrations were sometimes different in DBS compared to plasma indicating that the predictive models will likely need to be retrained to fit the protein concentrations in DBS.

Here, we have investigated the possibilities and limitations of predicting several phenotypic traits using DNA methylation and plasma proteins separately or in combination. We have developed a prediction schema starting with the prediction of sex. This was followed by the sex-specific stepwise prediction of individual age consisting of an initial classification into broad age categories followed by category-specific linear regression models. We then developed sex-specific models for multiple anthropometric traits and, finally, lifestyle-related traits (e.g., current smoking status and the use of lipid-lowering drugs). These characteristics were selected as possible traits of interest within forensic casework after state-of-the-art research on previous results obtained in the field. The different models were evaluated in a stepwise manner to account for the influence of some traits on others and to minimise the number of features in order to produce reasonably sized models suitable for application in routine forensic casework.

2. Materials and methods

2.1. Samples

We used samples (Table 1) from the Northern Sweden Population Health Study (NSPHS) [18]. The NSPHS is a cross-sectional study aiming to investigate the effects of lifestyle and genetics on health and medical

conditions. The original samples were collected in two phases; 719 samples were collected in 2006, and an additional 350 samples were collected in 2009. For each participant, serum and plasma were prepared from blood samples and stored at -70°C on site. A questionnaire was used to collect data on medications and lifestyles. Anthropometrical measurements were carried out, and the questionnaire was filled out at the local health care centre in the presence of a district nurse. Current smoking status and use of specific medical drugs was recorded as 'yes' or 'no' and the latest encoded using the Anatomical Therapeutic Chemical (ATC) classification system. A total of 811 individuals were included in the study based on the availability of molecular data (i.e., DNA methylation and plasma proteins) (Table 1). For the stepwise age prediction, the individuals were classified as being 'younger' (age below 40, $N = 257$), 'middle aged' (ages 40–55, $N = 211$) or 'older' (ages above 55, $N = 343$) to reflect broad biological processing of ageing such as andropause or menopause and to provide groups of similar sizes. Individual data were handled under an ethics permit (Regionala Etikprövningsnämnden, Uppsala, Dnr. 2005:325, with approval of an extended project period on 19-03-2016). All the analyses and data storage were conducted on a secure server provided by Uppsala University.

2.2. Plasma proteins

The characterisation of plasma proteins from the NSPHS was performed as previously described [13]. In brief, concentrations of proteins in plasma were determined using a proximity extension assay (PEA) [19]. The PEA is an affinity-based assay. For each protein, a pair of oligonucleotide-labelled antibody probes bound to the targeted protein. If the two probes were in close proximity, a PCR target sequence was formed by a proximity-dependent DNA polymerisation event, and the resulting sequence was subsequently detected and quantified using real-time PCR. The resulting abundance levels were given in Normalized Protein eXpression (NPX) on a log₂-scale where higher NPX corresponded to higher protein concentrations. Each proximity extension assay has a lower detection limit calculated at run-time based on the controls that are included in each run, and here, measurements below these per-protein limits were removed from further analysis. The assay characteristics, including detection limit specifications, assay performance and validations, are available from the manufacturer (www.olink.com). Here, plasma proteins were characterised using five pre-assembled 'panels' defined by the manufacturer (Olink Target 96 Cardiovascular II and III, Inflammation, Neurology and Oncology II). The details of this particular dataset have been published previously, and after quality control, 477 unique proteins from 811 individuals were kept for analysis [13].

2.3. DNA methylation

The DNA methylation data used here has been previously described [3]. In brief, an Illumina Human Methylation 450k BeadChip was used according to standard recommendations from the manufacturer

Table 1
General statistics of protein (top rows) and methylation (bottom rows) datasets.

Sex	N	Age ^{a,b}	Height ^{a,b}	Weight ^{a,b}	BMI ^{a,b}	Smokers ^b	LL drug users ^b
All	811	49.3 (20.0)	164.1 (9.6)	72.1 (15.2)	26.7 (4.8)	101	58
Men	377	49.5 (19.8)	171.0 (7.4)	79.1 (14.4)	27.0 (4.5)	46	38
Women	434	49.1 (20.2)	158.2 (6.9)	66.0 (13.1)	26.4 (5.0)	55	20
Sex	N	Age ^a	Height ^a	Weight ^a	BMI ^a	Smokers	LL drug users
All	617	46.6 (20.5)	164.2 (9.6)	71.6 (15.3)	26.5 (4.8)	85	58
Men	285	46.7 (20.3)	171.0 (7.5)	78.8 (14.8)	26.9 (4.5)	38	38
Women	332	46.6 (20.7)	158.4 (7.0)	65.5 (12.9)	26.1 (5.0)	47	20

^a Data presented as mean (standard deviation).

^b Data presented in the following units: Age – Years, Height – cm, Weight – kg, BMI (body mass index) – kg/m², Smokers – current smoker, LL drug users – current user of lipid-lowering drugs.

(Illumina, San Diego, CA, USA). Annotation datafiles for the location of DNA methylation sites were downloaded from Illumina (Human-Methylation450_15017482_v.1.1.csv.gz accessed on the 8th of February 2021, www.illumina.com). The methylation sites included here in the modelling were preselected from the complete set based on previous genetic or epigenetic associations in the literature in relation to the traits of interest. The number of genes and methylation sites preselected for each of the analysed traits are presented in Table 2. For sex, height, adiposity and smoking all methylation markers from the array located within genes previously indicated in, for example, genome-wide association studies were selected, and for age, methylation markers previously included in age-prediction models were selected. All sources and references for this selection are presented in Table 2.

2.4. Statistical analysis

All data handling, model development, statistical evaluations of models and results were performed with R Studio (version 1.4.1717) [20] using the following add-on packages: Caret [21], dplyr [22], tidyverse [23], pROC [24], ggVennDiagram [25], fBasics [26], data.table [27], Glmnet [28,29], ggplot2 [30] and gridExtra [31]. After quality control, the final dataset consisted of 617 individuals with a complete data: 447 plasma proteins, DNA methylation levels at 4034 CpG sites and information on sex, age, height, body mass index (BMI), waist-to-height ratio (WTH), smoking status (current smoker) and self-reported usage of common medical drugs. No imputation of missing values was carried out and samples with any missing values were excluded from further analyses.

The models were trained on 75 % of the cohort using ‘glmnet’ and ‘naïve Bayes’ functions depending on the outcome (continuous or discrete). The models were optimised using cross-validation with five folds and tuning grids for the optimisation of alpha (range from 0.1 to 0.9) and lambda (range from 0.001 to 0.5) parameters on the training data only. The final models generated were then evaluated on the remaining 25 % of the cohort. The predictive models for the different traits were developed using the ‘glmnet’ method for regression models and the naïve Bayes method for classification models. For the naïve Bayes classification, supervised feature selection was used to first select a subset of informative predictors using the recursive feature selection tool (‘rfe’ in R). Model performance was evaluated in both training and testing sets using Pearson’s correlation and prediction errors (mean error and standard deviation error) for the glmnet function. Differences in model performance were evaluated using two-sided Wilcoxon ranked sum tests. Model performance for the naïve Bayes method was assessed based on accuracy, sensitivity, and specificity according to the following formulas: $sensitivity = \text{true positives} / \text{total of positives}$, $specificity = \text{true negatives} / \text{total of negatives and accuracy} = (sensitivity + specificity) / 2$. Sex prediction was done in terms of predicting ‘male’. Accuracies for two-class predictions were calculated with a cut-off at the ‘best-point’, that is, the closest point on the curve to perfect classification by Euclidean distance, determined in the training data and then applied to the testing sets.

Table 2

Literature-based pre-selection of markers for different traits (genes, methylation sites and source). For age, the selection was done on published CpG sites directly, not on genes.

Trait	Number of genes selected	Number of CpG sites selected	Refs.
Sex	30	597	[47–49]
Age		680	[5,6]
Height	51	988	[50–53]
Adiposity status	48	970	[54–57]
Smoking status	89	280	[58,59]

3. Results

3.1. Overall prediction strategy

The selected traits were modelled after considering possible factors affecting the resulting predictions. Traits such as age and sex were thus taken into account when developing the models for other traits. The final goal was to create a model pipeline where sex would be the first predicted trait, followed by age and then other characteristics using sex-dependent models (Supplementary Fig. S1). After stringent quality control (Methods) the final dataset consisted of 617 individuals with complete data records. Models were trained on a random selection of 75 % of the individuals and then evaluated in the remaining 25%. Feature selection was carried out only in the training proportion of the data.

3.2. Sex prediction

As sex is encoded in DNA, a model based on DNA methylation on the X chromosome will be highly accurate for the prediction of sex. In our data, a model was developed using a naïve Bayes classifier consisting of two CpG sites (cg07887243 and cg19246080, Supplementary Table 1) located in the MAOB and ARSD genes, both located on chromosome X. This model achieved perfect classification when evaluated with receiver operating characteristics (ROC) and area under the curve (AUC), with an AUC of 1.0 in both the training and testing datasets (Fig. 1A). We also developed a model based on the protein data that used six proteins (Leptin, MMP3, GH, ST2, MB, CD38; Supplementary Table 2). The protein-based model achieved an AUC of 0.89 in the training set and 0.89 in the testing set (Fig. 1B) with no statistical difference observed in the AUC performance between training and testing sets (DeLong’s test, $p = 0.72$), suggesting that our methodology produced robust models. All prediction measures for sex are listed in Supplementary Table 3.

3.3. Age prediction

The prediction of age was evaluated separately for men and women using two different approaches. First, a linear regression model was used to predict age directly from the data across the full age span of the available samples (‘direct model’, Supplementary Table 4). Second, a stepwise model consisting of an initial classification with naïve Bayes into three groups (< 40 years, > 55 years or ‘unclassified’) before an age-group-specific linear regression model was used to predict individual age. These age groups were selected as representatives of ‘younger’ (< 40), ‘middle aged’ (40–55) and ‘older’ (> 55). Individuals not confidently assigned (resulting probability > 0.8) to the extreme groups were labelled as unclassified in the initial step. For each of these two age groups, a linear regression model was developed and trained in the age span of the corresponding group. For the unclassified individuals, we reverted to a linear regression model trained across the whole age span of the data (Fig. 2). We then evaluated age predictions based on plasma proteins and DNA methylation separately or in combination. The models were trained both on the complete training set and then separately for women and men to assess the effect of sex on the prediction of age. Finally, we evaluated the performance of sex-specific models by applying the model trained on women to men and vice versa.

With the protein-only models, age was predicted with a Pearson’s correlation coefficient between the actual and predicted ages of $R = 0.94$ for women and $R = 0.92$ for men in the training sets and $R = 0.93$ in both testing sets (all p -values < machine precision, 2.2×10^{-16}), with a maximum standard error across any of the models of 7.6 years (Table 3). The initial age-group classification used only three proteins for men and eight for women, whereas the age-group specific linear models used 40 and 15 proteins respectively for younger and older groups, and 50 for the full age span model (Table 3). For DNA-methylation-only modelling, correlations between actual and predicted ages increased to $R = 0.99$ for both women and men in the

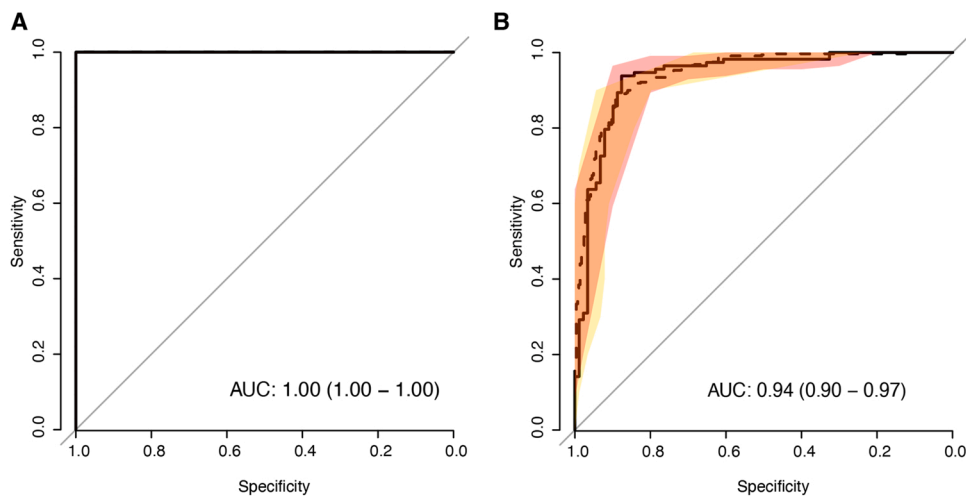


Fig. 1. Receiver operating characteristics (ROC) curves for sex predictions. **(A)** Methylation-based sex prediction models using training (dashed line, invisible) and testing (solid line) sets. The indicated area under the curve (AUC) is with respect to testing data, the black line represents the point estimate. **(B)** Protein-based sex prediction models using training (dashed line) and testing (solid line) sets. The coloured areas indicate the 95 % confidence interval surrounding the ROC curve obtained in the testing data (sensitivity in red and specificity in gold). The indicated AUC is with respect to the testing data, the black line represents the point estimate and coloured areas 95 % confidence interval.

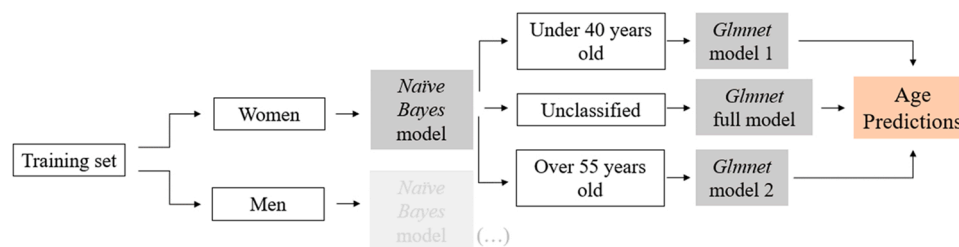


Fig. 2. Schematic view of the pipeline model developed for age predictions. The pipeline is detailed for women but mirrored also in men.

Table 3

Age prediction models based on protein and methylation levels, alone or in combination.

Model	Sub-grouping	Set	R ¹	Error ²	P-value (train. vs. test.) ³
PROTEINS	Females	Training	0.94	-0.53 (6.91)	0.24
	Females	Testing	0.93	0.43 (7.34)	
	Males	Training	0.92	0.18 (7.58)	
METHYLATION	Males	Testing	0.93	0.78 (7.28)	0.49
	Females	Training	1.00	-0.10 (2.09)	
	Females	Testing	0.98	-1.26 (4.06)	
PROTEINS + METHYLATION	Males	Training	0.99	-0.10 (2.96)	0.14
	Males	Testing	0.95	-1.30 (6.35)	
	Females	Training	1.00	-0.068 (1.79)	
PROTEINS + METHYLATION	Females	Testing	0.99	-1.12 (3.32)	0.21
	Males	Training	0.99	0.079 (2.56)	
	Males	Testing	0.95	-0.96 (6.51)	

^a Correlation factor estimate between actual and predicted (Pearson).

^b Data presented as mean (standard deviation) of errors in actual vs. predicted age in years.

^c Difference in error distribution between training and testing (two-sided Wilcoxon's ranked test).

training set and $R = 0.98$ and 0.95 , respectively, in the testing sets (Pearson's correlation, all p -values < machine precision, 2.2×10^{-16}), with a maximum standard error of 6.5 years (Table 3). Ten CpG markers were used in the age-group classifier developed for men and 15 in the model for women. In addition, 41 and 33 CpG markers were used for the continuous prediction of age within younger and older age groups, respectively, and 49 predictors were used for the full-age-span model. All protein and DNA methylation markers included in the analysis are listed in Supplementary Tables 2 and 3. Finally, models built from a combination of proteins and DNA methylation markers from the separate models were developed. The combined model achieved significant correlations between the actual and predicted ages of $R = 0.99$ for both women and men in the training set and between 0.95 and 0.99 for the testing sets (Pearson's correlation, all p -values < machine precision, 2.2×10^{-16}) and with a maximum standard deviation error between 3.3

and 6.5 years (Table 3). All selected proteins and DNA methylation sites are listed in Supplementary Tables 1 and 2 with detailed performance measures in Supplementary Table 5.

Stepwise model results were then compared to direct regression models with similar performances in the vast majority of comparisons (75 %, nominal p -value > 0.05 for difference in error distribution, Supplementary Table 6). However, when using direct regression models across the whole age span, the ages of older individuals tended to be predicted as younger than their actual age and conversely, younger individuals were predicted to be older than their actual age (Table 3). This phenomenon was not observed with the stepwise model. In addition, no difference was noted in the error distributions achieved between specific age-group models (Supplementary Tables 6 and 7) suggesting that initial modelling with age groups yielded more consistent error distributions across age ranges than a direct prediction model. Although age-

group-related mean errors and standard deviations showed significant differences in a few cases, no visible trends were found within the different modelling groups as observed above with, for example, older individuals being predicted as younger. Similar patterns were observed when comparing direct and pipeline models for correlation and mean error (Supplementary Table 7).

Finally, we evaluated sex-specific effects on age prediction and found a significant decrease in the prediction performance. For example, increased errors were found between actual and predicted ages from protein-based models when a woman's age was predicted using a model trained on men (p -value < 0.009 for training and testing sets, respectively; two-sided Wilcoxon's ranked test). For methylation-based models, a significant difference was observed when models trained on women were used to predict the age of men (p -value < 0.04 with training and testing sets, respectively; two-sided Wilcoxon's ranked test), implying that sex-specific age models provide more accurate predictions in these cases. All test statistics for sex-specific comparisons are listed in Supplementary Table 8.

3.4. Anthropometric prediction

We next set out to predict individual height, BMI (continuous and categorical) and WTH (continuous and categorical) ratios. The prediction models for height based on proteins alone achieved correlations (Pearson's R) between actual and predicted heights ranging from 0.74 to 0.81 in the training set and between 0.65 and 0.72 in the test set. These results corresponded to predictions (standard error) of the actual height within 5.1 and 5.8 cm for men and women, respectively, in the test data (Supplementary Table 9). The models based on DNA methylation achieved correlations (Pearson's R) between 0.58 and 0.85 in the training data and between 0.53 and 0.78 in the test set. These results corresponded to predictions (standard error) within 5.7–6.7 cm of the actual height in men and women in the testing proportion of the data. Lastly, a model based on both protein and DNA methylation levels was developed that achieved correlations (Pearson's R) between 0.82 and 0.92 in the training set and between 0.57 and 0.75 in the testing set. This corresponded to predictions within 5.7–6.7 cm of the actual height in the test data. All sex-specific models had slightly lower correlations than those obtained with the combined models, although no statistical differences were found in error distributions (all $p > 0.1$, two-sided Wilcoxon's ranked test). These models used between 15 and 68 features (Supplementary Tables 1 and 2) with fewer features, in general, for models based on DNA methylation compared to protein-based models. All prediction measures and statistical support for height are listed in Supplementary Table 9.

The same strategy was employed to predict BMI. Consistent performance was observed in training and testing sets with protein-based models, resulting in significant correlations (Pearson's R) between actual and predicted values; 0.76–0.85 with training sets and 0.79 and 0.82 with testing sets. The prediction errors were within 3.1–3.4 BMI units (kg/m^2) for the test data (Supplementary Table 10). The results showed a greater deviation when predictions were made using DNA methylation markers. Here correlations ranged from 0.46 to 0.70 with training sets and from 0.085 to 0.43 with test sets, corresponding to a prediction within 3.8–5.8 BMI units (kg/m^2) of the actual BMI values in the test data.

The combined models achieved correlations (Pearson's R) ranging from 0.86 to 0.90 in the training set and from 0.69 to 0.83 in the testing set corresponding to a prediction error within 2.1–2.4 for training data and within 2.8–3.3 for test data. All prediction measures and statistical support for continuous BMI predictions are listed in Supplementary Table 10.

Adiposity status was predicted based on BMI categories established by the World Health Organization (<https://www.euro.who.int/>). [32] Study individuals were classified into six groups using a naïve Bayes model: underweight (below $18.5 \text{ kg}/\text{m}^2$), normal weight

(18.5 – $25 \text{ kg}/\text{m}^2$), overweight (25 – $30 \text{ kg}/\text{m}^2$), obese I (30 – $35 \text{ kg}/\text{m}^2$), obese II (35 – $40 \text{ kg}/\text{m}^2$) and obese III (over $40 \text{ kg}/\text{m}^2$). Due to the low number of individuals in the most extreme categories, this modelling was only performed after combining men and women. The protein-only-based model used five proteins (LEPTIN, FURIN, FABP4, NCAN and IL6) with a multiclass accuracy of 0.64 for the training set and 0.55 for the testing set. The DNA methylation-based model used only two predictors (cg18473521 (HOXC4), cg19761273 (CSNK1D)) yielding less accurate predictions, with a multiclass accuracy of 0.54 and 0.48 for training and testing sets, respectively. Lower performance was observed for the testing set compared to the training set, suggesting that it might have been overfitted to the training proportion of the data. The model based on both proteins and DNA methylation markers used seven predictors in total and achieved a multiclass accuracy of 0.66 in the training data and 0.51 in the testing set. All prediction measures and statistical support for categorical BM predictions are listed in Supplementary Table 11.

Lastly, we used the data to predict WTH ratios with the same strategy as above. Here, protein-based models showed a similar performance using training and testing sets with correlations (Pearson's R) between predicted and actual values ranging from 0.81 to 0.84 in the training test and 0.80 and 0.82 in the test data, respectively (Supplementary Table 12). The predictions based on DNA methylation were less consistent than those from using protein models, with correlations ranging from 0.69 to 0.84 for the training set and 0.35 and 0.55 for the testing set. Next, we attempted to classify individuals into discrete groups, those with a 'high' (> 0.6) or a 'low' (< 0.6) WTH ratio. Using a naïve Bayes model based on five proteins, the obtained AUCs were 0.88 and 0.86 for training and testing sets, whereas we obtained AUCs of 0.78 and 0.79 for the model based on two DNA methylation markers and 0.86 and 0.89 for a combined model (Fig. 3A). In all tests, there were no significant difference between training and the testing sets (p -value > 0.4 , DeLong's test). All prediction measures and statistical support for continuous and categorical WTH predictions are listed in Supplementary Tables 12 and 13.

All proteins and DNA methylation markers included in the anthropometric models are listed in Supplementary Tables 1 and 2.

3.5. Lifestyle prediction

Models for predicting current smoking status (yes/no) of individuals were developed using a naïve Bayes method for men and women separately and in combination. To compensate for the imbalance in the dataset in smokers compared to non-smokers (Table 1), the data was down sampled to 79 smokers and 79 non-smokers in a training set, and 22 smokers and 22 non-smokers in a testing set. An analysis performed independent of sex resulted in a methylation-based model with eight markers (Supplementary Table 1) that achieved an AUC of 0.91 and 0.89 in training and testing sets, respectively (Table 4, Supplementary Table 14). Nine markers were used with each sex-specific model; the AUC of the female model was 0.96 with the training set and 0.88 with the testing set. The male model achieved an accuracy of 0.90 with the training set and 0.83 with the testing set. No significant differences were found between training and testing sets (all p -values > 0.32 , Supplementary Table 14). Similarly, protein-based models were created using from four to 18 markers (for men, women or combined) (Supplementary Tables 2 and 14) achieving AUCs ranging from 0.69 to 0.98 in the training and testing sets. Again, no significant differences were found between training and testing sets (all p -values > 0.20 , Supplementary Table 14) except for the female model that performed significantly better ($p = 0.036$) in the testing set compared to the training set (Supplementary Table 14). Lastly, models that combined protein and DNA-methylation markers achieved AUCs that ranged from 0.83 to 0.97 for both training and testing sets, without significant differences between such sets (Fig. 3B, all p -values > 0.21 , Supplementary Table 14). All prediction measures for smoking status are listed in Supplementary

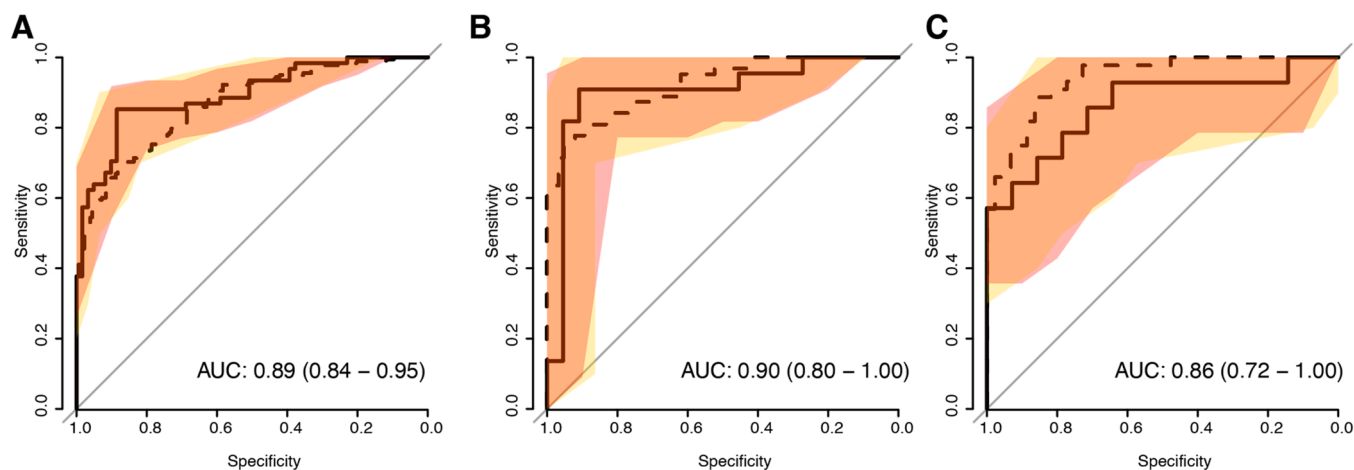


Fig. 3. Receiver operating characteristics (ROC) curves for waist-to-hip ratio, smoking status and use of lipid-lowering drugs. (A) Prediction of low/high waist-to-hip ratios using combined (methylation and protein) prediction models with performance in the training data indicated by a dashed line and in the testing data by a solid line. The coloured areas indicate the 95 % confidence interval surrounding the ROC curve obtained in the testing data (sensitivity in red and specificity in gold). The indicated area under the curve (AUC) is with respect to the testing data, the black line represents the point estimate and coloured areas 95 %. (B) As for (A) but for smoking status. (C) As for (A) but for use of lipid-lowering drugs.

Table 4

Smoking status prediction models based on protein levels and methylation markers alone or in combination.

Model	Subgrouping	Set	AUC ^a	P-value (train vs. test, DeLong)
PROTEINS	All	Training	0.89	0.71
		Testing	0.87	
METHYLATION	All	Training	0.91	0.67
		Testing	0.89	
PROTEINS + METHYLATION	All	Training	0.92	0.75
		Testing	0.90	

^a AUC, area under the curve.

Table 14.

In order to estimate the impact of smoking status on the prediction of age, we retrained the age prediction pipelines for men and women on a subset of the data containing only non-smokers. The models were then evaluated separately for male and female smokers and non-smokers (Supplementary Table 15). We found no statistical difference in the distribution of prediction errors in women ($p = 0.08$ for the training set and $p = 0.21$ for the testing set; two-sided Wilcoxon signed-rank test). However, on average, smokers were predicted to be 1.23 years older than their biological age. Among men, a clear trend existed towards smokers being predicted to be older than their biological age, with on average of 5.2 years in the training data and 5.3 years in the testing data. However, this difference was only statistically significant in the training proportion of the data ($p = 9.8 \times 10^{-6}$ in the training set and $p = 0.33$ in the testing set; two-sided Wilcoxon signed-rank test, Supplementary Table 15).

Finally, we built models predicting the use of lipid-lowering drugs. Drug users were defined as any individual using any drug catalogued under the C10A-code (ATC/DDD code index by the World Health Organization). In our dataset, these agents corresponded to ‘HMG CoA reductase inhibitors’ (C10AA), ‘fibrates’ (C10AB), and ‘other lipid modifying agents’ (C10AX). Because the number of individuals for this modelling was small, we did not build sex-specific models, but only a combined model. Using naïve Bayes, the resulting protein-based model consisted of 11 proteins and achieved AUCs of 0.94 and 0.86 for training and testing sets, respectively, with no statistical difference between sets (DeLong’s test, $p = 0.31$, Supplementary Table 16). The DNA

methylation-based model was built using 27 DNA methylation markers and achieved AUCs of 0.89 and 0.87 for training and testing sets, respectively (without a statistical difference found between sets, DeLong’s test, $p = 0.80$). Lastly, the model with protein and DNA methylation levels achieved AUCs of 0.94 and 0.86 for training and testing sets without a statistical difference found between sets (Fig. 3C, DeLong’s test, $p = 0.31$). For lipid-lowering drugs, all features included in the models are listed in Supplementary Tables 1 and 2 and all performance measures are listed in Supplementary Table 16.

4. Discussion

It is very common to find stains of biological material (e.g., blood, semen, saliva) at crime scenes. Therefore, being able to obtain information about the individual who deposited the biological material is important for quickly moving an investigation forward. Great progress has been made over the last few decades within forensic genetics in respect of making DNA analysis a primary tool for human identification. A routine analysis is based on short tandem repeats (STR), relying on a match between the STR profile from a crime scene sample with the profile of a suspect or a profile in a criminal database [33,34]. In the absence of suspects or database matches, other investigative tools are helpful to support the investigation with investigative leads. In these non-suspect cases, a mass population screening (or dragnet) may be a possible solution. However, such an approach is time consuming, would require enormous resources and be based on clear ethical guidelines. The main objective of this study was to provide prediction models for specific human traits that can help with investigative leads in forensic investigations. Predictions of appearance and lifestyle factors can limit the number of interesting individuals in a mass screening or limit the number of persons of interest (POIs) in a way that makes mass screening unnecessary. To date, we have developed models based on proteins and methylation markers in blood to predict seven traits (sex, age, height, BMI and WTH ratio, smoking status, and lipid-lowering drug usage) with good to excellent accuracies.

Sex and age models showed good predictive values using both protein and methylation levels; however, the age models showed marked accuracies and correlations using methylation levels. Our approach with first assigning individuals to an age group and then predicting continuous age also improved the overall performance of the predictions, specifically for those based on proteins although differences were not always statistically significant. In addition, when comparing predictions

based on proteins in stepwise modelling to a direct approach with continuous age-prediction, we did not observe a tendency to predict older individuals as younger (and vice versa). Several previous studies predicted age from DNA methylation and a common association is made with markers located in the *ELOVL2* gene, for instance in Johansson et al. [3]. Here the 'cg16867657' marker located in the promoter of the *ELOVL2* gene was present in all age-predicting models (both stepwise and direct). In conclusion, our data indicates that, when possible, methylation-based models should be the method of choice for the prediction of age and sex over protein-based models. However, when DNA is not available or usable, protein-based models can also offer informative predictions. Using methylation-based models resulted in accurate predictions when also assessing smoking status, suggesting that it would be possible to predict if blood collected at a crime scene came from a smoker or not. In this study, only self-reported current smoking status data was available; we could not investigate whether it was possible to predict high or low usage (i.e., the number of cigarettes smoked per day) nor distinguish former smokers from non-smokers. Previous studies have shown that it is also possible to separate former smokers from non-smokers [35] with high accuracy (AUC = 0.77) in addition to current smokers vs. non-smokers (AUC = 0.90) from only 13 DNA methylation markers for both sexes. In comparison to our current smoker models based on eight to nine DNA methylation markers, seven of the 13 markers were common. There were, however, some noteworthy differences. For instance, the two markers 'cg22132788' and 'cg12803068' located in the *MYO1G* gene were included in the 13-marker model but selected here only for the model predicting smoking status in women. Changes in DNA methylation patterns due to smoking in these two markers have previously been described to be present in children of at least 5.5 years of age and linked to prenatal exposure to maternal tobacco smoking [36]. In our data, protein-based models resulted in better prediction accuracies and correlations than DNA methylation for height, BMI, WTH ratio and lipid-lowering drug usage. For these traits, methylation-based models showed low correlations and accuracies, resulting in erroneous predictions. Here, DNA methylation levels originated from whole blood. That DNA methylation markers may result in a better performance if samples were obtained from specific tissues cannot be excluded. For BMI and the WTH ratio, a group classification was developed for well-established groups based on BMI and according to 'high' or 'low' WHT ratios. In the context of these classifications, it may be of interest to have at least a vague idea of such characteristics of the sample donor in forensic casework (for example, an eyewitness statement).

The main strength of this study was the availability of both plasma protein and DNA methylation measurements from a large number of well-characterised individuals. This allowed us to compare the performance of both data types and to use separated proportions of the cohort for training and testing, yielding realistic estimates of the performance. Overall, very high accuracies and correlations were obtained for predictions of both sex and age using DNA methylation alone or in combination with proteins, while good results were obtained for protein-based models (i.e., adipose status) and for methylation-based models in relation to smoking status. In parallel, a reduction in the number of predictors was carried out with the idea of also facilitating routine analysis in limited and degraded forensic samples. Across all models built in this study, 176 unique proteins and 283 unique methylation sites were used in the final models developed for the studied traits (Supplementary Table 17). As a comparison, the first molecular clock model suggested by Horvath [6] in 2013 was based on 353 unique methylation sites but later studies have used a lot fewer sites suggesting that improvements can likely be made in terms of number of ongoing variables. In comparison to those models, our stepwise model, based on DNA methylation only, predicted 50 % of individuals within 2.2 and 2.4 years for women and men, respectively, in the test proportion of the data compared to 3.6 years overall reported by Horvath. Here, the models based on proteins only predicted 50 % of individuals in the test

proportion of the data within 3.9 and 3.6 years for women and men, respectively. The model introduced by Hannum et al. [5] reported an overall RMSE of 4.9 years. Here, the stepwise model based on DNA methylations only had an RMSE of 4.2 and 6.4 years for women and men, respectively, in the test proportion of the data. More recently, models based on smaller sets of DNA methylation markers for predictions of age have been reported. Vidaki et al. [37] for instance, developed a model for forensic age prediction based on only 16 CpG sites to achieve a mean absolute error (MAE) of 4.4 years in their test set. Using the same error measures here, we achieved a MAE of 3.1 and 3.2 years in the test proportion of the data for women and men, respectively, using a stepwise model based on DNA methylation only. The same performance measures for the models based on proteins only were 5.2 and 5.3 years for women and men, respectively. The largest model here, the combined model predicting height, was built using a total of 130 proteins and DNA methylation markers. In the testing proportion of the data used here, this model explained 56.4 % of the variance in height. For traits such as height, which has a strong heritable component, genetic variance such as single nucleotide polymorphisms (SNPs) can also be used to explain observed variance between individuals. For height in particular, a recent investigation by Yengo et al. [38] attributed 40 % of the observed variance among individuals of European ancestry to 12, 111 SNPs. Although a limited comparison, using epigenetic and protein biomarkers seems to be a more promising route to highly accurate predictions of height than genetic variation encoded in DNA. Across all models built here, a total of 40 and 63 of the DNA methylation markers and protein biomarkers were present in at least two different prediction models, respectively (Fig. 4). The 'cg01820374' was the DNA methylation marker present in the highest number of models (age, BMI, waist-to-height ratio and lipid-lowering drug usage). This marker is located within the *LAG3* gene and has previously been associated specifically with predicting age [6], as a prognostic marker in triple negative breast cancer [39] as well as being associated with cardiovascular disease and all-cause mortality [40]. The protein present in the highest number of models (age, height, BMI, waist-to-height ratio and smoking status) was the 'neurocan core protein' (UniProtID: O14594, 'NCAN'). This protein has been associated [41] with a variety of biological processes such as cell adhesion, and skeletal system and central nervous development. DNA variation in the coding gene has previously been associated with bipolar disorder [42] although, to the best of our knowledge, no previous specific association has been shown in the traits modelled here. We also compared the models developed using only proteins with models developed using only methylation markers as predictors, which may provide a good alternative to the use of DNA for the prediction of most of the traits. Lastly, we also introduced a workflow for the sequential prediction of various traits to allow us to account for a possible interaction between different traits (e.g., sex and age).

Our study had several limitations. First, all analyses were based on samples of large amounts of biological material collected under controlled conditions that might not be comparable to samples collected at a crime scene. The samples used here have also been stored for several years in freezers that could affect molecular measurements, especially for measurable protein concentrations [43]. In addition, forensic samples are often challenged by environmental exposure leading to damage, degradation and decreased amounts of DNA, and most likely also degraded proteins. Therefore, forensic DNA assays often rely on a small set of targets, short amplification targets and optimised protocols that could affect the precision of the proposed modelling. Second, the individuals participating in the study were all from the northern parts of Sweden and a future replication of the study is warranted in other cohorts. Finally, the underlying number of observations for some of the traits, especially when considering sex-specific models, was low, which affects the certainty of the prediction measurements and the ability to obtain robustly performing models across both training and testing proportions of the data. Despite the limitations in the current study, we demonstrate the potential in using a combination of protein and

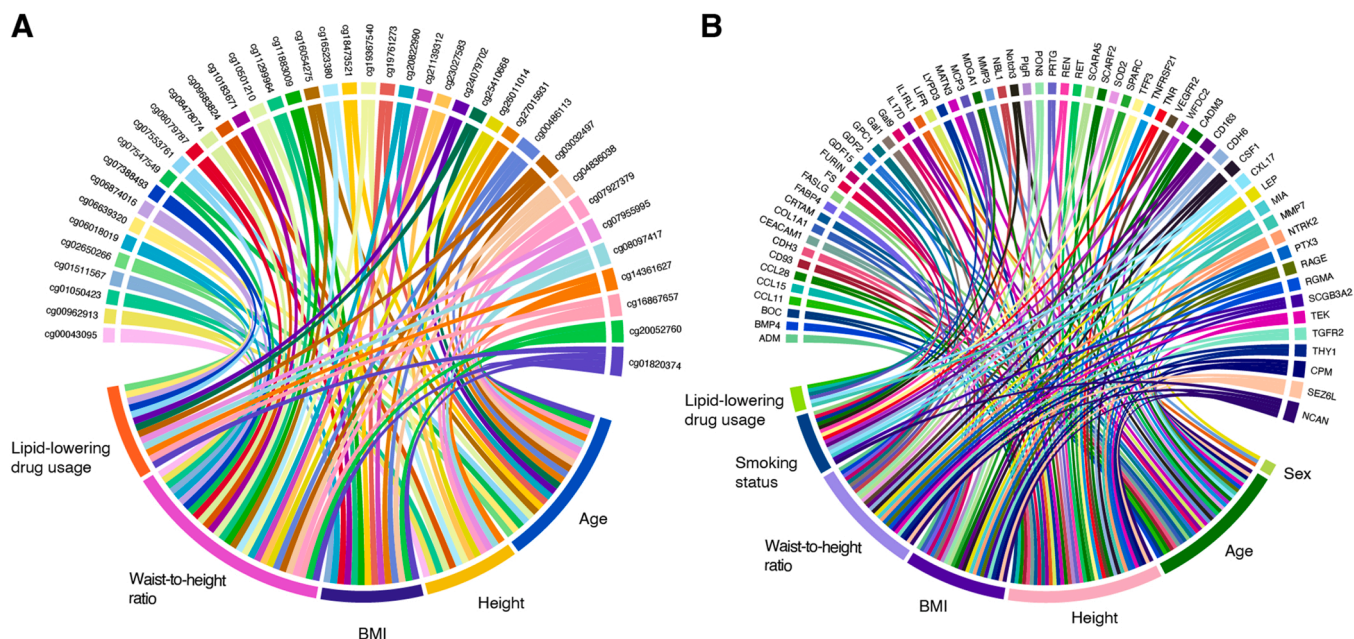


Fig. 4. Relationship between included markers and prediction models (A) DNA methylation markers (top) present in at least two prediction models (bottom). A connection between any two pairs represent an included marker in a specific model. (B) As for (A) but for plasma proteins.

methylation markers for predictions of both externally visible characteristics and lifestyle factors.

Here a total of 283 DNA methylation markers and 176 proteins were used across all models. Although this is a fairly large set of features, recent developments, especially in the affinity-based proteomics area, now allow for highly multiplexed characterisation of hundreds of proteins from minute amounts of material [44], including from dried blood spots [17]. Similarly, for DNA-methylation, previous studies have shown [14,15] that although it is technically possible to analyse hundreds of thousands of methylation markers using high-throughput assays from dried blood, there are also difficulties, and some markers cannot be reliably measured using this technology. In addition, not having previous knowledge of exactly which markers that can be reliably measured under which circumstances adds to the uncertainty of the prediction models. Future studies, ideally with larger sample sizes and denser data, will likely identify additional markers providing an equally good or more accurate prediction of individual characteristics, albeit based on fewer variables. When optimal markers have been identified and replicated, and the number of markers to use has been minimised, an optimal forensic assay can be designed. It is, however, likely that different sets of markers offer robust predictive performance in different sample types and substrates, such as dried blood on a solid surface or cloth. Here, we have employed a novel approach to the prediction of age starting from a broad characterisation of age groups based on semi-arbitrary cut-offs (e.g., under 40 and over 55 years). Although these precise cut-offs themselves have limited forensic profiling capabilities, the methodology could be expanded to additional age categories given a larger cohort. This could provide the capability to predict age with certainty and with a more acute forensic or jurisdictional relevance such as above or under the age of 18 or 21. Precise prediction of life-style related phenotypes through DNA-methylation, termed epigenomic lifestyle prediction [8] could offer the opportunity to build a very rich profile of an unknown individual. Traits, such as smoking status or the use of lipid-lowering drugs as analysed here, perhaps do not contribute to the potential identification of an individual based on visible characteristics but may still add valuable information [45]. With high-throughput DNA-methylation characterization, a large variety of traits ranging from alcohol consumption, disease risk and even dietary habits could potentially be predicted [8,45]. Although prediction beyond visible traits could help

identify an individual, there are several ethical concerns regarding individual privacy that needs to be considered [46] and predictions of this type should be used selectively and ideally based on national guidelines.

In conclusion, we have developed a stepwise approach for the de-novo prediction of personal characteristics from plasma protein and DNA methylation markers. These models are accurate and may provide valuable information and investigative leads in future forensic casework. These models may also be valuable in epidemiological studies, when such information is lacking, and these traits may be predicted from the analyses of biological samples.

CRediT authorship contribution statement

MOL analyzed the data, implemented the analysis pipeline and drafted the manuscript. ÅJ contributed data and reviewed the manuscript. UG contributed data and reviewed the manuscript. MA contributed to the manuscript writing and framing of the results. SE conceived of the study, supervised the data analysis, and contributed to the manuscript writing..

Funding

This data used in this study was funded by the Swedish Medical Research Council (ÅJ, UG), the Foundation for Strategic Research (UG), the Swedish Research Council (SE 2022-00857, MA 2022-02056) the Swedish Cancer Society (SE 220604FE) and the Swedish Collegium for Advanced Study (SE). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The datasets generated and/or analysed during the current study are

available from the authors on reasonable request.

Acknowledgements

Methylation profiling was performed by the SNP&SEQ Technology Platform in Uppsala, which is supported by Uppsala University, Uppsala University Hospital, Science for Life Laboratory (SciLifeLab) - Uppsala and the Swedish Research Council. PEA measurements were carried out by Olink Proteomics AB in Uppsala, Sweden. Parts of the computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project sens2016007.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2023.102871.

References

- [1] L. Chaitanya, et al., The HRisPlex-S system for eye, hair and skin colour prediction from DNA: introduction and forensic developmental validation, *Forensic Sci. Int. Genet.* 35 (2018) 123–135.
- [2] M. Kayser, P.M. Schneider, DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations, *Forensic Sci. Int. Genet.* 3 (2009) 154–161.
- [3] Å. Johansson, S. Enroth, U. Gyllensten, Continuous aging of the human DNA methylome throughout the human lifespan, *PLoS One* 8 (2013).
- [4] W. Besingi, Å. Johansson, Smoke-related DNA methylation changes in the etiology of human disease, *Hum. Mol. Genet.* 23 (2014) 2290–2297.
- [5] G. Hannum, et al., Genome-wide methylation profiles reveal quantitative views of human aging rates, *Mol. Cell* 49 (2013) 359–367.
- [6] S. Horvath, DNA methylation age of human tissues and cell types, *Genome Biol.* 14 (2013) R115.
- [7] C. Liu, et al., A DNA methylation biomarker of alcohol consumption, *Mol. Psychiatry* 23 (2) (2016) 422–433 (23, 2018).
- [8] A. Vidaki, M. Kayser, Recent progress, methods and perspectives in forensic epigenetics, *Forensic Sci. Int. Genet.* 37 (2018) 180–195.
- [9] S.C.E. Maas, et al., Validating biomarkers and models for epigenetic inference of alcohol consumption from blood, *Clin. Epigenet.* 13 (2021).
- [10] S. Enroth, S.B. Enroth, Å. Johansson, U. Gyllensten, Protein profiling reveals consequences of lifestyle choices on predicted biological aging, *Sci. Rep.* 1 (5) (2015) 1–10.
- [11] L. Folkersen, et al., Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease, *PLoS Genet.* 13 (2017).
- [12] S. Enroth, Å. Johansson, S.B. Enroth, U. Gyllensten, Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs, *Nat. Commun.* 5 (2014) 4684.
- [13] S. Enroth, et al., Systemic and specific effects of antihypertensive and lipid-lowering medication on plasma protein biomarkers for cardiovascular diseases, *Sci. Rep.* 8 (2018) 5531.
- [14] R.M. Walker, et al., Assessment of dried blood spots for DNA methylation profiling, *Wellcome Open Res.* 4 (2019).
- [15] P.A. Dugue, et al., Reliability of DNA methylation measures from dried blood spots and mononuclear cells using the HumanMethylation450k BeadArray, *Sci. Rep.* 6 (2016) 1–10.
- [16] J. Björkstén, et al., Stability of proteins in dried blood spot biobanks, *Mol. Cell. Proteom.* 16 (2017).
- [17] K. Broberg, et al., Evaluation of 92 cardiovascular proteins in dried blood spots collected under field-conditions: off-the-shelf affinity-based multiplexed assays work well, allowing for simplified sample collection, *BioEssays* (2021) 1–9, <https://doi.org/10.1002/bies.202000299>.
- [18] W. Igl, A. Johansson, U. Gyllensten, The Northern Swedish Population Health Study (NSPHS)—a paradigmatic study in a rural population combining community health and basic research, *Rural Remote Health* 10 (2010) 1363.
- [19] E. Assarsson, et al., Homogenous 96-Plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability, *PLoS One* 9 (2014), e95192.
- [20] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2020.
- [21] M. Kuhn, Variable selection using the caret package, *Caret Vignettes* (2012) 1–24.
- [22] H. Wickham, R. François, L. Henry, K. Müller, dplyr: A Grammar of Data Manipulation, Preprint at, 2021.
- [23] H. Wickham, et al., Welcome to the {tidyverse}, *J. Open Source Softw.* 4 (2019) 1686.
- [24] X. Robin, et al., pROC: an open-source package for R and S+ to analyze and compare ROC curves, *BMC Bioinform.* 12 (2011) 77.
- [25] C.-H. Gao, ggVennDiagram: A 'ggplot2' Implement of Venn Diagram, Preprint at, 2021.
- [26] D. Wuerzt, T. Setz, Y. Chalabi, fBasics: Rmetrics – Markets and Basic Statistics, Preprint at, 2020.
- [27] M. Dowle, A. Srinivasan, A.data.table: Extension of 'data.frame', Preprint at, 2021.
- [28] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, *J. Stat. Softw.* 33 (2010) 1–22.
- [29] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for Cox's proportional hazards model via coordinate descent, *J. Stat. Softw.* 39 (2011) 1–13.
- [30] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York, 2016.
- [31] B. Auguie, gridExtra: Miscellaneous Functions for 'Grid' Graphics, Preprint at, 2017.
- [32] WHO, Body Mass Index, World Health Organization, 2021. (<https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi>).
- [33] E. Místek, L. Halámková, I.K. Lednev, Phenotype profiling for forensic purposes: nondestructively potentially on scene attenuated total reflection Fourier transform-infrared (ATR FT-IR) spectroscopy of bloodstains, *Forensic Chem.* 16 (2019), 100176.
- [34] M. Kayser, Forensic DNA phenotyping: predicting human appearance from crime scene material for investigative purposes, *Forensic Sci. Int. Genet.* 18 (2015) 33–48.
- [35] S.C.E. Maas, et al., Validated inference of smoking habits from blood with a finite DNA methylation marker set, *Eur. J. Epidemiol.* 34 (2019) 1055–1074.
- [36] P. Rzehak, et al., Maternal smoking during pregnancy and DNA-methylation in children at age 5.5 years: epigenome-wide-analysis in the European Childhood obesity project (CHOP)-study, *PLoS One* 11 (2016), 155554.
- [37] A. Vidaki, et al., DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing, *Forensic Sci. Int. Genet.* 28 (2017) 225–236.
- [38] L. Yengo, et al., A saturated map of common genetic variants associated with human height, *Nature* 610 (7933) (2022) 704–712.
- [39] Y. Gao, et al., Identification of a DNA methylation-based prognostic signature for patients with triple-negative breast cancer, *Med. Sci. Monit.* 27 (2021) e930025–1.
- [40] L. Perna, et al., Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort, *Clin. Epigenet.* 8 (2016).
- [41] M. Uhlen, et al., A pathology atlas of the human cancer transcriptome, *Science* 357 (2017).
- [42] S. Cichon, et al., Genome-wide association study identifies genetic variation in neurocan as a susceptibility factor for bipolar disorder, *Am. J. Hum. Genet.* 88 (2011) 372.
- [43] S. Enroth, G. Hallmans, K. Grankvist, U. Gyllensten, Effects of long-term storage time and original sampling month on biobank plasma protein concentrations, *EBioMedicine* 12 (2016).
- [44] E. Assarsson, et al., Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability, *PLoS One* 9 (2014).
- [45] M. Shabani, P. Borry, I. Smeers, B. Bekaert, Forensic epigenetic age estimation and beyond: ethical and legal considerations, *Trends Genet.* 34 (2018) 489–491.
- [46] C. Dupras, E.M. Bunnik, Toward a framework for assessing privacy risks in multi-omic research and databases, *Am. J. Bioeth.* 21 (2021) 46–64.
- [47] B. Ho, et al., X chromosome dosage and presence of SRY shape sex-specific differences in DNA methylation at an autosomal region in human cells, *Biol. Sex Differ.* 9 (2018) 1–10.
- [48] E. Hall, et al., Sex differences in the genome-wide DNA methylation pattern and impact on gene expression, microRNA levels and insulin secretion in human pancreatic islets, *Genome Biol.* 15 (2014) 522.
- [49] I. Yusipov, et al., Age-related DNA methylation changes are sex-specific: a comprehensive assessment, *Aging* 12 (2020) 24057–24080.
- [50] K. Tatton-Brown, et al., Mutations in epigenetic regulation genes are a major cause of overgrowth with intellectual disability, *Am. J. Hum. Genet.* 100 (2017) 725–736.
- [51] K. Tatton-Brown, et al., Germline mutations in the oncogene EZH2 cause Weaver syndrome and increased human height, *Oncotarget* 2 (2011) 1127–1133.
- [52] P. Muthuirulan, T.D. Capellini, Complex phenotypes: mechanisms underlying variation in human stature, *Curr. Osteoporos. Rep.* 17 (2019) 301–323.
- [53] R. Tripaldi, L. Stuppia, S. Alberti, Human height genes and cancer, *Biochim Biophys. Acta Rev. Cancer* 1836 (2013) 27–41.
- [54] D. Fratantonio, F. Virgili, B. Benassi, Diet and Epigenetics: Dietary Effects on DNA Methylation, Histone Remodeling and mRNA Stability. *Comprehensive Foodomics*, 1, Elsevier, 2021.
- [55] X. Lu, et al., An epigenome-wide association study identifies multiple DNA methylation markers of exposure to endocrine disruptors, *Environ. Int.* 144 (2020), 106016.
- [56] A.B. Crujeiras, A. Diaz-Lagares, DNA Methylation in Obesity and Associated Diseases. *Epigenetic Biomarkers and Diagnostics*, Elsevier Inc., 2016, <https://doi.org/10.1016/B978-0-12-801899-6.00016-4>.
- [57] C. Schiano, et al., Epigenetic-sensitive pathways in personalized therapy of major cardiovascular diseases, *Pharm. Ther.* 210 (2020), 107514.
- [58] D. Fragou, E. Pakkidi, M. Aschner, V. Samanidou, L. Kovatsi, Smoking and DNA methylation: correlation of methylation with smoking behavior and association with diseases and fetus development following prenatal exposure, *Food Chem. Toxicol.* 129 (2019) 312–327.
- [59] L.P. Breitling, R. Yang, B. Korn, B. Burwinkel, H. Brenner, Tobacco-smoking-related differential DNA methylation: 27K discovery and replication, *Am. J. Hum. Genet.* 88 (2011) 450–457.