ISSN 1230-7483
e-ISSN 2720-5983

*Problems of Forensic Sciences 2024, vol. 137, 17–29*

DOI: 10.4467/12307483PFS.24.002.19858

# DNA TESTING FOR INVESTIGATIVE PURPOSES: DESCRIPTION OF THE PERPETRATOR

Wojciech BRANICKI

*Institute of Zoology and Biomedical Research, Jagiellonian University, Kraków, Poland*
*Institute of Forensic Research, Kraków, Poland*

## Abstract

The results of research projects to understand the diversity of the human genome have opened up new avenues of biomedical research and provided new tools for human identification studies. Genome wide association studies and epigenome wide association studies have enabled the identification of DNA markers that have been implemented and validated as predictive tools in the field of forensic DNA phenotyping. In the age of genomics, the study of biological traces can reveal the biogeographical ancestry, physical appearance, age and lifestyle of the perpetrator. The combination of different methods, including forensic genetic genealogy and prediction of phenotypic features, offers the possibility of significantly narrowing down the pool of suspects, thereby significantly improving the process of solving criminal cases. In general, a limitation is the availability of effective methods for large-scale DNA analysis that would ensure the forensic level sensitivity of the test.

## Keywords

DNA markers; Forensic genomics; Predictive DNA analysis in forensics; DNA intelligence.

*Received 30 January 2024; accepted 25 March 2024*

## 1. Information about biogeographical ancestry encoded in the human genome

The evolution of modern humans has involved their expansion and adaptation to extremely diverse environmental conditions and to the exchange of genes with other human species. This has led to the phenotype and genotype diversity we observe among today's human populations. This diversity is highly valuable in terms of directing forensic investigations and identifying culprits. A better understanding of the human genome variation has led to the establishment of the concept of the 'biological witness', according to which biological traces in the form of DNA isolated from a biological sample discovered at a crime scene may provide reliable information about the criminal that can be used to complement or replace witness testimonies.

Inferring the biogeographical ancestry (BGA) by analysing DNA markers narrows down the suspects considerably, and is a particularly useful forensic tool, especially in ethnically diverse populations. Typical markers containing information about BGA include single-nucleotide polymorphism (SNP), insertion deletion polymorphism (INDEL) and the haplotypes of chromosome Y and mitochondrial DNA that are specific to the population inhabiting a given geographical region [1, 2]. Due to the mode of inheritance, haploid markers have major limitations with respect to the inferrence of BGA. They can be used to reveal a suspect's biogeographical ancestry on the mother's or father's linage and are less genetically diverse. In

contrast, markers of autosomal DNA are inherited from both parents and provide information about the structure of a population, which can be utilised to detect admixtures of DNA from different populations and migration events.

Basic research on population biodiversity conducted by anthropologists and geneticists has yielded very reliable data for the development of BGA tests. Particularly important in this respect have been the HapMap and the 1000 genomes projects [3, 4], as a result of which many markers and methods for the inferrence of BGA for forensic purposes were developed [5].

Snipper, a dedicated app (http://mathgene.usc.es/snipper/), provides algorithms for the interpretation of data from 34 SNP and 46 INDEL loci and ensures the resolution of BGA inference at the level of Europe, Africa, East Asia, America and Oceania. Genetic data needed to predict BGA can be collected using traditional methods (minisequencing followed by capillary electrophoresis), or more advanced ones incorporating high-throughput sequencing [6]. Snipper has different display options for the results of a BGA analysis, including the options to use a Bayesian calculator, multinomial regression or genetic distance. Complementary data for biogeographical inference can be obtained from the FROG-kb database (https://frog.med.yale.edu/FrogKB/) [7, 8], which allows users to calculate the relative probabilities of biogeographical ancestry based on profiles determined with various SNP marker tests. Recently, the VISAGE consortium has proposed a method that incorporates combined data that includes autosomal SNP (including 21 microhaplotypes), ChrY and ChX. The benefits of this extended panel of 226 markers are an increased inference resolution that includes regions in the Middle East and South Asia, improved analysis of mixed BGA, and inference in case of some DNA mixtures [9].

Forensic DNA laboratories also employ commercial solutions, which implement the DNA variants selected by forensic geneticists tested using high-throughput sequencing, such as ForenSeq DNA Signature Prep Kit or Precision ID Ancestry Panel. An interesting solution is the PhenoTrivium which, in addition to 163 autosomal markers for the analysis of BGA, also contains 120 markers located on the Y chromosome [10]. Complementary to the list of markers for the inferrence of biogeographical ancestry are the 41 pigmentation markers included in the HIrisPlex-S algorithm, which will be discussed in more detail in Section 2.

Lastly, genomic predictors of BGA are available that provide a very good global resolution. An example is the Geographic Population Structure (GPS) algorithm. The GPS classifier draws biogeographical information from 40,000–130,000 SNP polymorphisms [11]. From a technical standpoint, analysing such a vast number of markers in forensic DNA laboratories is problematic, due to the necessity of using a DNA microarray requiring large amounts of DNA. However, research is underway on genomic methods that will also meet the forensic standards. Determining the biogeographical ancestry helps in the development of a culprit's genetic portray and provides indirect information about his or her physical phenotype. Describing an individual's physical traits in more detail requires analysis of genes and genetic variants that directly affect the appearance, many of which have already been identified.

## 2. Prediction of human physical traits

Research on differences in phenotypes of monozygotic twins has shown that genes are important contributors to physical traits. Many studies have confirmed the high heritability of the human physical phenotype (65–95%), as well as the major importance of additive gene effects in the shaping thereof [12]. However, the genetic prediction of physical traits is very difficult due to the polygenic nature of physical traits. Research indicates that even eye colour, which was initially classified among the traits with a Mendelian inheritance, is determined by dozens of genes [13]. A particularly complex physical trait is height which, according to the literature, is affected by genetic variants located in over 7,000 segments of the human genome [14]. Genes that control the human phenotype are identified through genome-wide association studies (GWAS). Detecting rare variants of DNA that contribute to a person's physical traits often requires the assessment of many hundreds of thousands of people with diverse phenotypes. In order to develop predictive algorithms with which the phenotype of an unknown person can be determined within a particular degree of probability based on a DNA sample, it is necessary to use additional datasets of a sufficient size.

There are many reports concerning algorithms that enable the prediction of pigmentation traits. Whereas a majority of these algorithms were developed for determining eye colour (the least genetically complicated trait), some can also predict a person's hair colour or skin colour. A unique tool in this regard is the aforementioned HIrisPlex-S. It combines two elements necessary for an analysis: a genetic test to collect the data; and an algorithm to interpret that data. With respect to data collection, HIrisPlex-S can be

used in both the traditional version (minisequencing) and in a version based on high-throughput sequencing. In turn, the statistical interpretation in HIrisPlex-S utilises three separate predictive algorithms [15, 16]. The simplest algorithm for the prediction of eye colour uses only 6 polymorphic SNP positions across 6 genes, including *HERC2* and *OCA2*. The complex of these two genes is the most significant contributor to human eye colour, with variant C in the *HERC2* determining blue eyes [17]. There is a long list of publications that present predictive methods for the pigmentation phenotype, especially the colour of the iris. However, only the HIrisPlex-S algorithms have been commercialised [18, 5]. Predicting the colour of one's eyes, hair and skin with HIrisPlex-S only requires 41 SNP variants, which indicates that the genetic architecture of pigmentation traits is relatively simple. Nonetheless, it should be noted that the assessment of physical traits through genetic tests is a probabilistic classification. The accuracy of the prediction depends on the phenotype category and involves a risk of wrong classification, even for eye colour. Research clearly indicates that the risk of a wrong classification in the case of a blue or brown eye colour is small. However, the risk increases for the intermediate iris colours. A practical measure of the accuracy of predictive models (genetic classifiers) for qualitative variables is the area under curve (AUC), which describes both the sensitivity and specificity of a model. An AUC of 1 denotes a 100% correct classification. A value of 0.5 means that explanatory variables (in the case of genetic classifiers, genetic variants) do not increase the probability of classifying a given sample to the correct phenotype category. In the case, the AUC is 0.91 for blue eyes and 0.93 for brown eyes. For the intermediate colours, the value drops to 0.72 [18]. Another important consideration is that each marker included in a predictive model affect the results differently, whereas a lack of some predictors prevents a classification altogether. In the case of eye colour, a critical marker is the position rs12913832 in the *HERC2* gene, without which a predictive analysis is impossible.

In the subject literature, predictive methods for other physical traits are also proposed, which are without a doubt useful for describing an unknown culprit, as they allow for the prediction of extreme height [19], hair shape [20] and progress of greying and balding [21, 22]. Although these traits can be used to differentiate persons, the genetic prediction methods for such phenotypes are complex (i.e. they require more markers to analyse) and are poorly validated, which makes them less popular in practice [18]. A greatly anticipated method among forensic experts is one that could predict a person's facial appearance. Such a method would make it possible to directly identify a culprit by analysing a biological sample collected at the crime scene. However, the research in this area is highly demanding, primarily due to the complexity of the facial phenotype. Nonetheless, the first methods have already been proposed, despite the difficulties. They take advantage of differences in the facial morphology depending on one's biogeographical ancestry and sex and, to a lesser extent, also utilise an analysis of the genes responsible for shaping the facial phenotype [23].

A key element of predictive DNA testing for forensic purposes is collecting the genetic data, i.e. performing a genetic analysis of the biological traces. In this case, predictive tests face the same difficulties as those related to a traditional human identification analysis based on STR markers. The difficulty with examining biological traces lies with the degradation and contamination of samples collected at the crime scene. Unfortunately, genome-wide sequencing – a seemingly universal solution for biological trace analyses – is impractical. Genome-wide sequencing has performed well in paleogenetic studies, in which the researchers have accepted the fact that most of the resulting sequence reads are derived from the DNA of the microorganisms accompanying human remains. Forensic researchers use various targeted genetic panels to analyse selected segments of the human DNA. Intense research is being conducted on the genetic prediction of the physical phenotype for forensic purposes. However, the use of genetic methods in this area continues to be the subject of scientific, legal and ethical debates. In some countries, such as the Netherlands, the use of predictive methods for physical traits is regulated by law.

## 3. Significance of information about age in forensic investigations

Establishing the age of a culprit may considerably speed up an investigation, both in criminal cases and in the identification of human remains. Forensic scientists have for many years sought a method to predict human age. A breakthrough took place after effective markers were found in the human methylome as part of the research on genomic variation. DNA methylation is a chemical modification that provides control over the expression of genes. Although methylation may occur in response to environmental factors, it is largely pre-programmed and related to the development of every organism. Of greatest importance for

genome function is the attachment of a methyl group to cytosines in CG dinucleotide sequences, which are observed especially in gene promoter regions, where they form so-called CpG islands. An increase in the methylation level in these regions downregulates the gene expression.

A study demonstrated that about 30% of methylation changes within the human genome correlate significantly with age [24]. This finding opened up a new direction in research involving the development of methods for estimating human age based on genome methylation levels. Epigenetic clocks are mathematical algorithms, which are used to infer a person's age from information obtained about methylation levels in selected CpG sites. Consequently, age predictors were created by applying machine learning methods to sets of data about DNA methylation levels. The evolution of this line of research has led to the development of new generations of clocks, calibrated based on the participants' health information, including significant biochemical parameters of the blood [25]. As a result, second- and third-generation clocks are now able to determine a person's rate of ageing. With these methods, researchers can assess a person's biological age, physical condition, health or even approximate time of death [26, 27, 28, 29].

The methods of epigenetic age estimation have been developed based on sets of data on DNA methylation, collected using the specific DNA microarrays. However, an analysis of microarrays requires large amounts of high-quality DNA which is known to be difficult to achieve in forensics. The most popular epigenetic clocks use information about DNA methylation in 10 or even 1030 cytosines, located across various segments of the genome [26, 30, 29]. The earliest clocks were calibrated based on the participants' calendar age, which makes them the most suitable for estimating an individual's chronological age, and consequently, for use in forensics.

The study of biological traces containing small amounts of DNA has required the development of specific methods for measuring DNA methylation. Although these methods are highly sensitive, they can only analyse small number of CpG sites. As a result, the markers must be carefully selected in order to correlate well with the age. One of the first methods of this type was proposed by a Polish research consortium [31]. This predictive algorithm utilises data about the methylation levels of 5 CpG sites in the *ELOVL2*, *MIR29B2C*, *TRIM59*, *KLF14* and *FHL2* genes. The method underwent an effective validation through numerous studies conducted among different populations, including non-European ones. Furthermore, a set of 5 markers from the study by Zbieć-Piekarska et al. (2015) became the foundation for the development of advanced algorithms that allow for the estimation of age based on blood, saliva and bones [31, 32]. In the original study by Zbieć-Piekarska et al. (2015), the methylation data was collected using pyrosequencing. Later studies also used other methods, including high-throughput sequencing [31, 33].

In cases of sexual offences, semen analysis is often used and age prediction in such samples can be important at the investigation stage. However, research has shown that the DNA methylation pattern of sperm is completely different from that of somatic cells [34]. Pioneering research on the development of an age analysis method involving semen was conducted by a team from South Korea, who identified three CpG sites that significantly correlated with age (cg06304190 in the *TTC7B* gene, cg06979108 in the *NOX4* gene and cg12837463) [35]. The VISAGE research consortium has also conducted studies on this subject [36].

Collecting DNA methylation data is very difficult because the analysis is quantitative in nature, and the conversion stage of the non-methylated cytosines performed by bisulfite treatment has a degrading effect on the DNA. Moreover, the PCR reaction of the converted DNA is much more difficult to design and optimise. As a result, methods of DNA methylation analysis require a large amount of template DNA, while the simultaneous amplification of multiple loci presents its own significant challenge. An interesting solution that employs the AmpliSeq method for a simultaneous analysis of 161 CpG sites involved in four epigenetic age clocks was proposed in a recent study [37].

It is worth noting that determining age is exceedingly useful for the interpretation of predictions of age-related traits, such as hair greying or balding. Without information about a person's age, the genetic data may even lead to false interpretations of the phenotypes, and the reports provided to investigative authorities could be misleading. Therefore, it seems that the DNA analysis methods used at the investigation stage have practical value as long as they are applied jointly, which in some countries is restricted by the legal regulations [38].

A separate topic that has appeared in the literature is the application of epigenetic methods for age estimation in migration cases, which involve determining the age of persons who do not possess any documents containing such information [39]. The difficulty in using epigenetic methods for this purpose is due to their limited accuracy due to the variation in the rate of ageing of the human population, the error in measuring DNA methylation due to imperfect methods, and at

the same time the need to achieve high assay accuracy with virtually no margin of error, which is acceptable in studies of a predictive nature.

Methylation age is estimated increasingly frequently in forensic DNA laboratories, due to its high level of practicality and accuracy (an error of 3–5 years). Furthermore, the development of DNA sequencing methods is allowing for further improvements in the current methods of age estimation.

## 4. Lifestyle information contributes to the profile of a culprit

Age estimation is only one of many practical uses of DNA methylation analysis in forensics. The DNA methylation pattern is largely heritable and constitutes part of the ageing processes. However, a person's ageing rate is also affected by the environment, and DNA methylation itself is considered to be a bridge connecting genes and environmental factors. Consequently, various external factors may modify the methylation levels in different CpG sites. Research indicates that the analysis of DNA methylation in the human genome may reveal traces of the effect of the environment, which in turn allows for a selection of the epigenetic markers of various environmental factors [40]. Sites within the genome that have been affected by a given environmental factor can be identified using epigenome-wide association studies (EWAS), which are similar to GWAS. In particular, EWAS can be used to find particular methylation changes caused by specific environmental factors. The subjects are unrelated individuals with varying degrees of exposure to specific environmental factors. The method has provided very promising results in the case of the effects of smoking and alcohol use.

Smoking is known to have an exceptionally strong influence on DNA methylation levels. Many CpG sites have been identified that allow a tested person to be classified as a smoker or non-smoker. Furthermore, samples from current smokers, past smokers and never smokers have been classified correctly. These models use the DNA methylation data contained in several or more CpG markers [41, 42, 43, 44]. A predictor designed by Maas et al. uses information about the DNA methylation levels across 13 CpG sites and is highly accurate with an AUC of 0.901. The algorithm can also detect former smokers and never smokers [44]. The best-validated smoking marker is the *AHRR* gene. According to a study by Maas et al. (2019), a single cytosine in the *AHRR* gene enables the prediction with an AUC of 0.88. The product of the gene takes part in the degradation of environmental toxins.

Alcohol use is another factor that strongly affects the human methylome. A model for the epigenetic prediction of high alcohol consumption, developed using 144 CpG sites, showed a very satisfactory accuracy with an AUC > 0.9. The predictor was developed based on an analysis of the blood methylome of over 9,000 persons of European origin. A total of 363 cytosines correlating with alcohol consumption were identified, and predictive models were proposed based on 5, 23, 78 and 144 markers [44]. A study conducted on an independent sample confirmed the effectiveness of this model, although a slightly lower accuracy was obtained (AUC = 0.78). Interestingly, good results (AUC = 0.83) were obtained with alternative predictive models based on fewer markers (5 and 23 CpG sites), which were proposed in a study by Liu et al. (2018) and Maas et al. (2021) [45, 46]. It is worth noting that an analysis involving few markers is preferable in the case of forensic investigations.

Future models based on analyses of DNA methylation can be expected to prove helpful in revealing a culprit's other traits and habits. Basic research is underway aimed at identifying CpG sites that indicate correlations between the methylation levels and the use of psychoactive substances, culinary preferences, physical activity and exposure to stress.

Of particular interest is the potential of methylome analysis to predict body weight. Body weight is an important part of the description of an unknown culprit and is a vital complement to research on the prediction of a person's physical appearance. A study demonstrated that a model based on 397 CpG sites explained as much as 32% of variation in the BMI, which allowed for predictions of the BMI at an acceptable level [47]. Errors in the classification occurred in persons with characteristic biochemical blood profiles. It means that the identified CpG sites are markers of the observed BMI, as well as markers of the body mass indicator on a molecular level. Consequently, because a high BMI predisposes a person to developing many diseases, this predictor may prove to have clinical significance. Information about a person's physique is certainly useful when describing a culprit; however, it can also have a medical aspect, which forensic science has been avoiding. Nonetheless, such limitations may disappear in future, as medical data continues to be collected on a large scale in databases, which may in some cases greatly speed up identifying a culprit.

## 5. Summary

DNA isolated from biological traces is a valuable source of information for forensic investigations. Contemporary forensic genetics involves not only the identification of persons by comparing DNA profiles, but also directing investigations by revealing relevant information about unknown culprits. The application of a DNA analysis at the investigative stage in forensics is limited by the technical obstacles and gaps in fundamental knowledge about the practical importance of variation in the human genome. However, research on the genetic architecture of traits with a high potential for identification, such as facial features, is progressing. High-throughput sequencing technology is still difficult to use in forensic DNA laboratories. Furthermore, preparing DNA libraries and the sequencing process itself take up more time than the traditional methods of DNA variation analysis and standard genetic analysers. The necessity to store large amounts of data and to perform often uneasy bioinformatics analyses, followed by interpreting the results obtained for hundreds or thousands of DNA markers, means that not every laboratory will opt to implement new technologies if they only prove useful in a few criminal cases. One should also keep in mind the possibility of analysing non-human DNA for identification purposes. Recent studies have shown that analysing the environmental microbiome may provide information about the geographical origin of a sample [48]. In light of the above considerations, it can be concluded that forensic genetics is still a developing field, which continues to provide new research tools for the human identification.

## References


1. Halder I, Shriver M, Thomas M, Fernandez JR, Frudakis T. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum Mutat. 2008;29(5):648-58. doi: 10.1002/humu.20695.
2. Bardan F, Higgins D, Austin JJ. A custom hybridisation enrichment forensic intelligence panel to infer biogeographic ancestry, hair and eye colour, and Y chromosome lineage. Forensic Sci Int Genet. 2023;63:102822. doi: 10.1016/j.fsigen.2022.102822.
3. International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437(7063):1299-320. doi: 10.1038/nature04226.
4. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73. doi: 10.1038/nature09534.
5. Kayser M, Branicki W, Parson W, Phillips C. Recent advances in forensic DNA phenotyping of appearance, ancestry and age. Forensic Sci Int Genet. 2023;65:102870. doi: 10.1016/j.fsigen.2023.102870.
6. Eduardoff M, Gross TE, Santos C, de la Puente M, Ballard D, Strobl C, et al. Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM™. Forensic Sci Int Genet. 2016;23:178-189. doi: 10.1016/j.fsigen.2016.04.008.
7. Kidd KK, Soundararajan U, Rajeevan H, Pakstis AJ, Moore KN, Ropero-Miller JD. The redesigned forensic research/reference on genetics-knowledge base, FROG-kb. Forensic Sci Int Genet. 2018;33:33-37. doi: 10.1016/j.fsigen.2017.11.009.
8. Rajeevan H, Soundararajan U, Pakstis AJ, Kidd KK. FrogAncestryCalc: a standalone batch likelihood computation tool for ancestry inference panels catalogued in FROG-kb. Forensic Sci Int Genet. 2020;46:102237. doi: 10.1016/j.fsigen.2020.102237.
9. Ruiz-Ramírez J, de la Puente M, Xavier C, Ambroa-Conde A, Álvarez-Dios J, Freire-Aradas A, et al. Development and evaluations of the ancestry informative markers of the VISAGE enhanced tool for appearance and ancestry. Forensic Sci Int Genet. 2023;64:102853. doi: 10.1016/j.fsigen.2023.102853.
10. Diepenbroek M, Bayer B, Schwender K, Schiller R, Lim J, Lagacé R, Anslinger K. Evaluation of the Ion AmpliSeq™ PhenoTrivium panel: MPS-based assay for ancestry and phenotype predictions challenged by casework samples. Genes (Basel). 2020;11(12):1398. doi: 10.3390/genes11121398.
11. Elhaik E, Tatarinova T, Chebotarev D, Piras IS, Maria Calò C, De Montis A, et al. Geographic population structure analysis of worldwide human populations infers their biogeographical origins. Nat Commun. 2014 Apr 29;5:3513. doi: 10.1038/ncomms4513.
12. Polderman TJ, Benyamin B, de Leeuw CA, Sullivan PF, van Bochoven A, Visscher PM, et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nat Genet. 2015;47(7):702-9. doi: 10.1038/ng.3285.
13. Simcoe M, Valdes A, Liu F, Furlotte NA, Evans DM, Hemani G, et al. Genome-wide association study in almost 195,000 individuals identifies 50 previously unidentified genetic loci for eye color. Sci Adv. 2021;7(11):eabd1239. doi: 10.1126/sciadv.abd1239.
14. Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, et al. A saturated map of common genetic variants associated with human height. Nature. 2022;610(7933):704-712. doi: 10.1038/s41586-022-05275-y.

15. Chaitanya L, Breslin K, Zuñiga S, Wirken L, Pośpiech E, Kukla-Bartoszek M, et al. The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. Forensic Sci Int Genet. 2018;35:123-135. doi: 10.1016/j.fsigen.2018.04.004.
16. Breslin K, Wills B, Ralf A, Ventayol Garcia M, Kukla-Bartoszek M, Pospiech E, et al. HIrisPlex-S system for eye, hair, and skin color prediction from DNA: massively parallel sequencing solutions for two common forensically used platforms. Forensic Sci Int Genet. 2019;43:102152. doi: 10.1016/j.fsigen.2019.102152.
17. Sturm RA, Duffy DL, Zhao ZZ, Leite FP, Stark MS, Hayward NK, et al. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. Am J Hum Genet. 2008;82(2):424-31. doi: 10.1016/j.ajhg.2007.11.005.
18. Pośpiech E, Teisseyre P, Mielniczuk J, Branicki W. Predicting physical appearance from DNA data-towards genomic solutions. Genes (Basel). 2022;13(1):121. doi: 10.3390/genes13010121.
19. Liu F, Zhong K, Jing X, Uitterlinden AG, Hendriks AEJ, Drop SLS, Kayser M. Update on the predictability of tall stature from DNA markers in Europeans. Forensic Sci Int Genet. 2019;42:8-13. doi: 10.1016/j.fsigen.2019.05.006.
20. Pośpiech E, Chen Y, Kukla-Bartoszek M, Breslin K, Aliferi A, Andersen JD, et al. Towards broadening forensic DNA phenotyping beyond pigmentation: improving the prediction of head hair shape from DNA. Forensic Sci Int Genet. 2018;37:241-251. doi: 10.1016/j.fsigen.2018.08.017.
21. Pośpiech E, Kukla-Bartoszek M, Karłowska-Pik J, Zieliński P, Woźniak A, Boroń M, et al. Exploring the possibility of predicting human head hair greying from DNA using whole-exome and targeted NGS data. BMC Genomics. 2020;21(1):538. doi: 10.1186/s12864-020-06926-y.
22. Chen Y, Hysi P, Maj C, Heilmann-Heimbach S, Spector TD, Liu F, Kayser M. Genetic prediction of male pattern baldness based on large independent datasets. Eur J Hum Genet. 2023;31(3):321-328. doi: 10.1038/s41431-022-01201-y.
23. Sero D, Zaidi A, Li J, White JD, Zarzar TBG, Marazita ML, et al. Facial recognition from DNA using face-to-DNA classifiers. Nat Commun. 2019;10(1):2557. doi: 10.1038/s41467-019-10617-y.
24. Jones MJ, Goodman SJ, Kobor MS. DNA methylation and healthy human aging. Aging Cell. 2015;14(6):924-32. doi: 10.1111/acel.12349.
25. Noroozi R, Ghafouri-Fard S, Pisarek A, Rudnicka J, Spólnicka M, Branicki W, et al. DNA methylation-based age clocks: from age prediction to age reversion. Ageing Res Rev. 2021;68:101314. doi: 10.1016/j.arr.2021.101314.
26. Zhang Y, Wilson R, Heiss J, Breitling LP, Saum KU, Schöttker B, et al. DNA methylation signatures in peripheral blood strongly predict all-cause mortality. Nat Commun. 2017;8:14617. doi: 10.1038/ncomms14617.
27. Belsky DW, Caspi A, Arseneault L, Baccarelli A, Corcoran DL, Gao X, et al. Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm. Elife. 2020;9:e54870. doi: 10.7554/eLife.54870.
28. McGreevy KM, Radak Z, Torma F, Jokai M, Lu AT, Belsky DW, et al. DNAmFitAge: biological age indicator incorporating physical fitness. Aging (Albany NY). 2023;15(10):3904-3938. doi: 10.18632/aging.204538.
29. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging (Albany NY). 2019;11(2):303-327. doi: 10.18632/aging.101684.
30. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14(10):R115. doi: 10.1186/gb-2013-14-10-r115.
31. Zbieć-Piekarska R, Spólnicka M, Kupiec T, Parys-Proszek A, Makowska Ż, Pałeczka A, et al. Development of a forensically useful age prediction method based on DNA methylation analysis. Forensic Sci Int Genet. 2015 Jul;17:173-179. doi: 10.1016/j.fsigen.2015.05.001.
32. Woźniak A, Heidegger A, Piniewska-Róg D, Pośpiech E, Xavier C, Pisarek A, et al. Development of the VISAGE enhanced tool and statistical models for epigenetic age estimation in blood, buccal cells and bones. Aging (Albany NY). 2021;13(5):6459-6484. doi: 10.18632/aging.202783.
33. Hong SR, Shin KJ, Jung SE, Lee EH, Lee HY. Platform-independent models for age prediction using DNA methylation data. Forensic Sci Int Genet. 2019;38:39-47. doi: 10.1016/j.fsigen.2018.10.005.
34. Jenkins TG, Aston KI, Pflueger C, Cairns BR, Carrell DT. Age-associated sperm DNA methylation alterations: possible implications in offspring disease susceptibility. PLoS Genet. 2014;10(7):e1004458. doi: 10.1371/journal.pgen.1004458.
35. Lee HY, Jung SE, Oh YN, Choi A, Yang WI, Shin KJ. Epigenetic age signatures in the forensically relevant body fluid of semen: a preliminary study. Forensic Sci Int Genet. 2015;19:28-34. doi: 10.1016/j.fsigen.2015.05.014.
36. Pisarek A, Pośpiech E, Heidegger A, Xavier C, Papież A, Piniewska-Róg D, et al. Epigenetic age prediction in semen – marker selection and model development. Aging (Albany NY). 2021;13(15):19145-19164. doi: 10.18632/aging.203399.
37. Pośpiech E, Pisarek A, Rudnicka J, Noroozi R, Boroń M, Masny A, et al. Introduction of a multiplex amplicon sequencing assay to quantify DNA methylation in target cytosine markers underlying four selected epigenetic clocks. Clin Epigenetics. 2023;15(1):128. doi: 10.1186/s13148-023-01545-2.

38. Zieger M. Forensic DNA phenotyping in Europe: how far may it go? J Law Biosci. 2022 Sep 14;9(2):lsac024. doi: 10.1093/jlb/lsac024.
39. Ritz-Timme S, Schneider PM, Mahlke NS, Koop BE, Eickhoff SB. Age estimation based on DNA methylation. Ready for use to establish the chronological age of young migrants without valid identity documents? Rechtsmedizin. 2018;28(3):202-208.
40. Vidaki A, Kayser M. From forensic epigenetics to forensic epigenomics: broadening DNA investigative intelligence. Genome Biol. 2017;18(1):238. doi: 10.1186/s13059-017-1373-1.
41. Philibert RA, Beach SR, Lei MK, Brody GH. Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking. Clin Epigenetics. 2013;5(1):19. doi: 10.1186/1868-7083-5-19.
42. McCartney DL, Hillary RF, Stevenson AJ, Ritchie SJ, Walker RM, Zhang Q, et al. Epigenetic prediction of complex traits and death. Genome Biol. 2018;19(1):136. doi: 10.1186/s13059-018-1514-1.
43. Sugden K, Hannon EJ, Arseneault L, Belsky DW, Broadbent JM, Corcoran DL, et al. Establishing a generalized polyepigenetic biomarker for tobacco smoking. Transl Psychiatry. 2019;9(1):92. doi: 10.1038/s41398-019-0430-9.
44. Maas SCE, Vidaki A, Wilson R, Teumer A, Liu F, van Meurs JBJ, et al. Validated inference of smoking habits from blood with a finite DNA methylation marker set. Eur J Epidemiol. 2019;34(11):1055-1074. doi: 10.1007/s10654-019-00555-w.
45. Liu C, Marioni RE, Hedman ÅK, Pfeiffer L, Tsai PC, Reynolds LM, et al. A DNA methylation biomarker of alcohol consumption. Mol Psychiatry. 2018;23(2):422-433. doi: 10.1038/mp.2016.192.
46. Maas SCE, Vidaki A, Teumer A, Costeira R, Wilson R, van Dongen J, et al. Validating biomarkers and models for epigenetic inference of alcohol consumption from blood. Clin Epigenetics. 2021;13(1):198. doi: 10.1186/s13148-021-01186-3.
47. Do WL, Sun D, Meeks K, Dugué PA, Demerath E, Guan W, et al. Epigenome-wide meta-analysis of BMI in nine cohorts: Examining the utility of epigenetically predicted BMI. Am J Hum Genet. 2023;110(2):273-283. doi: 10.1016/j.ajhg.2022.12.014.
48. Danko D, Bezdan D, Afshin EE, Ahsanuddin S, Bhattacharya C, Butler DJ, et al. A global metagenomic map of urban microbiomes and antimicrobial resistance. Cell. 2021;184(13):3376-3393.e17. doi: 10.1016/j.cell.2021.05.002.

**ORCID**
Wojciech Branicki 0000-0002-7412-5733

**Corresponding author**
Prof. Wojciech Branicki
Instytut Ekspertyz Sądowych
ul. Westerplatte 9
PL 31-033 Kraków
e-mail: wbranicki@ies.gov.pl

# BADANIA DNA DLA CELÓW DOCHODZENIOWO-ŚLEDCZYCH – OPIS SPRAWCY PRZESTĘPSTWA

## 1. Informacja o pochodzeniu biogeograficznym zapisana w genomie człowieka

Ewolucja człowieka współczesnego wiązała się z jego ekspansją i dostosowaniem do skrajnie różnych warunków środowiskowych, a także wymianą genów z innymi napotkanymi gatunkami ludzi. Doprowadziło to do zróżnicowania na poziomie fenotypu i genotypu, które obserwujemy we współcześnie żyjących populacjach ludzkich. Zróżnicowanie to ma dużą wartość w ukierunkowaniu śledztwa i procesie identyfikacji przestępców. Lepsze poznanie zmienności genomu człowieka doprowadziło do powstania koncepcji świadka biologicznego, która zakłada, że ślad biologiczny, a dokładnie DNA wyizolowany z próbki biologicznej ujawnionej na miejscu zdarzenia kryminalnego, może stanowić źródło cennych informacji o przestępcy, umożliwiających uzupełnienie lub zastąpienie zeznań świadków przestępstwa.

Określenie pochodzenia biogeograficznego poprzez analizę markerów DNA w opisie sprawcy przestępstwa istotnie zawęża grono osób podejrzanych i jest szczególnie przydatnym narzędziem dochodzeniowo-śledczym, zwłaszcza w populacjach wieloetnicznych. Markery zawierające informację o biogeografii to zazwyczaj polimorfizm pojedynczo nukleotydowy SNP, ale również polimorfizm insercyjno-delecyjny INDEL, a także haplotypy chromosomu Y i DNA mitochondrialnego specyficzne dla populacji zamieszkujących różne regiony geograficzne [1, 2]. W związku ze sposobem dziedziczenia markery haploidalne mają istotne ograniczenia w zastosowaniu do badania pochodzenia biogeograficznego. Pozwalają na ujawnienie pochodzenia biogeograficznego w linii matki lub ojca, a także charakteryzują się mniejszym zróżnicowaniem genetycznym. Przeciwnie, dziedziczone po obojgu rodzicach markery autosomalnego DNA dostarczają informacji na temat struktury populacji, co pozwala na wykrywanie domieszek DNA pochodzącego z różnych populacji, a także zdarzeń migracji.

Badania podstawowe prowadzone przez antropologów i genetyków nad bioróżnorodnością populacji ludzkich dostarczyły bardzo dobrych danych do opracowania testów na pochodzenie biogeograficzne. Jak wspomniano wcześniej, szczególnie duże znaczenie miały badania przeprowadzone w ramach projektów HapMap, a następnie 1000 genomów [3, 4]. W konsekwencji zaproponowano wiele markerów i metod umożliwiających analizę pochodzenia biogeograficznego dla celów sądowych [5].

Dedykowana dla tego celu aplikacja Snipper (http://mathgene.usc.es/snipper/) udostępnia algorytmy pozwalające na interpretację danych z 34 SNP oraz 46 loci INDEL i zapewnia rozdzielczość na poziomie Europa – Afryka – Azja Wschodnia – Ameryka – Oceania. Dane genetyczne potrzebne do przeprowadzenia predykcji pochodzenia biogeograficznego można zebrać za pomocą metod klasycznych (minisekwencjonowanie połączone z elektroforezą kapilarną) lub bardziej zaawansowanych opartych na technologii sekwencjonowania wysokoprzepustowego [6]. Snipper umożliwia przedstawienie wyniku analizy pochodzenia biogeograficznego na różne sposoby, w tym zastosowanie kalkulatora bayesowskiego, regresji wielomianowej oraz dystansu genetycznego. Komplementarnych danych w zakresie wnioskowania nt. biogeografii dostarcza baza FROG-kb (https://frog.med.yale.edu/FrogKB/) [7, 8]. Umożliwia ona obliczenie względnych prawdopodobieństw pochodzenia biogeograficznego na podstawie profili oznaczonych za pomocą różnych paneli markerów SNP. Niedawno konsorcjum VISAGE zaproponowało metodę, która korzysta z połączonych danych SNP autosomalnych (w tym 21 mikrohaplotypów), ChrY i ChrX. Ten poszerzony panel 226 markerów pozwala na zwiększenie rozdzielczości predykcji o regiony Środkowego Wschodu oraz Azji Południowej, skuteczniejszą analizę mieszanego pochodzenia biogeograficznego, a także na wnioskowanie w przypadku niektórych mieszanin DNA [9].

W kryminalistycznych laboratoriach DNA stosowane są także komercyjne rozwiązania, które polegają na zastosowaniu zestawów do analizy wyselekcjonowanych przez genetyków sądowych wariantów DNA metodą sekwencjonowania wysokoprzepustowego, jak ForenSeq DNA Signature Prep Kit czy Precision ID Ancestry Panel. Ciekawym rozwiązaniem jest panel PhenoTrivium, który poza 163 markerami autosomalnymi umożliwiającymi analizę pochodzenia biogeograficznego zawiera dodatkowo 120 markerów zlokalizowanych na chromosomie Y [10]. Listę markerów pozwalających na ustalenie pochodzenia biogeograficznego uzupełnia 41 markerów pigmentacyjnych wchodzących w skład algorytmu HIrisPlex-S, o którym więcej informacji pojawi się w rozdziale 2.

Znane są wreszcie genomiczne predyktory pochodzenia biogeograficznego zapewniające bardzo dobrą rozdzielczość w skali globalnej, do których należy algorytm GPS (*geographic population structure*). Klasyfikator GPS czerpie informację o biogeografii z 40 000–130 000 polimorfizmów SNP [11]. Technicznie analiza tak dużej liczby markerów w kryminalistycznych laboratoriach DNA jest problematyczna ze względu na konieczność zastosowania metody mikromacierzy DNA wymagającej dużych ilości DNA, ale prowadzone są prace nad

metodami genomicznymi, które jednocześnie zaspokajałyby wymagania kryminalistyki. Określenie pochodzenia biogeograficznego jest pomocne w opracowaniu portretu genetycznego sprawcy przestępstwa, pośrednio informując o jego ogólnym fenotypie fizycznym. Do dokładnego opisu cech fizycznych konieczna jest analiza genów i wariantów genetycznych, z których wiele zostało już zidentyfikowanych, a które bezpośrednio wpływają na wygląd.

## 2. Predykcja cech wyglądu człowieka

Badania nad zróżnicowaniem fenotypu bliźniąt monozygotycznych wykazały, że geny mają istotne znaczenie w kształtowaniu cech wyglądu. Wysoki poziom odziedziczalności fenotypu fizycznego człowieka (na poziomie 65–95%) potwierdzono w licznych badaniach naukowych. Jednocześnie wykazano duże znaczenie addytywnych efektów genów w kształtowaniu fenotypu fizycznego [12]. Genetyczna predykcja cech fizycznych sprawia jednak duże trudności ze względu na ich poligeniczność. Z przeprowadzonych badań wynika, że nawet kolor oczu, który początkowo genetycy klasyfikowali w grupie cech o dziedziczeniu mendlowskim, jest determinowany przez kilkadziesiąt genów [13]. Cechą fizyczną o wyjątkowej złożoności jest wzrost człowieka, który – jak wynika z przeprowadzonych badań – pozostaje pod wpływem wariantów genetycznych zlokalizowanych w ponad siedmiu tysiącach segmentów genomu człowieka [14]. Identyfikację genów kontrolujących fenotyp człowieka prowadzi się poprzez pełnogenomowe badania asocjacyjne (GWAS). Wykrycie rzadkich wariantów DNA istotnych w determinacji cech wymaga często zbadania wielu setek tysięcy osób o zróżnicowanym fenotypie. Do opracowania algorytmów predykcyjnych, dzięki którym po zbadaniu próbki DNA można z określonym prawdopodobieństwem ustalić fenotyp fizyczny nieznanej osoby, konieczne jest zastosowanie dodatkowych zbiorów danych o odpowiedniej liczebności.

W literaturze można odnaleźć wiele raportów na temat algorytmów, które umożliwiają predykcję cech pigmentacyjnych. Najwięcej z nich opracowano dla koloru oczu (najmniej skomplikowanej pod względem genetycznym cechy), ale są też dostępne metody do predykcji koloru włosów i skóry. Wyjątkowym narzędziem jest wspomniana wcześniej metoda HIrisPlex-S. Jest ona złożona z dwóch elementów niezbędnych do przeprowadzenia całego procesu analizy, tj. testu genetycznego umożliwiającego zebranie danych oraz algorytmu pozwalającego na ich interpretację. Na etapie zbierania danych genetycznych metoda może być stosowana zarówno w wersji klasycznej (minisekwencjonowanie), jak również w wersji opartej na zastosowaniu sekwencjonowania wysokoprzepustowego, a na etapie statystycznej interpretacji danych korzysta z trzech odrębnych algorytmów predykcyjnych [15, 16]. Najprostszy algorytm do predykcji koloru oczu oparty jest na zaledwie 6 polimorficznych pozycjach SNP z 6 genów, w tym *HERC2* i *OCA2*. Kompleks tych dwóch genów jest najważniejszy w kształtowaniu koloru oczu, a wariant C w genie *HERC2* jest odpowiedzialny za niebieski kolor oczu u człowieka [17]. Lista publikacji, w których przedstawiono metody predykcji fenotypu pigmentacyjnego, zwłaszcza koloru tęczówki oka, jest długa, ale to algorytmy HIrisPlex-S zostały skomercjalizowane [18, 5]. Predykcja koloru oczu, włosów i skóry za pomocą metody HIrisPlex-S wymaga analizy zaledwie 41 wariantów SNP, co wskazuje na stosunkowo mało skomplikowaną architekturę genetyczną cech pigmentacyjnych. Należy pamiętać, że określanie cech wyglądu poprzez badania genetyczne to probabilistyczna klasyfikacja. Dokładność predykcji zależy od kategorii fenotypowej i wiąże się z ryzykiem błędnej klasyfikacji nawet w przypadku koloru oczu. Z badań jasno wynika, że ryzyko nieprawidłowej klasyfikacji niebieskiego koloru oczu jest niewielkie i podobnie rzecz się ma z kolorem brązowym. Jednak ryzyko to rośnie dla pośrednich kolorów tęczówki oka. Praktyczną miarą dokładności działania modeli predykcyjnych (klasyfikatorów genetycznych) dla zmiennych jakościowych jest parametr AUC (*area under curve*), który opisuje jednocześnie czułość i specyficzność modelu. Wartość AUC równa 1 oznacza 100% prawidłowych klasyfikacji, natomiast wartość 0,5 oznacza, że zmienne wyjaśniające (w przypadku klasyfikatorów genetycznych – warianty genetyczne) nie zwiększają prawdopodobieństwa zaklasyfikowania próbki do właściwej kategorii fenotypowej. Dla niebieskiego koloru oczu AUC = 0,91, dla brązowego AUC = 0,93, natomiast dla kategorii kolorów pośrednich wartość ta spada do AUC = 0,72 [18]. Istotne jest również to, że poszczególne markery włączone w model predykcyjny mają różny wpływ na wynik badania, a brak niektórych predyktorów uniemożliwia przeprowadzenie klasyfikacji. W przypadku koloru oczu krytycznym markerem jest właśnie pozycja rs12913832 w genie *HERC2* – bez niej analiza predykcyjna jest niemożliwa do przeprowadzenia.

W literaturze można znaleźć również wiele doniesień na temat metod predykcji innych cech wyglądu fizycznego. Z pewnością są one przydatne w opisie nieznanego sprawcy przestępstwa, gdyż umożliwiają predykcję ekstremalnie wysokiego wzrostu [19], kształtu włosów [20], stopnia zaawansowania siwienia i łysienia [21, 22]. Są to cechy przydatne w różnicowaniu ludzi, ale metody predykcji genetycznej tych fenotypów są bardziej złożone (np. wymagają analizy większej ilości markerów), słabiej zwalidowane i przez to rzadziej stosowane w praktyce [18]. Duże oczekiwania wiąże się z predykcją wyglądu twarzy w kontekście kryminalistycznym. Taka metoda mogłaby umożliwić bezpośrednią identyfikację

sprawcy przestępstwa po zbadaniu próbki biologicznej ujawnionej na miejscu zdarzenia. Prace badawcze w tym obszarze są jednak bardzo wymagające, głównie ze względu na złożoność fenotypu twarzy. Mimo trudności zaproponowano już pierwsze metody predykcji wyglądu twarzy. Polegają one w dużej mierze na wykorzystaniu różnic w jej morfologii zależnych od pochodzenia biogeograficznego, a także płci, a w mniejszym stopniu na badaniu genów odpowiedzialnych za kształtowanie fenotypu twarzy [23].

Kluczowym elementem predykcyjnego badania DNA dla potrzeb sądowych jest zebranie danych genetycznych, a więc przeprowadzenie analizy genetycznej śladu biologicznego. Testy predykcyjne muszą w tym wypadku zmierzyć się z tymi samymi problemami, które wiążą się z klasyczną analizą identyfikacyjną opartą na zastosowaniu markerów STR. Trudność badania śladów biologicznych wiąże się z problemami degradacji i kontaminacji próbek zabezpieczanych na miejscu zdarzenia kryminalnego. Niestety mało praktyczne jest sekwencjonowanie pełnogenomowe, które wydawać by się mogło uniwersalnym rozwiązaniem analizy śladów biologicznych. Zastosowanie tej metody dobrze sprawdziło się w badaniach paleogenetycznych, gdzie pogodzono się z faktem, że większość odczytów sekwencji z przeprowadzonej analizy pochodzi z DNA mikroorganizmów towarzyszących szczątkom ludzkim. W badaniach sądowych stosowane są różnego rodzaju celowane panele genetyczne, które umożliwiają analizę wybranych segmentów ludzkiego DNA. Genetyczna predykcja fenotypu fizycznego w kryminalistyce jest obszarem badawczym, w którym wciąż prowadzone są intensywne badania. Zastosowanie metod genetycznej predykcji cech fenotypowych w praktyce wciąż jest przedmiotem dyskusji na gruncie naukowym, prawnym i etycznym. W niektórych krajach, jak na przykład w Holandii, stosowanie metod predykcji cech wyglądu zostało uregulowane prawnie.

## 3. Znaczenie informacji o wieku człowieka na etapie śledztwa

Odpowiedź na pytanie, w jakim wieku był sprawca przestępstwa, może istotnie przyspieszyć śledztwo. Określenie wieku może być pomocne zarówno w śledztwach prowadzonych w sprawach kryminalnych, jak i w sprawach identyfikacji szczątków ludzkich. W naukach sądowych od wielu lat poszukiwano skutecznych metod predykcji wieku człowieka. Przełom nastąpił dzięki badaniom podstawowym nad zmiennością genomu, a skuteczne markery odnaleziono w metylomie człowieka. Metylacja DNA to chemiczna modyfikacja, która umożliwia kontrolę ekspresji genów u człowieka. Może ona stanowić odpowiedź na czynniki środowiskowe, ale w dużej mierze jest zaprogramowana i wiąże się z naturalnym rozwojem każdego organizmu. Największe znaczenie dla funkcjonowania genomu ma przyłączanie grupy metylowej do cytozyn w sekwencjach dinukleotydowych CG, które są obserwowane zwłaszcza w regionach promotorów genów, gdzie tworzą tzw. wyspy CpG. Podniesienie poziomu metylacji w tych regionach prowadzi do obniżenia ekspresji genu.

Z przeprowadzonych badań wynika, że ok. 30% zmian metylacji w genomie człowieka zachodzi w sposób istotnie skorelowany z wiekiem [24]. Wiedza ta umożliwiła rozwój nowego kierunku badawczego polegającego na konstruowaniu metod umożliwiających estymację wieku człowieka na podstawie poziomu zmetylowania genomu. Zegary epigenetyczne to algorytmy matematyczne, które umożliwiają zamianę informacji o stopniu metylacji w odpowiednio wyselekcjonowanych miejscach CpG na wiek badanej osoby. Predyktory wieku są wobec tego efektem zastosowania metod uczenia maszynowego do zbiorów danych o poziomie metylacji DNA. Ewolucja tego kierunku badań doprowadziła do opracowania kolejnych generacji zegarów, które zostały skalibrowane z uwzględnieniem danych o stanie zdrowia uczestników, na przykład istotnych parametrów biochemicznych krwi [25]. W rezultacie zegary drugiej i trzeciej generacji umożliwiają określenie tempa starzenia badanej osoby. Zastosowanie takich metod pozwala na estymację wieku biologicznego, kondycji fizycznej, stanu zdrowia, a wręcz uzyskania informacji o przybliżonym czasie zgonu [26, 27, 28, 29].

Metody szacowania wieku epigenetycznego opracowano w oparciu o zbiory danych o metylacji DNA zebrane przy użyciu specyficznych mikromacierzy DNA. Jednakże analiza mikromacierzy wymaga dużych ilości wysokiej jakości DNA, co jest trudne do uzyskania w kryminalistyce. Najpopularniejsze zegary epigenetyczne wykorzystują informację o metylacji DNA w 10, a nawet 1030 cytozynach, zlokalizowanych w różnych segmentach genomu [26, 30, 29]. Najwcześniejsze zegary kalibrowano na podstawie wieku kalendarzowego uczestników, co czyni je najbardziej odpowiednimi do szacowania wieku chronologicznego jednostki, a co za tym idzie, do stosowania w kryminalistyce.

Badanie śladów biologicznych, które zawierają niewielkie ilości DNA, wymusiło opracowanie specjalnych metod pomiaru metylacji DNA. Metody te zapewniają wysoką czułość badania, ale za ich pomocą możliwa jest analiza mniejszej ilości miejsc CpG – dlatego markery te muszą być skrupulatnie wyselekcjonowane, aby wykazywać wysoką korelację z wiekiem. Jedną z pierwszych tego typu metod zaproponowało polskie konsorcjum badawcze [31]. Opracowany algorytm predykcyjny korzysta z informacji o stopniu zmetylowania 5 miejsc CpG zlokalizowanych w genach *ELOVL2*, *MIR29B2C*, *TRIM59*, *KLF14* i *FHL2*. Metodę dobrze zwalidowano dzięki licznym badaniom przeprowadzonym w różnych

populacjach, także pozaeuropejskich. Zestaw 5 markerów z pracy Zbieć-Piekarskiej i in. (2015) stał się podstawą do opracowania bardziej zaawansowanych algorytmów umożliwiających estymację wieku w krwi, ślinie oraz kościach [31, 32]. W oryginalnej pracy Zbieć-Piekarskiej i in. (2015) dane metylacyjne zbierano metodą pirosekwencjonowania, ale w późniejszych pracach zastosowano też inne metody, w tym sekwencjonowanie wysokoprzepustowe [31, 33].

Sprawy przestępstw na tle seksualnym wiążą się z koniecznością analizy nasienia, a predykcja wieku w takich próbkach może mieć istotne znaczenie na etapie dochodzeniowo-śledczym. Badania wykazały jednak, że wzór metylacji DNA w plemnikach jest całkowicie odmienny od tego, który znamy dla komórek somatycznych [34]. Pionierskie prace nad opracowaniem metody analizy wieku w nasieniu przeprowadził zespół z Korei Południowej, który zidentyfikował 3 miejsca CpG o istotnej korelacji z wiekiem (cg06304190 w genie *TTC7B*, cg06979108 w genie *NOX4* oraz cg12837463) [35]. Badania prowadzone były również przez konsorcjum badawcze VISAGE [36].

Zbieranie danych metylacyjnych sprawia niemałe trudności w związku z tym, że analiza jest ilościowa, a etap konwersji niemetylowanych cytozyn za pomocą wodorosiarczynu ma działanie degradujące na DNA. Dodatkowo reakcja PCR skonwertowanego DNA jest dużo trudniejsza do zaprojektowania i optymalizacji. Dlatego metody analizy metylacji DNA wymagają większej ilości matrycy, a jednoczesna amplifikacja wielu loci stanowi duże wyzwanie. Ciekawe rozwiązanie, polegające na zastosowaniu metody AmpliSeq do jednoczesnej analizy 161 miejsc CpG istotnych dla działania 4 epigenetycznych zegarów wieku, zaproponowano w jednej z niedawnych prac [37].

Warto zauważyć, że określenie wieku jest niezwykle przydatne w interpretacji wyniku analizy predykcyjnej cech związanych z wiekiem, np. siwienia czy łysienia. Sama informacja genetyczna pozbawiona wiedzy o wieku może wręcz prowadzić do fałszywych interpretacji tych fenotypów, a wskazówki przekazane organom dochodzeniowo-śledczym mogą wprowadzić w błąd. Dlatego wydaje się, że metody analizy DNA stosowane na etapie dochodzeniowo-śledczym mają praktyczne zastosowanie, gdy stosowane są łącznie, co jest ograniczane w niektórych krajach za pomocą regulacji prawnych [38].

Odrębnym tematem, który pojawił się w literaturze, jest zastosowanie epigenetycznych metod estymacji wieku w przypadku spraw migracyjnych, w których konieczne jest określenie wieku osób nieposiadających dokumentów zawierających taką informację [39]. Trudność wdrożenia metod epigenetycznych w tym celu wynika z ich ograniczonej dokładności spowodowanej zróżnicowaniem tempa starzenia w populacji ludzi, błędem pomiaru metylacji DNA wynikającym z niedoskonałości metod, a przy tym konieczności osiągnięcia wysokiej dokładności badania praktycznie pozbawionej marginesu błędu, który jest dopuszczalny w badaniach o charakterze predykcyjnym.

Estymacja wieku metylacyjnego jest coraz częściej stosowana w kryminalistycznych laboratoriach DNA, co wynika z dużej praktycznej użyteczności, a przy tym dokładności badania z błędem ok. 3–5 lat. Co więcej, rozwój metod sekwencjonowania DNA umożliwia dalsze udoskonalanie wdrożonych metod estymacji wieku.

## 4. Informacja o stylu życia przybliża wizerunek sprawcy przestępstwa

Estymacja wieku to jedynie jedno z praktycznych zastosowań badania metylacji DNA dla celów kryminalistycznych. Wzór metylacji DNA jest w dużej mierze odziedziczalny i stanowi element procesów starzenia się człowieka. Na tempo starzenia ma jednak również wpływ środowisko, a sama metylacja DNA uważana jest za pomost pomiędzy genami a czynnikami środowiskowymi. Dlatego różne czynniki zewnętrzne mogą wpływać na zmianę poziomu metylacji w różnych miejscach CpG. Z przeprowadzonych badań wynika, że analiza metylacji DNA w genomie człowieka umożliwia odnalezienie śladów działania środowiska. To z kolei pozwala na selekcję epigenetycznych markerów różnych czynników środowiskowych [40]. Identyfikację miejsc w genomie naznaczonych działaniem określonego czynnika środowiskowego prowadzi się za pomocą tzw. analizy asocjacyjnej pełnego epigenomu (EWAS). Badanie przypomina analizę GWAS. Badane są osoby niespokrewnione, które są w różnym stopniu narażone na oddziaływanie konkretnych czynników środowiskowych. Metoda umożliwiła odnalezienie specyficznych zmian metylacyjnych u osób narażonych na działanie różnych czynników. Badania przyniosły szczególnie obiecujące rezultaty w przypadku badania wpływu palenia papierosów i spożywania alkoholu.

Palenie papierosów wiąże się z wyjątkowo silnym wpływem na poziom metylacji DNA. Zidentyfikowano wiele miejsc CpG, które pozwalają na klasyfikację badanej osoby w grupie palaczy lub osób niepalących. Co więcej, analiza pozwala na prawidłową klasyfikację próbek pochodzących od osób aktualnie palących, palących w przeszłości i niepalących. Modele korzystają z danych o metylacji DNA zawartych w kilku do kilkunastu markerach CpG [41, 42, 43, 44]. Predyktor opracowany przez Maasa i wsp. korzysta z informacji o poziomach metylacji DNA w 13 pozycjach CpG i umożliwia predykcję z dokładnością wyrażoną poprzez AUC na poziomie 0,901. Algorytm pozwala również na wykrycie byłych palaczy i niepalących przez całe życie [44]. Najlepiej zwalidowanym markerem palenia jest gen *AHRR*.

Z badań Maasa i in. (2019) wynika, że pojedyncza cytozyna w tym genie umożliwia predykcję na poziomie AUC = 0,88. Produkt genu *AHRR* ma udział w degradacji toksyn środowiskowych.

Silny jest także wpływ spożycia alkoholu na metylom człowieka. Model do epigenetycznej predykcji wysokiej konsumpcji alkoholu zbudowany z zastosowaniem 144 miejsc CpG wykazał dokładność na bardzo satysfakcjonującym poziomie AUC > 0,9. Predyktor opracowano z uwzględnieniem wyników analizy metylomu krwi dla ponad 9 tysięcy osób o pochodzeniu europejskim. W pracy zidentyfikowano 363 cytozyny skorelowane z konsumpcją alkoholu i zaproponowano modele predykcyjne oparte na 5, 23, 78 i 144 markerach [44]. Badania przeprowadzone na niezależnej próbie potwierdziły działanie modelu, choć zanotowano nieco mniejszą dokładność (AUC = 0,78). Co ciekawe, dobre wyniki (AUC = 0,83) uzyskano dla alternatywnych modeli predykcyjnych opartych na mniejszej liczbie markerów (5 i 23 CpG) zaproponowanych w pracy Liu i in (2018) oraz Maasa i in. (2021) [45, 46]. Warto zauważyć, że analiza mniejszej liczby markerów jest korzystniejsza dla badań prowadzonych w kryminalistyce.

Można założyć, że modele oparte na wynikach analizy metylacji DNA będą w przyszłości pomocne w ujawnianiu innych cech i przyzwyczajeń sprawcy przestępstwa. Badania podstawowe prowadzone są nad identyfikacją miejsc CpG wykazujących korelację poziomu metylacji ze stosowaniem substancji psychotropowych, upodobaniami kulinarnymi czy też stopniem aktywności fizycznej i poziomem narażenia na stres.

Szczególnie interesujący jest potencjał analizy metylomu w kierunku predykcji masy ciała. Cecha ta ma duże znaczenie przy opisie nieznanego sprawcy przestępstwa i jest istotnym uzupełnieniem badań nad predykcją wyglądu fizycznego. W jednej z prac pokazano, że model oparty na 397 miejscach CpG wyjaśnia aż 32% zmienności wskaźnika masy ciała (BMI), co umożliwia predykcję BMI na akceptowanym poziomie [47]. Błędy w klasyfikacji stwierdzono u osób o charakterystycznych profilach biochemicznych krwi. Oznacza to, że zidentyfikowane CpG są markerami BMI, który podlega obserwacji, oraz wskaźnika masy ciała na poziomie molekularnym. W związku z tym, że wysoki BMI stanowi ryzyko rozwoju wielu chorób, wydaje się, że zastosowanie takiego predyktora może mieć istotne znaczenie kliniczne. Informacje na temat budowy ciała z pewnością są użyteczne w opisie sprawcy przestępstwa, ale – jak widać – mogą mieć również charakter medyczny, którego kryminalistyka stara się przynajmniej na razie unikać. Ograniczenia tego typu mogą jednak zniknąć w przyszłości, gdyż dane o charakterze medycznym są gromadzone na dużą skalę w bazach danych, a bez wątpienia w niektórych okolicznościach mogą skutecznie przyspieszyć wykrycie sprawcy przestępstwa.

## 5. Podsumowanie

DNA wyizolowany ze śladu biologicznego jest istotnym źródłem informacji przydatnych na etapie prowadzonego śledztwa. Współczesna genetyka sądowa to nie tylko identyfikacja poprzez porównanie profili DNA, ale również ukierunkowanie śledztwa poprzez ujawnienie istotnych informacji o nieznanym sprawcy przestępstwa. Zastosowanie analizy DNA na etapie dochodzeniowo-śledczym ograniczają problemy techniczne i wciąż istniejące luki w wiedzy podstawowej na temat praktycznego znaczenia zróżnicowania genomu człowieka. Wciąż jednak rośnie wiedza podstawowa nad architekturą genetyczną cech o dużym potencjale identyfikacyjnym, takich jak wygląd twarzy. Technologia sekwencjonowania wysokoprzepustowego jest wciąż niełatwa do zastosowania w kryminalistycznym laboratorium DNA. Przygotowanie bibliotek DNA i sam proces sekwencjonowania zabierają więcej czasu w porównaniu z zastosowaniem klasycznych metod analizy zmienności DNA i analizatorów genetycznych. Konieczność przechowywania dużych ilości danych i często niełatwej analizy bioinformatycznej, a następnie interpretacji wyników uzyskanych dla setek czy tysięcy markerów DNA sprawia, że nie każde laboratorium zdecyduje się na wdrożenie nowych technologii, które znajdą zastosowanie w nielicznych sprawach kryminalnych. Należy też pamiętać o możliwości analizy DNA niepochodzącego od człowieka dla celów wykrywczych. Niedawne prace pokazały, że badanie mikrobiomu środowiskowego może dostarczyć informacji o pochodzeniu geograficznym próbki badawczej [48]. W świetle powyższych rozważań można stwierdzić, że genetyka sądowa wciąż się rozwija, dostarczając nowych narzędzi badawczych, które ułatwiają identyfikację człowieka.