The last imprint: Differential methylation assay for forensic identification of brain tissue and death determination

Supplementary file

The score for each genomic position is calculated using the following Equation (S1):

$$\text{Score} = \sum_{1}^{n}[\sum_{1}^{k} b*\log(|d - Mt[n][k]| + p) + \sum_{1}^{w} \log(|(1-d) - Mb[n][k]| + p)] \quad \text{(S1)}$$

where $Mt_1 \ldots Mt_k$ refers to n vectors in length k, describing the methylation state in k non-brain tissue samples in n genomic positions; $Mb_1 \ldots Mb_w$ refers to n vectors in length k, describing the methylation state in w brain tissue samples in n genomic positions; p is the penalty parameter; b is the "brain preference" parameter, representing the relative importance of methylation values in brain samples; and d is the ideal methylation state for each tissue, for each score type. The value of d is defined based on the sample type and the corresponding score classification. For brain samples, the hypo score was assigned a value of 0, while the hyper score was assigned 1. In contrast, for non-brain samples, the hypo score was assigned 1, and the hyper score was assigned 0.

Potential genomic regions containing at least five CpG sites within 300 bp were ranked based on CpG count and the mean informativeness score of the top 50% most informative sites using Equation (S2):

$$\text{Score} = \frac{\min(amplicon\_len,\ \max\_rewarding\_size)}{\min(\max\_\text{len},\ \max\_\text{rewarding}\_\text{size})} * |\max\_score| * length\_weight + \max\_bias * \max\_score + \frac{top\_half\_sum}{amplicon\_len} \quad \text{(S2)}$$

The parameters and their corresponding final values for Equation (S2) are defined as follows. The maximum length (max_len) was set to 300. A value of 1 was used to represent the bias toward the best position within the amplicon range (max_bias). The maximum amplicon size rewarded (max_rewarding_size) was set at 150. The bias toward amplicon length was given a weight of 0.0 (length_weight), while 0.5 was assigned to represent the fraction of top positions given higher importance (top_half_fraction).

For information regarding potential genomic position ranges, the sum of the best top_half_fraction positions was recorded as Top_half_sum. The score of the best position within the range was referred to as top_score, while the distance between the first and last CpG sites was defined as amplicon_len. In addition, the maximum and minimum scores within the range were denoted as max_score and min_score, respectively.