Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology) 2023 24(9):839-852
www.jzus.zju.edu.cn; www.springer.com/journal/11585
E-mail: jzus_b@zju.edu.cn

Research Article
https://doi.org/10.1631/jzus.B2200555

Construction and evaluation of in-house methylation-sensitive SNaPshot system and three classification prediction models for identifying the tissue origin of body fluid

Yating FANG[1,2], Man CHEN[1], Bofeng ZHU[1,3]✉

[1]*Guangzhou Key Laboratory of Forensic Multi-Omics for Precision Identification, School of Forensic Medicine, Southern Medical University, Guangzhou 510515, China*
[2]*School of Basic Medical Sciences, Anhui Medical University, Hefei 230031, China*
[3]*Microbiome Medicine Center, Department of Laboratory Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou 510515, China*

Abstract: The identification of tissue origin of body fluid can provide clues and evidence for criminal case investigations. To establish an efficient method for identifying body fluid in forensic cases, eight novel body fluid-specific DNA methylation markers were selected in this study, and a multiplex single base extension reaction (SNaPshot) system for these markers was constructed for the identification of five common body fluids (venous blood, saliva, menstrual blood, vaginal fluid, and semen). The results indicated that the in-house system showed good species specificity, sensitivity, and ability to identify mixed biological samples. At the same time, an artificial body fluid prediction model and two machine learning prediction models based on the support vector machine (SVM) and random forest (RF) algorithms were constructed using previous research data, and these models were validated using the detection data obtained in this study (n=95). The accuracy of the prediction model based on experience was 95.79%; the prediction accuracy of the SVM prediction model was 100.00% for four kinds of body fluids except saliva (96.84%); and the prediction accuracy of the RF prediction model was 100.00% for all five kinds of body fluids. In conclusion, the in-house SNaPshot system and RF prediction model could achieve accurate tissue origin identification of body fluids.

Key words: DNA methylation; Body fluid; Forensic identification; Single base extension reaction (SNaPshot); Machine learning

1 Introduction

Body fluid (or stain) is a kind of common biological evidence in forensic identification. Body fluid materials extracted from crime scenes contain a wealth of biological information closely related to the crime. Identifying the tissue origin of body fluid can establish a link between the biological evidence and the crime (Sijen and Harbison, 2021). For example, if the DNA extracted from the vagina of the female victim can be identified as containing the semen DNA of the suspect, it will be of great significance in the trial of rape. The traditional methods to identify the tissue origins of body fluids mainly focus on identifying the specific tissue cells or proteins in different body fluids by serological or immunological technologies, such as blood identification based on Luminol test. These methods are simple and convenient, and form the basis for the development of forensic body fluid identification. However, in addition to their low specificity, other limitations include the high demand for sample quality and quantity, which is not always possible due to the damage of body fluid sample; also, detecting body fluid samples under special conditions such as degradation and mixing can be challenging (Martin et al., 2006; Virkler and Lednev, 2009). Moreover, false-positive or false-negative results occur easily, resulting in a higher risk of misjudgment. With the rapid technological development of molecular biology, forensic researchers have begun to look for body fluid-specific markers at the molecular level for body fluid identification, including DNA methylation markers (Choung et al., 2021; Haddrill, 2021).

✉ Bofeng ZHU, zhubofeng7372@126.com
Bofeng ZHU, https://orcid.org/0000-0002-9038-2342

Received Nov. 3, 2022; Revision accepted Mar. 6, 2023;
Crosschecked Apr. 27, 2023; Published online June 27, 2023

DNA methylation is one of the most important epigenetic markers. It refers to the process of transferring a methyl group to the specific base under the catalysis of DNA methyltransferase (Mattei et al., 2022). The technology of detection requires a smaller amount of body fluid and only needs a single DNA extraction step for body fluid identification and subsequent DNA individual identification analysis. Therefore, DNA methylation is an ideal molecular marker for identifying the tissue origins of different body fluids (Lee et al., 2022). Previous studies have reported a variety of body fluid-specific DNA methylation markers. Nonetheless, the obtained results indicated that forensic identifications based on these methylation markers were influenced by multiple factors, such as genetic differences (Kader et al., 2020; Zhao et al., 2022). Moreover, there are still many body fluid-specific DNA methylation markers in the whole genome, which have not been explored, especially the specific markers for identifying menstrual blood and vaginal fluid (Park et al., 2014; Lee et al., 2015; Forat et al., 2016; Kader et al., 2020). In addition, the expression level of a DNA methylation marker is a number between "0" and "1" and its interpretation so far depends on empirical judgment, which may lead to inaccurate conclusions based on false-positive or false-negative results.

Methylation-sensitive single base extension reaction (SNaPshot) technology comprises site-specific DNA methylation analysis based on the capillary electrophoresis platform. It not only realizes the quantitative detection of multiple DNA methylation markers at the same time, but can also be implemented in primary forensic laboratories, which could greatly improve the detection efficiency in forensic practice (Dias et al., 2020; Hao et al., 2021).

In this study, we identified a novel set of eight body fluid-specific DNA methylation markers for venous blood, semen, vaginal secretion, saliva, and menstrual blood in the Chinese Han population based on previous results, and constructed a multiplex detection system for these markers based on methylation-sensitive SNaPshot technology. Moreover, an artificial prediction model and two machine-learning prediction models were constructed to meet the need of accurate prediction of the tissue origin of body fluid. Thus, we aim to provide a reliable method and detection technology platform for the accurate identification of the five common types of body fluids in forensic practice.

2 Materials and methods

2.1 Selection and primer design of DNA methylation markers for body fluid identification

For five kinds of body fluids (venous blood, semen, vaginal secretion, saliva, and menstrual blood) which are common in forensic practice, eight body fluid-specific DNA methylation markers (CpGs) were selected by the Illumina Infinium Methylation EPIC BeadChip Array (Illumina, San Diego, CA, USA) according to the following principles: (1) the absolute value of methylation expression level between one target body fluid type and the other four body fluid types is greater than 0.2; (2) the *P*-value calculated by the linear model (limma package) between one target body fluid type and non-target body fluid types is less than 0.05 or the adjusted *P*-value is less than 0.01; (3) the body fluid specificity of the selected CpGs was previously verified by the method of pyrosequencing (n= 100). At the same time, one completely unmethylated cytosine was selected as an internal quality control to test the efficiency of DNA bisulfite conversion.

The MethPrimer 2.0 online software (https://www.urogene.org/methprimer2) and Primer Premier 5.0 software (Primer, Canada) were used to complete the primer design of target DNA methylation marker amplification and single base extension (SBE). The design principles were as follows: (1) primer lengths of 15–35 bp; (2) range of melting temperature (T_m) values of primers from 45 °C to 70 °C; (3) lengths of all amplified products less than 300 bp; (4) no hairpin structure in the primers; (5) no dimer formation between the primers; (6) different fragment sizes of the SBE primers by adding CT at the 5' end of the primers to distinguish different CpGs within the SNaPshot system. All primers were synthesized by Sangon Biotech Incorporation (Shanghai, China), and they were shown in Table 1. The preparation of multiplex primers was carried out after the verification of primer specificity and efficacy of primer binding.

2.2 Body fluid sampling and DNA exaction

A total of 95 unrelated healthy Chinese Han volunteers (20–45 years old) provided 95 body fluid samples (15 venous blood samples, 20 semen samples, 20 vaginal secretion samples, 20 saliva samples, and 20 menstrual blood samples) and signed the written informed consents. All volunteers declared that they

Table 1 Information of PCR primers and SBE primers of eight CpGs in the multiplex SNaPshot system

ID	Target ID[a]	PCR/SBE	Primer sequence (5'→3')	Product size (bp)
CpG1	cg25922751	PCR-forward	AATATTGAGGTTTGATTTATGGTG	127
		PCR-reverse	AAACACCRCAACCTAATACAAAT	
		SBE-forward	GTAGGTTAGGGYGGGGATATATATT	26
CpG2	cg24301930	PCR-forward	GTATGGTTTTTGYGTTTATAGATGA	111
		PCR-reverse	AAAATAAACCCCRAACTAACCTC	
		SBE-forward	$(CT)_5$AAGTTGYGTTAGTTAGTAAGGTTTT	36
CpG3	cg03902386	PCR-forward	GTGATGAAAGATTATTAGTATTTGATGT	84
		PCR-reverse	CAATTTCCACATAAAAACACRAAAC	
		SBE-forward	$(CT)_{10}$AAGGTTGTGTTTTTTTAGTG	41
CpG4	cg03282313	PCR-forward	GTGATATGTAAGATTTGAATGAAGG	94
		PCR-reverse	ATAACAAAACCATAACTCAAATTCA	
		SBE-forward	$(CT)_{10}$TGGGTATATGGTTAGTGGTGGAATT	46
IQC	Internal quality control	PCR-forward	GTYGTTAGGTAGTAGTATTAGTAGGTTTAGT	125
		PCR-reverse	CAATACAAATAATATATCAAAAAACCA	
		SBE-forward	$(CT)_{12}$TAGGTAGTAGTATTAGTAGGTTTAG	50
CpG5	cg24772753	PCR-forward	GAYGATTTTTYGTGGTGGAGTATT	178
		PCR-reverse	AACCCAACACCAAACCTCTTT	
		SBE-forward	(CT)CCAACACCAAACCTCTTTATAAAAC	28
CpG6	cg05558714	PCR-forward	AGGAATTTTAYGGTAGGGATAGATT	95
		PCR-reverse	TTTCCATACTACTTTCCAAAAACA	
		SBE-forward	$(CT)_3$TCACTCTTTAAATAAAACCAATCCC	32
CpG7	cg05614346	PCR-forward	GAAGGAGGTTTAGTTGAGAGTTTG	85
		PCR-reverse	AAACCAAACCCACACCTACTAC	
		SBE-forward	$(CT)_5$CCTACTACCTACCTAACRAAAATAC	36
CpG8	cg07452397	PCR-forward	GTTTYGTTTTGGTAATAATTTTTG	102
		PCR-reverse	AAAACTACACTAACTCTTTCTTATCCA	
		SBE-forward	$(CT)_7$TTTCTTATCCATAAATAACACAAAC	40

[a] Target ID is a unique ID in the Illumina Infinium Methylation EPIC BeadChip array. CpGs: DNA methylation markers; ID: identification; PCR: polymerase chain reaction; SBE: single base extension; SNaPshot: single base extension reaction.

did not smoke, or take antibiotics or other drugs within three months at the time of body fluid sample collection. Semen and vaginal fluid samples were collected after the volunteers had no sexual or contraceptive activity within 72 h. Menstrual blood samples were collected on the second day of the volunteers' menstrual cycle. Saliva samples were taken on the premise that the volunteers did not eat within 1 h or gargle. Each venous blood sample was collected in an anticoagulant tube containing ethylene diamine tetraacetic acid (EDTA), then shaken, and stored at −20 ℃. The semen and saliva samples were stored in clean tubes at −20 ℃. Vaginal fluid and menstrual blood samples were stored on sterile cotton swabs, dried at room temperature, and then stored at −20 ℃. After the pet owners signed the written informed consents, two canine saliva and two cat saliva samples were collected and stored at −20 ℃ after drying. A total of six venous blood samples were separately collected from mice, rats, and pigeons, and stored at −20 ℃ in the anticoagulant tubes containing EDTA.

The genomic DNA of body fluid samples was extracted by the QIAamp DNA Investigator Kit (Qiagen, Hilden, Germany). DNA quantification was performed using the Qubit 4.0 Fluorometer (Applied Biosystems, Foster City, CA, USA). The bisulfite conversion of genomic DNA was carried out by the EZ DNA Methylation Gold Kit (Zymo Research,

Orange, CA, USA) according to the manufacturer's recommendations.

2.3 Target DNA methylation marker amplification and SBE

The Platinum Multiplex PCR Master Mix Kit (Applied Biosystems) was used to complete the multiplex PCR reaction (total volume 20 μL), which included 1 μL of converted DNA, 10 μL of Platinum Multiplex PCR Master Mix (2×), 2 μL of GC Enhancer, 2 μL of multiplex PCR primer (1 μmol/L), and 5 μL of sterile deionized water. The PCR reaction was performed on the Veriti 96-well Thermal Cycler (Applied Biosystems) under the following conditions: predenaturation at 95 ℃ for 2 min; denaturation at 95 ℃ for 30 s, annealing at 56 ℃ for 90 s, and extension at 72 ℃ for 30 s, for 35 cycles; then final extension at 60 ℃ for 30 min and hold at 4 ℃. The amplification products were purified by incubation at 37 ℃ for 1 h and 80 ℃ for 20 min in a 10-μL reaction system including 5.2 μL of amplified products, 2.5 μL of shrimp alkaline phosphatase (rSAP; 1 U/μL), 1.0 μL of rSAP buffer (10×), 0.3 μL of Exonuclease I (5 U/μL), and 1.0 μL of Exonuclease I buffer (10×).

The SNaPshot Multiplex Kit (Applied Biosystems) was used to conduct the multiplex SBE reaction (total volume of reaction system was 10 μL): 5 μL of SNaPshot Multiplex Mix, 1 μL of multiplex SBE primer, 1 μL of purified amplification product, and 3 μL of sterile deionized water. The reaction conditions were as follows: 96 ℃ for 10 s, 50 ℃ for 5 s, and 60 ℃ for 30 s, for 25 cycles. The SBE products were added with 1 μL of rSAP and 1.2 μL of rSAP buffer (10×), and then incubated at 37 ℃ for 1 h and 80 ℃ for 20 min to complete the purification.

2.4 Capillary electrophoresis for multiplex SNaPshot detection

According to the manufacturer's instructions, the DS-02 Matrix Standard Kit (Applied Biosystems) was used for the spectral correction of the five-color fluorescence of 3130*xl* genetic analyzer (Applied Biosystems) before the SBE product separation. Next, 2 μL of purified SBE products were mixed with 2 μL of formamide and 0.5 μL of GeneScan 120 LIZ Dye Size Standard (Applied Biosystems). Next, after denaturation at 95 ℃ for 5 min, the loading mixture was placed on ice for 3 min and put on a 3130*xl* genetic analyzer. Capillary electrophoresis was performed under the following conditions: hot plate temperature of 60 ℃; injection voltage of 2 kV; injection time of 12 s; electrophoresis voltage of 15 kV; and electrophoresis time of 25 min. Data Collection v3.0 software (Applied Biosystems) was used to collect the genotyping data of electrophoresis analyses. GeneMapper ID-x v1.3 software (Thermo Fisher Scientific, USA) was employed to analyze the SNaPshot electrophoresis data and collect the peak height of each amplicon. The methylation rate of each CpG was calculated according to Eq. (1) or (2) (Pan et al., 2020) shown below.

2.5 Tests of sensitivity, species specificity, and mixed body fluid sample

In order to test the sensitivity of this system, 10 ng/μL of DNA from each body fluid sample was extracted for bisulfite conversion and diluted with deionized water separately to 0.5, 1, 2, and 5 ng/μL. Venous blood or saliva DNA templates (n=10) were extracted from dogs, cats, rats, mice, and pigeons to test the species specificity of the in-house SNaPshot system. The DNA mixed samples of semen and menstrual blood, semen and saliva, or semen and vaginal fluid were prepared to test the forensic identification efficiency of the system for mixed body fluid samples. These DNA-mixed samples were divided into three groups with three different volume mixing ratios (1:1, 1:2, and 1:4).

2.6 Construction of three models for body fluid identification

The three different prediction models for five types of body fluids were constructed on basis of the pyrosequencing results of these eight CpGs obtained in our laboratory. A total of 90% of all data were

$$\text{Methylation level} = \frac{\text{peak height of cytosine}}{\text{peak height of cytosine} + \text{peak height of thymine}} \times 100\%, \quad (1)$$

$$\text{Methylation level} = \frac{\text{peak height of guanine}}{\text{peak height of guanine} + \text{peak height of adenine}} \times 100\%. \quad (2)$$

extracted as the training set to construct the prediction models, and the remaining 10% were used as the validation set to verify the accuracy of these three models. This study used the results of the in-house SNaPshot system as a testing set (n=95) to test the accuracy of the constructed prediction models. The heatmap of methylation expression levels was built by TBtools v1.082 software (Chen et al., 2020). The e1071 and randomForest packages (Alderden et al., 2018) in R software were used to complete the constructions and tests of support vector machine (SVM) and random forest (RF) models, respectively.

3 Results

3.1 Methylation-sensitive SNaPshot system for body fluid identification

The control DNA sample and control SBE primer in the SNaPshot Multiplex Kit were used for the quality control of capillary electrophoresis on a 3130*xl* genetic analyzer for the process of constructing an in-house SNaPshot system. The results showed that the product peaks of the positive controls obtained by this study were consistent with the product peaks in the electrophoretic profile provided by the kit instructions, and the negative control showed no specific product peaks in the electrophoretic profile (Fig. S1). A total of eight body fluid-specific CpGs and an internal quality control (IQC) were used to construct the methylation-sensitive SNaPshot system. The information and methylation levels of these CpGs were shown in Table 2.

As shown in Fig. 1, CpG1 and CpG2 exhibited specifically low expression in venous blood samples, and CpG3 and CpG4 also had specifically low expression in saliva samples. Moreover, CpG3, CpG4, and CpG7 exhibited specifically high expression in semen samples. The above five CpGs exhibited specifically high expression in the above three target body fluids. However, the four specific CpGs (CpG5, CpG6, CpG7, and CpG8) of menstrual blood and vaginal fluid showed smaller differences in the DNA methylation expression levels between target and non-target body fluids. Among these four CpGs, the methylation level of CpG5 was higher in menstrual blood but lower in other four body fluids. The CpG7 exhibited specifically moderate expression in menstrual blood, specifically high expression in semen, but low expression in other three body fluids. CpG6 and CpG8 exhibited lower expression in vaginal fluid and higher expression in the other four body fluids. Based on this system, methylation levels of these eight CpGs for 95 body fluid samples were detected and the results were shown in Fig. 2.

3.2 Artificial prediction model for body fluid identification

In order to facilitate the interpretation of the detection data of our multiplex SNaPshot system, the classification threshold of each CpG for the target body fluid was set based on the results obtained in our laboratory based on the researcher's experience as follows: (1) the detected sample could be identified as venous blood when the methylation values of CpG1 and CpG2 were less than 0.05; (2) when the methylation

Table 2 Information of the eight body fluid-specific DNA methylation markers

ID	Target ID[a]	Chromosome position[b]	Methylation level[c]				
			Venous blood	Saliva	Menstrual blood	Vaginal fluid	Semen
CpG1	cg25922751	12:124950720	**0.03±0.00**	0.32±0.16	0.29±0.12	0.57±0.23	**0.95±0.04**
CpG2	cg24301930	11:65408496	**0.05±0.01**	0.37±0.18	0.32±0.13	0.59±0.22	**0.90±0.02**
CpG3	cg03902386	5:36703583	0.32±0.09	**0.05±0.02**	0.29±0.09	0.28±0.11	**0.85±0.03**
CpG4	cg03282313	2:113545053	0.35±0.07	**0.05±0.02**	0.33±0.13	0.50±0.22	**0.96±0.03**
CpG5	cg24772753	2:171573419	0.09±0.08	0.09±0.08	**0.09±0.08**	0.09±0.08	0.09±0.08
CpG6	cg05558714	11:125828068	0.95±0.01	0.93±0.02	0.83±0.10	**0.43±0.25**	0.96±0.01
CpG7	cg05614346	5:176858608	0.03±0.01	0.03±0.00	**0.16±0.12**	0.04±0.03	**0.96±0.02**
CpG8	cg07452397	1:209740971	0.91±0.02	0.91±0.03	0.87±0.06	**0.53±0.23**	0.97±0.00

[a] Target ID is a unique ID in the Illumina Infinium Methylation EPIC BeadChip array. [b] Chromosome position indicates the genomic location by human reference genome 37 (GRCh37/hg19). [c] Methylation levels (mean±SD, n=20) are obtained from the data available in our laboratory, and the bold values indicate the methylation levels for the target body fluids. CpG: DNA methylation marker; ID: identification; SD: standard deviation.

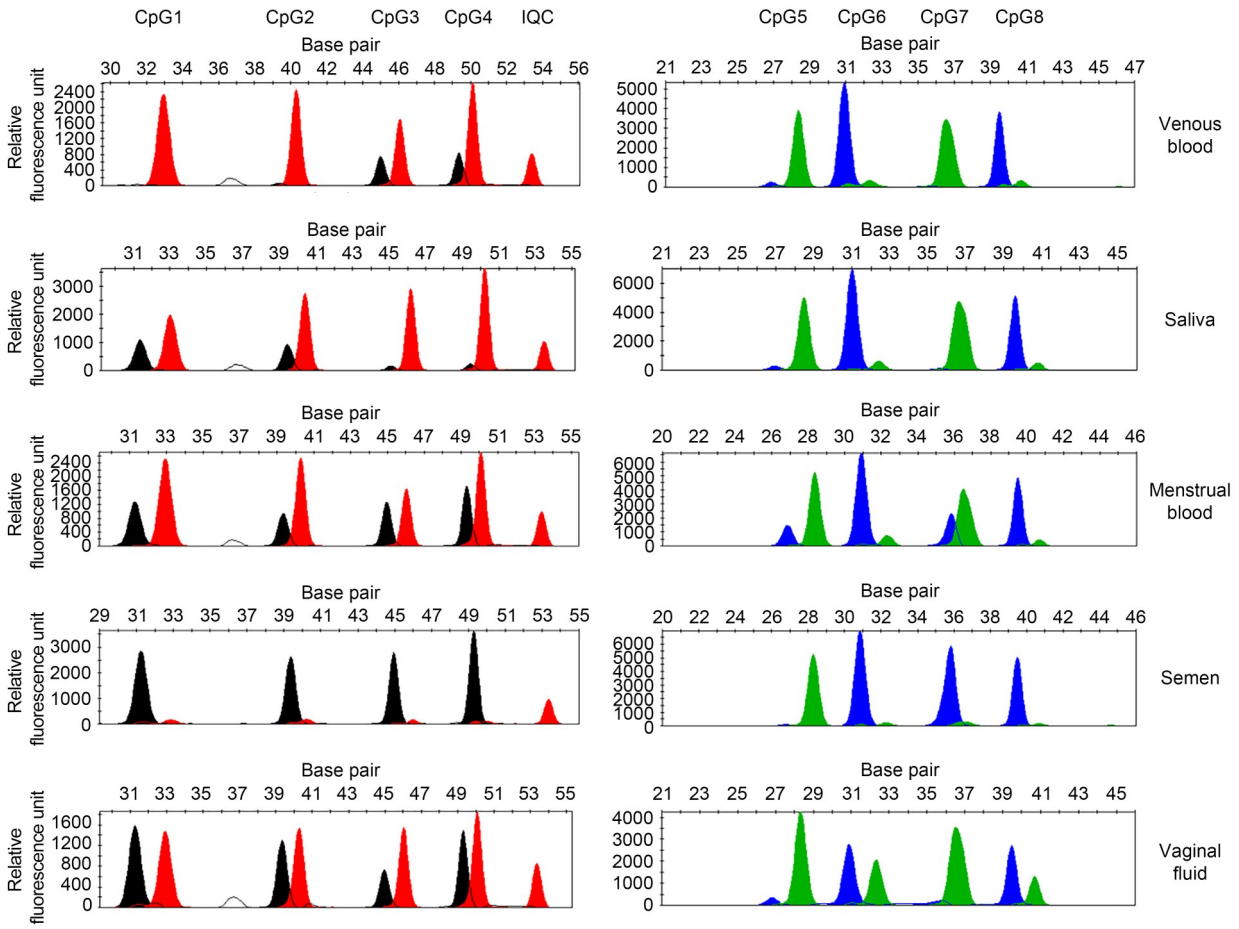

Fig. 1 Capillary electrophoresis profiles of eight body fluid-specific DNA methylation markers (CpGs) in five kinds of body fluids using the in-house multiplex SNaPshot system. The red peak represents thymine (T), the black peak represents cytosine (C), the blue peak represents guanine (G), and the green peak represents adenine (A). SNaPshot: single base extension reaction; IQC: internal quality control.

values of CpG3 and CpG4 were less than 0.10, it could be identified as saliva; (3) the menstrual blood could be identified if the methylation value of CpG5 was greater than 0.10 and the methylation value of CpG7 was between 0.10 and 0.50; (4) if the methylation values of CpG6 and CpG8 were less than 0.80, it could be identified as vaginal fluid; (5) the sample could be identified as semen when the methylation values of CpG3, CpG4, and CpG7 were greater than 0.80.

The 95 body fluid samples in this study were tested by the in-house SNaPshot system (the data were shown in Fig. S2). According to the above classification threshold of single CpG for target body fluids, it was found that the CpG1 and CpG2 successfully predicted all venous blood samples. Moreover, all semen samples were successfully predicted by CpG3, CpG4, and CpG7. Using the CpG3 and CpG4 simultaneously, all saliva samples were successfully predicted, but two venous blood samples were incorrectly predicted as saliva. All menstrual blood samples were successfully predicted by CpG5 and CpG7 at the same time, but three vaginal fluid samples were incorrectly predicted as menstrual blood. When CpG6 and CpG8 were used together to predict vaginal fluid samples, only one vaginal fluid sample was not successfully predicted, and four menstrual blood samples were also incorrectly predicted as vaginal fluid. Thus, the classification prediction accuracy of these CpGs was different for the five kinds of body fluids: it was the highest (100.00%) for venous blood and semen samples, followed by saliva (97.89%), menstrual blood (96.84%), and vaginal fluid (94.74%).

In addition, it is worth noting that CpG7 exhibited a specifically high expression in semen and low

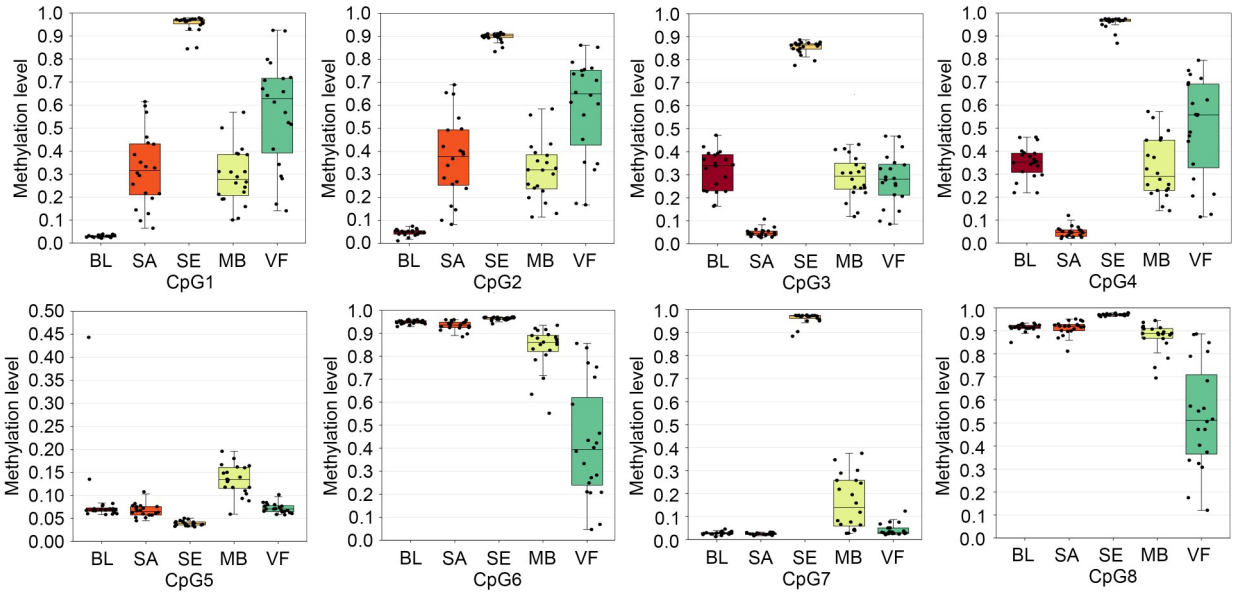

Fig. 2 Box plots of methylation levels for the eight body fluid-specific DNA methylation markers (CpGs) on the in-house multiplex SNaPshot system. BL: venous blood (*n*=15); SA: saliva (*n*=20); SE: semen (*n*=20); MB: menstrual blood (*n*=20); VF: vaginal fluid (*n*=20). SNaPshot: single base extension reaction.

expression in venous blood, saliva, and vaginal fluid samples. Therefore, when semen was mixed with only one of other three kinds of body fluids (venous blood, saliva, or vaginal fluid), the methylation value of CpG7 in the above mixed sample was between 0.10 and 0.50, and this could lead to misdiagnosis as menstrual blood on the premise that single-source menstrual blood also exhibited the methylation value of 0.10–0.50 at this CpG. Based on this possibility, a prediction model for body fluid classification has been developed, as shown in Fig. 3. Then, the SNaPshot dataset (*n*=95) was used as the testing dataset to test the artificial model. Finally, only four vaginal fluid samples were unsuccessfully classified in the testing set, and the overall classification prediction accuracy of the system was 95.79%.

3.3 SVM prediction model for body fluid identification

A binary classification SVM prediction model was constructed for each kind of body fluid. In the construction of the model, the target body fluid was assigned a value of 1, and the non-target body fluid was assigned a value of 0. Two SVM prediction models were constructed based on the linear kernel function and the radial kernel function for each kind of body fluid, respectively. Finally, a total of ten SVM models were constructed to identify five kinds of body fluids.

The prediction results of 95 body fluid samples based on ten SVM models were shown in Table 3. Compared with the models based on radial kernel function, the models based on linear kernel function exhibited fewer support vectors and higher prediction accuracy. Among the five prediction models for five kinds of body fluids based on the linear kernel function, the accuracy values of the prediction models for venous blood, menstrual blood, semen, and vaginal fluid were all 100.00%, as the Kappa, sensitivity, and specificity values were all 1.0000, while the saliva prediction performance of the SVM prediction model was relatively poor (the accuracy was 96.84%, the Kappa value was 0.9100, and the specificity was 0.8696).

3.4 RF prediction model for body fluid identification

Based on the dataset available in our laboratory, an RF prediction model for the five kinds of body fluids was successfully constructed. The out-of-bag (OOB) misclassification rate could be used to evaluate the classification effect of the RF model. As shown in Fig. 4, the misclassification rate of the trained OOB gradually decreased with the increase in the number of classification trees in the RF model. The OOB misclassification rate tended to be stable when the number of classification trees in the model was set to 100, which was 5.56%. The number of randomly selected variables was set to two at each split. The RF

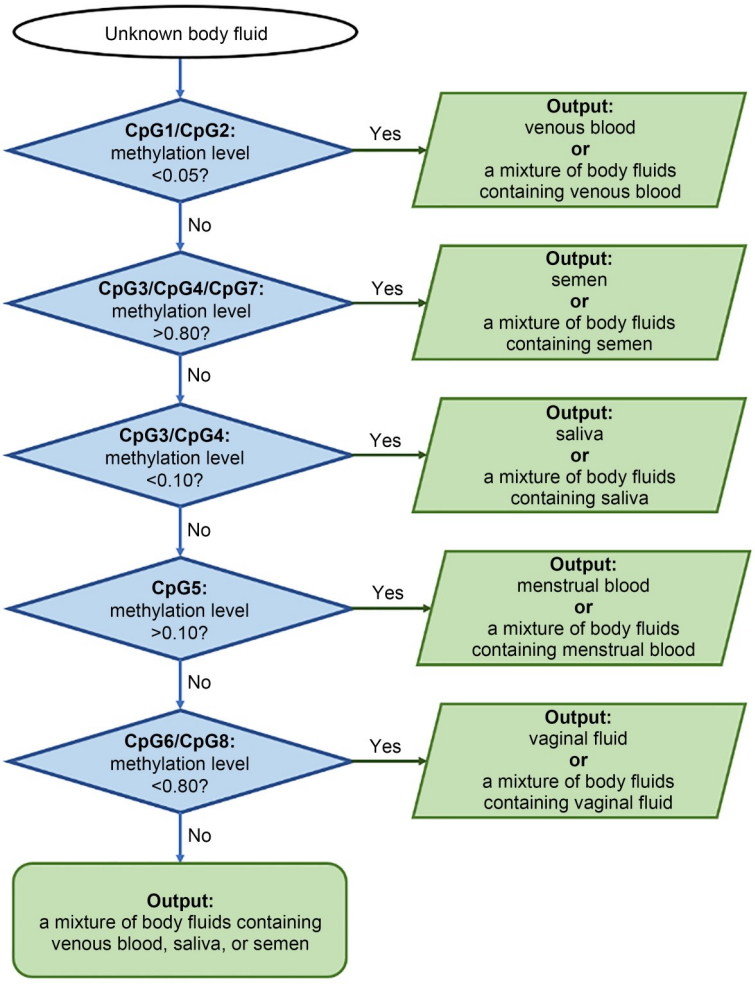

Fig. 3 Classification mode diagram of the artificial prediction model for five kinds of body fluids. CpG: DNA methylation marker.

Table 3 Performance measurement of ten support vector machine (SVM) models for five kinds of body fluids

Body fluid	Kernel function	Number of support vectors	Accuracy (%)	Kappa	Sensitivity	Specificity
Venous blood	Linear	11	100.00	1.0000	1.0000	1.0000
	Radial	24	97.89	0.9163	0.9756	1.0000
Menstrual blood	Linear	25	100.00	1.0000	1.0000	1.0000
	Radial	25	94.74	0.8257	0.9375	1.0000
Saliva	Linear	10	96.84	0.9100	1.0000	0.8696
	Radial	29	95.79	0.8820	1.0000	0.8333
Semen	Linear	5	100.00	1.0000	1.0000	1.0000
	Radial	10	100.00	1.0000	1.0000	1.0000
Vaginal fluid	Linear	11	100.00	1.0000	1.0000	1.0000
	Radial	23	96.84	0.8995	0.9615	1.0000

results showed that all 95 body fluid samples were successfully predicted with an accuracy of 100.00% (95% confidence interval (CI): 0.9619–1.0000) and a Kappa value of 1.0000.

In order to further analyze the performance of the RF prediction model for the five kinds of body fluids, a confusion matrix was used to visually compare the true values of the samples with the predicted values

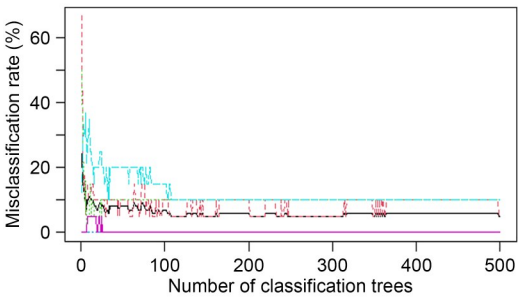

Fig. 4 Relationship between random forest size and the change of out-of-bag (OOB) misclassification rate.

(Fig. S3). In predicting each kind of body fluid, the target fluid was named as "positive" and all other non-target fluids were named as "negative." Meanwhile, the true-positive rate, true-negative rate, false-positive rate, false-negative rate, precision, and recall were calculated separately to perform the probability estimation of the prediction model (Table 4). The true-positive rate is defined as sensitivity, and the true-negative rate is specificity. False-positive rate represents the rate of falsely identifying negative sample as positive, also known as misdiagnosis rate, while false-negative rate refers to the rate of mistakenly identifying positive sample as negative, also known as missed diagnosis rate. Precision denotes the proportion of true positives among all samples identified as positive by the model. Recall represents the proportion of all actual positive samples that can be identified as positive by the model. Although the result of recall is the same as sensitivity, it is a retrospective measure of the model prediction results. The results showed that all body fluid samples were correctly predicted (the true-positive rates were all 100% and the false-negative rates were all 0%), and no other samples were incorrectly predicted (the false-positive rates were all 0% and the true-negative rates were all 100%). The precisions and recalls of the prediction model for the five kinds of body fluids were all 100%.

After the RF models were successfully constructed, the average decline accuracy (ADA) and the average Gini decline (AGD) were calculated to evaluate the importance of the eight CpGs in the prediction models. The ADA values for the identification of target body fluids at these CpGs were all greater than 0.20, except CpG4 (for semen), CpG7 (for menstrual blood and semen), and CpG8 (for vaginal fluid), while the ADA values of the five kinds of body fluids was between 0.07 and 0.15 (Table 5). CpG3 (cg03902386) was the most important among these CpGs in this

Table 4 Probability estimation of eight CpGs by the RF model for the identification of five kinds of body fluids

Body fluid	True-positive rate (%)	False-positive rate (%)	True-negative rate (%)	False-negative rate (%)	Precision (%)	Recall (%)
Menstrual blood	100	0	100	0	100	100
Saliva	100	0	100	0	100	100
Semen	100	0	100	0	100	100
Vaginal fluid	100	0	100	0	100	100
Venous blood	100	0	100	0	100	100

CpG: DNA methylation marker; RF: random forest.

Table 5 ADA values of eight CpGs in the RF model for the identification of five kinds of body fluids

ID	CpG ID	Menstrual blood	Saliva	Semen	Vaginal fluid	Venous blood	All five kinds of body fluids
CpG1	cg25922751	0.11	0.04	0.07	0.08	0.33	0.12
CpG2	cg24301930	0.13	0.04	0.10	0.06	0.30	0.12
CpG3	cg03902386	0.08	0.32	0.20	0.08	0.09	0.15
CpG4	cg03282313	0.08	0.32	0.14	0.09	0.08	0.14
CpG5	cg24772753	0.22	0.02	0.08	0.05	0.00	0.07
CpG6	cg05558714	0.07	0.06	0.07	0.27	0.08	0.11
CpG7	cg05614346	0.07	0.07	0.13	0.04	0.05	0.07
CpG8	cg07452397	0.05	0.04	0.16	0.15	0.03	0.08

ADA: average decline accuracy; CpG: DNA methylation marker; RF: random forest; ID: identification.

system, while the two menstrual blood-specific CpGs had relatively small values, which was similar to the results for AGD (Fig. S4).

3.5 Preliminary assessment for forensic application

The genomic DNAs from five kinds of common animals were extracted to test the species specificity of the multiplex SNaPshot system. The results showed that no effective peak was detected in the genomic DNAs of these animals (Fig. S5). According to the results of sensitivity tests of the multiplex SNaPshot system shown in Fig. 5, different amounts of genomic DNA input into the system would lead to the slight differences of DNA methylation rates. Moreover, all samples could be accurately identified by the artificial classification model except for 0.5 ng/μL of venous blood, and all body fluid samples could be accurately predicted by the RF prediction model.

In order to test the forensic efficacy of the system in identifying mixed body fluid samples, DNA templates from semen and menstrual blood, semen and saliva, and semen and vaginal fluid were separately mixed in volume ratios of 1:1, 1:2, and 1:4 with three sets of mixture samples per group. Based on the artificial model, we could successfully predict the following: menstrual blood components in nine DNA mixed samples of semen and menstrual blood; vaginal fluid components in nine DNA mixed samples of semen and vaginal fluid; the nine DNA mixed samples of semen and saliva as the mixed fluid samples not containing menstrual blood or vaginal fluid (Fig. 6). Moreover, the DNA methylation pattern of the mixed

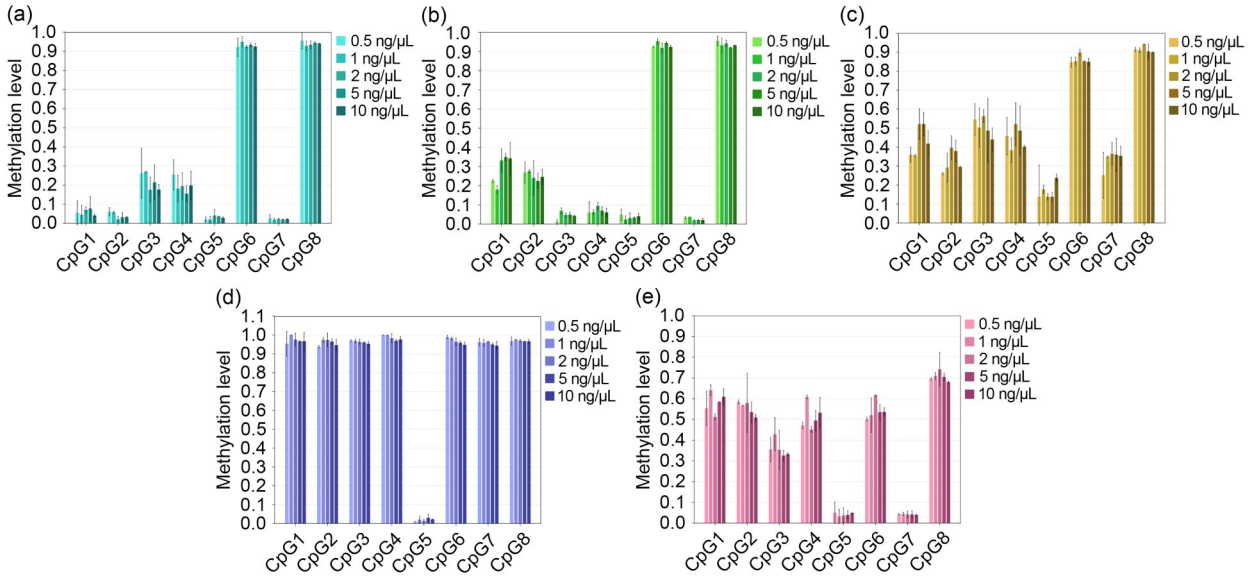

Fig. 5 DNA methylation levels detected by the in-house multiplex SNaPshot system with different DNA inputs of five kinds of body fluids: (a) venous blood; (b) saliva; (c) menstrual blood; (d) semen; (e) vaginal fluid. Data are expressed as mean±SD, *n*=2. CpG: DNA methylation marker; SNaPshot: single base extension reaction; SD: standard deviation.

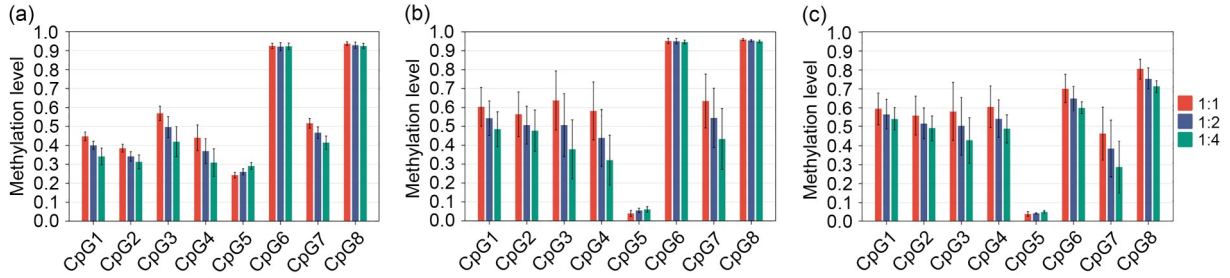

Fig. 6 DNA methylation levels of mixed DNA samples with different volume ratios from different body fluids detected by the in-house multiple SNaPshot system: (a) semen and menstrual blood; (b) semen and saliva; (c) semen and vaginal fluid. Data are expressed as mean±SD, *n*=3. CpG: DNA methylation marker; SNaPshot: single base extension reaction; SD: standard deviation.

samples gradually tended to resemble the contents of body fluids with larger proportions in the mixture.

4 Discussion

Previous studies have identified new body fluid-specific DNA methylation sites from the whole human genome, and conducted their systematic validation and forensic evaluation. However, those strategies might not be optimal for the tissue traceability of healthy, young, and middle-aged Chinese people (20–45 years old). The present study constructed a novel approach of body fluid identification based on DNA methylation markers for these people. It is worth mentioning that in addition to venous blood, menstrual blood also contains a lot of mucus, inflammatory cells, exfoliated endometrium and vaginal epithelial cells, etc. Both menstrual blood and vaginal secretion belong to the body fluids of vaginal origin. Saliva is a secretion derived from the oral cavity, which contains mucus, lysozyme, epithelial cells, and other components. Menstrual blood and saliva also contain epithelial cells, inflammatory cells, and so on. The components contained in these body fluids are heterogeneous and are easily affected by environmental factors. Therefore, in the process of sample collection, we need to strictly control the conditions and methods of sampling. In this study, the influences of medication, smoking, drugs, sexual behavior, as well as dietary factors on these body fluids were eliminated at the time of sampling. The same was applied for the sampling time within the menstrual cycle; sampling collection was performed on the second day of the menstrual cycle from volunteers without gynecological diseases.

We have previously identified eight novel body fluid-specific CpGs based on the Illumina Infinium Methylation EPIC BeadChip array to identify five kinds of body fluids, and verified the consistency of two genotyping results of the SNaPshot system and the pyrosequencing method. Although the latter method has advantages in quantitative and qualitative analyses of DNA methylation for specific gene fragments, its application in forensic science is limited due to its inability to achieve the composite detection of multiplex CpGs. In contrast, methylation-sensitive SNaPshot technology can not only realize docking with the forensic DNA laboratory platform, but also can achieve the simultaneous quantitative detection of multiplex CpGs, which greatly improves its detection efficiency. This might be conducive to popularization, application, and transformation of DNA methylation research achievements in forensic casework. Therefore, this study developed an in-house system containing eight novel body fluid-specific CpGs for body fluid identifications based on the methylation-sensitive SNaPshot technology, and conducted a thorough evaluation of this system.

In the analyses of the methylation-sensitive SNaPshot system results, CpGs were determined according to the location of product peaks in the electrophoresis profiles. In principle, the location of the product peak at the CpG should be consistent with the size of the SBE product fragment. However, due to the effect of different dyes on DNA fragment migration, the fragment size of the CpG product peak in the electrophoretic pattern could be different from the actual size of the SBE product fragment in electrophoresis, and this difference was more obvious for shorter SBE product fragments. Meanwhile, four kinds of dideoxynucleoside triphosphate (ddNTP) used in the SBE reaction were separately labeled with different fluorescent dyes. Thus, the CpG typing result was determined according to the color and location of the product peak. However, due to the different migration rates of the same sizes of DNA fragments by different dyes in a CpG marker, the two product peaks of the same CpG marker in a body fluid sample were not at the same location. Firstly, the black product peak (cytosine, C) appeared before the red product peak (thymine, T), and the blue product peak (guanine, G) appeared before the green product peak (adenine, A). Secondly, the height value of the product peak should be extracted after the product peak was determined, and the methylation rate of each CpG was calculated by the ratio of two peak heights, which was less affected by the amount of DNA input. Through preliminary test and evaluation, the converted DNA above 0.5 ng/μL was successfully detected by the system, and there was no significant difference in DNA methylation rate, which confirmed the high sensitivity of this system. At the same time, the results of our method could not be interfered with by the DNA of dogs, cats, rats, mice, or pigeons, which indicated that it had good species specificity.

After the establishment of the methylation-sensitive SNaPshot system, the next important step in our study was finding the means to interpret the

obtained data from the CpG typing result to predict the tissue source of the body fluid. Mixed samples can be identified by the artificial classification model. For example, the methylation levels of CpG1 and CpG2 in venous blood were less than 0.05, while those in other four kinds of body fluids were all higher. Therefore, if these two CpGs were not all less than 0.05, it might be either one of other four kinds of body fluids or a mixed body fluid (with or without venous blood). The same was true for semen or saliva identification from a single source. However, CpG5 exhibited methylation levels of less than 0.10 in other four kinds of body fluids except menstrual blood. Thus, if the methylation level of CpG5 was greater than 0.10, it only indicated that menstrual blood was contained in this body fluid (a single source of menstrual blood or a mixed body fluid including menstrual blood). The same was true for the identification of vaginal fluid. Among the 95 single-source body fluid samples for testing, the accuracy of this system in body fluid classification was 95.79% (only four vaginal fluid samples were not classified successfully). At the same time, 27 mixed samples were successfully classified in our study. It is worth noting that although this simple and intuitive classification mode was easy to operate, there was a risk that it might misjudge body fluid origin. For example, the venous blood sample with a small amount of DNA (<0.5 ng/μL) might lead to a false-positive result, because the model relies on the body fluid-specific classification thresholds of these CpGs. These thresholds are static and absolute, and the results of capillary electrophoresis data analyses may have small fluctuations due to the too small DNA injection volume, electrophoretic drift, etc. When the fluctuation occurs above and below the static threshold, it will easily lead to misjudgment. To reduce this misclassification rate, we suggest that multiple tests are performed for DNA below 0.5 ng/μL to obtain DNA methylation expression levels with smaller error.

In order to avoid the bias caused by the artificial threshold, machine-learning algorithms were introduced in this paper to achieve more accurate identification of body fluids (Deo, 2015; Bi et al., 2019). SVM is a novel learning method based on kernel function to realize the nonlinear mapping of data from low-dimensional space to high-dimensional space, which simplifies the derivation process in traditional statistical methods (Huang et al., 2018). However, it can only solve the problem of binary classification. In this study, we used dichotomous thinking to solve the classification problem of five kinds of body fluids, and constructed ten SVM prediction models based on two kernel functions. To improve the efficiency of the detection models for body fluid identifications, we also used the RF algorithm to construct a multi-classification prediction model, hoping to complete the prediction of all five kinds of body fluids simultaneously using one integrated model. RF is a kind of algorithm that uses ensemble strategy to train multiple decision trees randomly and synthesize all the results for predictions (Che et al., 2011; Chowdhury et al., 2019). The randomness is reflected in both sample and feature selection. In addition, the Bagging algorithm is used to create random samples in the RF algorithm to avoid overfit.

After the machine-learning models were constructed, the corresponding parameters were calculated to evaluate the performance of these classifiers. The prediction accuracy rate refers to the prediction success ratio of the model for body fluid identification. Since the dataset tested by the two kinds of machine-learning models in this study was uneven in each category (15 samples in the venous blood group and 20 samples in each of the other four body fluid groups), the Kappa value was calculated to evaluate the consistency between the predicted classification output by the model and the actual classification of body fluid sample, so as to correct the impact of the size differences among different kinds of sample datasets on the prediction accuracy (de Raadt et al., 2019). Meanwhile, we also calculated the model sensitivity and specificity to weigh whether the classifier was too conservative or too radical for classification prediction (Engstrand and Moeller, 1967; Benn Torres et al., 2019). Among the five SVM models based on linear kernel function, the accuracy, Kappa value, sensitivity, and specificity were all 1.0000 in the prediction models of venous blood, menstrual blood, semen, and vaginal fluid samples, while the prediction performance of the saliva prediction model was relatively low (the accuracy was 96.84%, and the Kappa value was 0.9100). This might be due to the low specificity of the SVM model (0.8696), which means that the model was too radical in the classification prediction of saliva. By comparison, the RF prediction model could not only identify five kinds of body fluids at the same time, but also exhibited higher prediction accuracy, sensitivity, and specificity

for the five kinds of body fluids (100.00%), which were better than those for the SVM models. Moreover, the prediction accuracy of the RF model was higher than that of the artificial prediction model (95.79%), which further confirmed that the machine-learning algorithm had high application value in the accurate identification of body fluid based on DNA methylation.

In this study, when the SNaPshot system based on eight novel CpGs was combined with the RF prediction model, the prediction accuracy values for venous blood, saliva, semen, menstrual blood, and vaginal secretion samples were all 100.00%. At the same time, the prediction probability for the five kinds of body fluids also reached 100.00%, which was significantly higher than that of previously reported systems and models (Tian et al., 2020; Huang et al., 2022). The SNaPshot system and classification models constructed in this study have some limitations. First of all, we failed to find more ideal menstrual blood specific CpGs. Moreover, the contributions of the two menstrual blood-specific CpGs were the least in the RF model. Secondly, the two machine-learning classification models could only be used for the identification of single-source body fluid because of the lack of mixed-sample data in the process of system construction. Therefore, it is necessary for future research to identify more high body fluid-specific DNA methylation markers, and detect more mixed samples of body fluids to build more refined classification patterns and machine-learning prediction models for the accurate determination of tissue origin inference of a single body fluid. Furthermore, to meet the need of forensic application, the SNaPshot system and prediction models constructed in this study still need to be further tested on a large number of different body fluid samples, such as samples from different age groups, and different sampling time during the menstrual cycle.

5 Conclusions

In this study, a methylation-sensitive SNaPshot system was constructed based on eight novel body fluid-specific CpGs. The in-house developed system showed good species specificity, high sensitivity, and a good ability to identify mixed samples. Based on previous results, we artificially set the classification mode of the SNaPshot system for five kinds of body fluids. Through the tests, the average accuracy of the classification standard for the selected body fluids was 95.79%. In addition, we constructed binary classification SVM prediction models and a multi-classification RF prediction model. Compared with the SVM models, the RF model could not only predict five kinds of body fluids at the same time, but also had better performance. When using the RF prediction model, the prediction accuracy was 100.00%, and the prediction sensitivity and specificity were both 1.0000. At the same time, the prediction accuracy of the multi-classification model was also higher than that of the artificial classification model. Overall, the methylation-sensitive SNaPshot system and the multi-classification prediction RF model constructed in the present study could accurately trace the tissue source of five kinds of body fluids and form a good basis for practical applications in forensic science.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 81930055 and 81772031).

Author contributions

Yating FANG performed the experimental research and data analysis, and wrote and edited the manuscript. Man CHEN performed the preliminary screening and full-text evaluation. Bofeng ZHU contributed to the study design, data analysis, and writing and editing of the manuscript. All authors have read and approved the final manuscript, and therefore, have full access to all the data in the study and take responsibility for the integrity and security of the data.

Compliance with ethics guidelines

Yating FANG, Man CHEN, and Bofeng ZHU declare that they have no conflict of interest.

The Ethics Committees of Anhui Medical University and Southern Medical University Health Science Center approved all processes (Approval No. 83230248). All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2013. Informed consents were obtained from all volunteers for being included in the study.

References

Alderden J, Pepper GA, Wilson A, et al., 2018. Predicting pressure injury in critical care patients: a machine-learning model. *Am J Crit Care*, 27(6):461-468. https://doi.org/10.4037/ajcc2018525

Benn Torres J, Martucci V, Aldrich MC, et al., 2019. Analysis of biogeographic ancestry reveals complex genetic histories for indigenous communities of St. Vincent and Trinidad.

Am J Phys Anthropol, 169(3):482-497. https://doi.org/10.1002/ajpa.23859

Bi QF, Goodman KE, Kaminsky J, et al., 2019. What is machine learning? A primer for the epidemiologist. *Am J Epidemiol*, 188(12):2222-2239. https://doi.org/10.1093/aje/kwz189

Che DS, Liu Q, Rasheed K, et al., 2011. Decision tree and ensemble learning algorithms with their applications in bioinformatics. *In*: Arabnia HR, Tran QN (Eds.), Software Tools and Algorithms for Biological Systems. Springer, New York, p.191-199. https://doi.org/10.1007/978-1-4419-7046-6_19

Chen CJ, Chen H, Zhang Y, et al., 2020. TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol Plant*, 13(8):1194-1202. https://doi.org/10.1016/j.molp.2020.06.009

Choung CM, Lee JW, Park JH, et al., 2021. A forensic case study for body fluid identification using DNA methylation analysis. *Leg Med*, 51:101872. https://doi.org/10.1016/j.legalmed.2021.101872

Chowdhury AR, Chatterjee T, Banerjee S, 2019. A random forest classifier-based approach in the detection of abnormalities in the retina. *Med Biol Eng Comput*, 57(1):193-203. https://doi.org/10.1007/s11517-018-1878-0

Deo RC, 2015. Machine learning in medicine. *Circulation*, 132(20):1920-1930. https://doi.org/10.1161/circulationaha.115.001593

de Raadt A, Warrens MJ, Bosker RJ, et al., 2019. Kappa coefficients for missing data. *Educ Psychol Meas*, 79(3):558-576. https://doi.org/10.1177/0013164418823249

Dias HC, Cordeiro C, Pereira J, et al., 2020. DNA methylation age estimation in blood samples of living and deceased individuals using a multiplex SNaPshot assay. *Forensic Sci Int*, 311:110267. https://doi.org/10.1016/j.forsciint.2020.110267

Engstrand RD, Moeller G, 1967. Confusion matrix analysis for form perception. *Hum Factors*, 9(5):439-446. https://doi.org/10.1177/001872086700900507

Forat S, Huettel B, Reinhardt R, et al., 2016. Methylation markers for the identification of body fluids and tissues from forensic trace evidence. *PLoS ONE*, 11(2):e0147973. https://doi.org/10.1371/journal.pone.0147973

Haddrill PR, 2021. Developments in forensic DNA analysis. *Emerg Top Life Sci*, 5(3):381-393. https://doi.org/10.1042/etls20200304

Hao T, Guo JL, Liu JD, et al., 2021. Predicting human age by detecting DNA methylation status in hair. *Electrophoresis*, 42(11):1255-1261. https://doi.org/10.1002/elps.202000349

Huang HZ, Liu XZ, Cheng JB, et al., 2022. A novel multiplex assay system based on 10 methylation markers for forensic identification of body fluids. *J Forensic Sci*, 67(1):136-148. https://doi.org/10.1111/1556-4029.14872

Huang SJ, Cai NG, Pacheco PP, et al., 2018. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics*, 15(1):41-51. https://doi.org/10.21873/cgp.20063

Kader F, Ghai M, Olaniran AO, 2020. Characterization of DNA methylation-based markers for human body fluid identification in forensics: a critical review. *Int J Legal Med*, 134(1):1-20. https://doi.org/10.1007/s00414-019-02181-3

Lee HY, An JH, Jung SE, et al., 2015. Genome-wide methylation profiling and a multiplex construction for the identification of body fluids using epigenetic markers. *Forensic Sci Int Genet*, 17:17-24. https://doi.org/10.1016/j.fsigen.2015.03.002

Lee JE, Lee JM, Naue J, et al., 2022. A collaborative exercise on DNA methylation-based age prediction and body fluid typing. *Forensic Sci Int Genet*, 57:102656. https://doi.org/10.1016/j.fsigen.2021.102656

Martin NC, Clayson NJ, Scrimger DG, 2006. The sensitivity and specificity of Red-Starch paper for the detection of saliva. *Sci Justice*, 46(2):97-105. https://doi.org/10.1016/s1355-0306(06)71580-5

Mattei AL, Bailly N, Meissner A, 2022. DNA methylation: a historical perspective. *Trends Genet*, 38(7):676-707. https://doi.org/10.1016/j.tig.2022.03.010

Pan C, Yi SH, Xiao C, et al., 2020. The evaluation of seven age-related CpGs for forensic purpose in blood from Chinese Han population. *Forensic Sci Int Genet*, 46:102251. https://doi.org/10.1016/j.fsigen.2020.102251

Park JL, Kwon OH, Kim JH, et al., 2014. Identification of body fluid-specific DNA methylation markers for use in forensic science. *Forensic Sci Int Genet*, 13:147-153. https://doi.org/10.1016/j.fsigen.2014.07.011

Sijen T, Harbison S, 2021. On the identification of body fluids and tissues: a crucial link in the investigation and solution of crime. *Genes*, 12(11):1728. https://doi.org/10.3390/genes12111728

Tian H, Bai P, Tan Y, et al., 2020. A new method to detect methylation profiles for forensic body fluid identification combining ARMS-PCR technique and random forest model. *Forensic Sci Int Genet*, 49:102371. https://doi.org/10.1016/j.fsigen.2020.102371

Virkler K, Lednev IK, 2009. Analysis of body fluids for forensic purposes: from laboratory testing to non-destructive rapid confirmatory identification at a crime scene. *Forensic Sci Int*, 188(1-3):1-17. https://doi.org/10.1016/j.forsciint.2009.02.013

Zhao C, Yang J, Xu H, et al., 2022. Genetic diversity analysis of forty-three insertion/deletion loci for forensic individual identification in Han Chinese from Beijing based on a novel panel. *J Zhejiang Univ-Sci B (Biomed & Biotechnol)*, 23(3):241-248. https://doi.org/10.1631/jzus.B2100507

Supplementary information

Figs. S1–S5