genes

MDPI

Review

Predicting Physical Appearance from DNA Data—Towards Genomic Solutions

Ewelina Pośpiech [1], Paweł Teisseyre [2,3], Jan Mielniczuk [2,3] and Wojciech Branicki [1,4,*]

[1] Malopolska Centre of Biotechnology, Jagiellonian University, 30-387 Kraków, Poland; ewelina.pospiech@uj.edu.pl
[2] Institute of Computer Science, Polish Academy of Sciences, 01-248 Warsaw, Poland; Pawel.Teisseyre@ipipan.waw.pl (P.T.); jan.mielniczuk@ipipan.waw.pl (J.M.)
[3] Faculty of Mathematics and Information Science, Warsaw University of Technology, 00-662 Warsaw, Poland
[4] Central Forensic Laboratory of the Police, 00-583 Warsaw, Poland
* Correspondence: wojciech.branicki@uj.edu.pl; Tel.: +48-126-645-024

Abstract: The idea of forensic DNA intelligence is to extract from genomic data any information that can help guide the investigation. The clues to the externally visible phenotype are of particular practical importance. The high heritability of the physical phenotype suggests that genetic data can be easily predicted, but this has only become possible with less polygenic traits. The forensic community has developed DNA-based predictive tools by employing a limited number of the most important markers analysed with targeted massive parallel sequencing. The complexity of the genetics of many other appearance phenotypes requires big data coupled with sophisticated machine learning methods to develop accurate genomic predictors. A significant challenge in developing universal genomic predictive methods will be the collection of sufficiently large data sets. These should be created using whole-genome sequencing technology to enable the identification of rare DNA variants implicated in phenotype determination. It is worth noting that the correctness of the forensic sketch generated from the DNA data depends on the inclusion of an age factor. This, however, can be predicted by analysing epigenetic data. An important limitation preventing whole-genome approaches from being commonly used in forensics is the slow progress in the development and implementation of high-throughput, low DNA input sequencing technologies. The example of palaeoanthropology suggests that such methods may possibly be developed in forensics.

Keywords: physical appearance; human genome variation; DNA-based prediction; investigative leads; forensic DNA intelligence; forensic genomics

Citation: Pośpiech, E.; Teisseyre, P.; Mielniczuk, J.; Branicki, W. Predicting Physical Appearance from DNA Data—Towards Genomic Solutions. *Genes* **2022**, *13*, 121. https://doi.org/10.3390/genes13010121

Academic Editor: Fulvio Cruciani

Received: 10 December 2021
Accepted: 4 January 2022
Published: 10 January 2022

1. Introduction

The information included in genomic data can be used to generate investigative leads that, when properly used, can speed up the process of human identification in forensic investigations. Such forensic DNA intelligence can use a variety of methods, including relatedness testing, the inference of ancestry, the prediction of physical phenotype, and age estimation [1–4]. As an inherently interdisciplinary science, forensic science today can benefit from the rapidly developing methods in the areas of genomics and machine learning, which is particularly beneficial for the further development of forensic DNA intelligence. Studies of human genome variation conducted today on an unprecedented scale are revealing how genes control phenotypes. This knowledge has fundamental meaning for understanding the genome–phenome relationship. Importantly, the growing knowledge of human genome variation allows for the development of algorithms that can more accurately predict phenotypes, providing more reliable investigative leads to help identify an unnamed perpetrator or victim and solve a case. It is worth noting that the DNA-based predictive tools developed in the forensic field are also useful in evolutionary anthropology. In this review paper, we will summarise how the advances in understanding

the genetic architectures of various human physical characteristics, and the progress in high-throughput genotyping technologies in combination with machine-learning methods, allow the prediction of physical appearance traits. We will also highlight the evolution of the approach to the genetic prediction of physical traits, which has moved from building predictive models based on variables that show genetic association to building models based on variables that improve predictive performance (Figure 1).

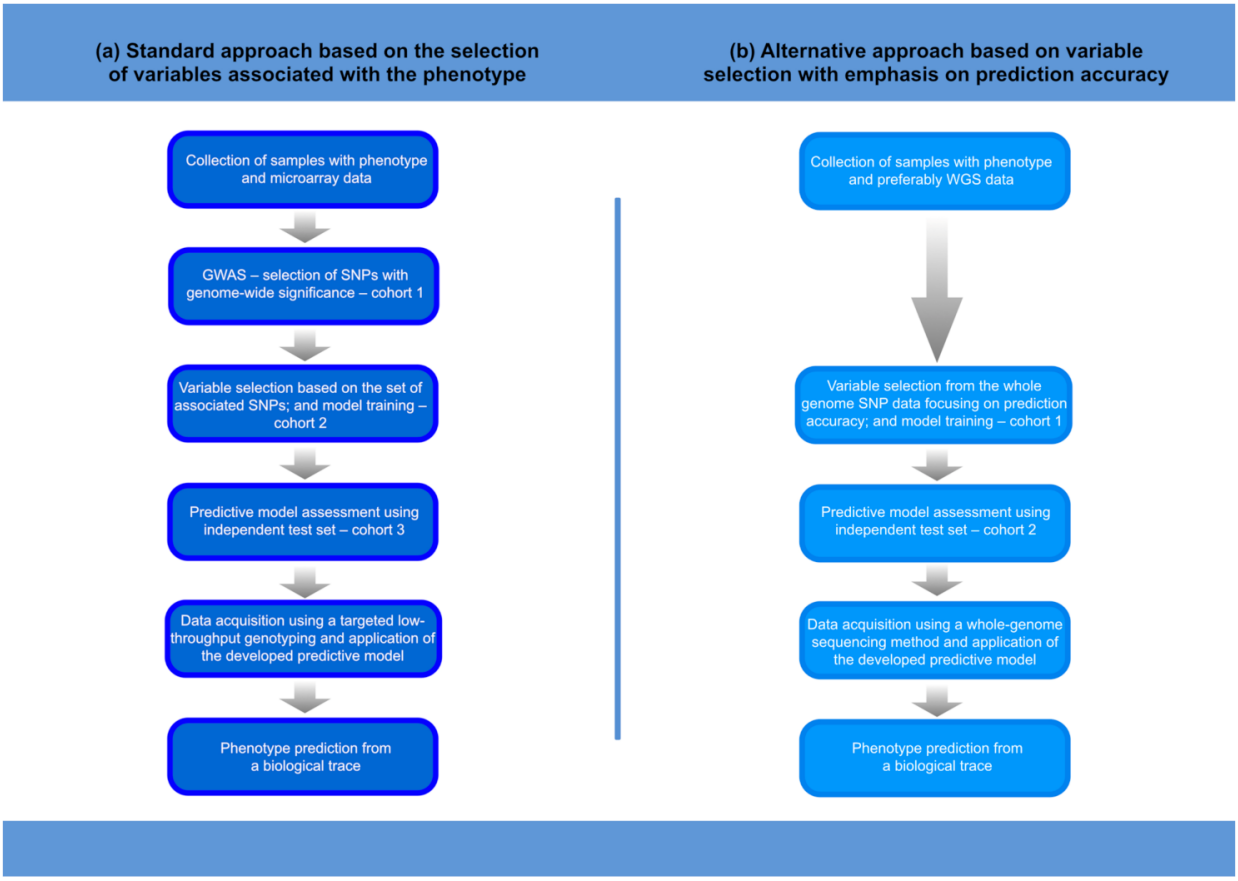

Figure 1. Procedure for the development and application of a phenotype prediction tool. The main differences in the procedures for developing a predictive model using the standard or alternative approach concern the selection of variables and the number of variables in the model. Consequently, the method of acquiring genetic data in the practical forensic applications of the next-generation predictive models may require whole-genome sequencing methods. Thus: (**a**) only phenotype-associated SNPs are included in prediction modelling, the models are not very extensive, and the methods of data acquisition can be less complex (SNaPshot, targeted MPS); (**b**) the selection of relevant variables (SNPs) is targeted towards improving the prediction accuracy of the model, and much more advanced variable selection methods are required. Some complex models may involve many thousands of SNPs, which, in biological traces, must be analysed using whole-genome sequencing methods that are effective for low DNA input samples.

2. Explaining the Heritability of Appearance Traits

A large meta-analysis of twin studies has confirmed that all human traits are heritable and showed that most of the traits can be explained by an additive genetic variation [5]. The extreme similarity of physical appearance of monozygotic twins clearly indicates the role of genes, but their identification is not simple due to the complex nature of appearance traits. Linkage mapping, which relies on the co-segregation of causal DNA variants with marker variants (SNP or STR) within pedigrees, has been very successful at identifying the gene variants affecting simple mendelian traits [6], but mostly failed to identify the DNA variants involved in the determination of complex traits [7]. The breakthrough in explaining the

heritability of complex phenotypes has come with the advent of genome-wide association studies (GWAS), which are effective at discovering common variants with small effect sizes on traits [7]. GWAS is used to identify associations between genotypes and phenotypes by testing for differences in the allele frequency of DNA variants between individuals who differ phenotypically. Technically, the analysis of hundreds of thousands of DNA variants in the genomes of these individuals enables finding those statistically associated with a specific phenotype [8].

2.1. Pigmentation Phenotype

GWAS data have been very effective at explaining the heritability of physical appearance traits. The heritability of human pigmentation traits has been assessed to be above 80% and, thus, provides a good starting point for DNA-based prediction, because it means that 80% of the variation in pigmentation in a population is due to genetic variation between individuals and that the influence of the environment is relatively small [9–11].

Many candidate genes for human pigmentation were identified before the GWAS era through animal models and the linkage to diseases with mendelian inheritance modes, such as oculocutaneous albinism. Genome-wide association scans confirmed the importance of these genes and identified many of the novel gene variants influencing the variability of normal human pigmentation. The collected data confirmed a very promising perspective for the genetic prediction of pigmentation traits. The less complex nature of some pigmentation phenotypes, such as blue and brown eye colours and red hair colour, and the availability of DNA variants with relatively large effect sizes, similar to the genetic effects observed for mendelian traits, were particularly encouraging. The region on chromosome 15, including the *OCA2* gene, was implicated in eye colour via linkage and subsequent fine-mapping analyses [12,13]. The evidence of a relationship between *OCA2* genotypes and eye colour became stronger with additional reports [14–16]. This was an Icelandic GWAS that implicated the involvement of neighbouring *HERC2* in the determination of eye colour and suggested that this genomic region was responsible for the regulation of *OCA2* gene expression [17]. This speculation was soon confirmed by other studies that showed that the DNA variant rs12913832 was responsible for brown and blue eye colour in humans [18,19]. The postulated interaction between these two genes in determining eye colour was also confirmed [20]. Most of the SNP heritability of red hair colour is explained by the single *MC1R* gene, which was also discovered long before the genome-wide association scans and confirmed in various population samples across the globe [21,22]. The effect of this gene was extended to skin colour and freckling [23,24]. These early genome-wide association scans for pigmentation also clearly demonstrated their agnostic power to detect novel, sometimes unexpected genotype–phenotype relationships, as in the case of the *IRF4* gene, which is now an important predictor of pigmentation phenotypes [17,25]. The success of GWAS was clear, but a significant proportion of heritability remained missing, which could be attributed mainly to the insufficiently large sample sizes used in genome-wide association scans, insufficient phenotyping regimes generating heterogeneity, the insufficient density of the GWA arrays and the significance of non-additive variation [26,27]. Indeed, the improved statistical power to detect small effect-size variants more effectively in the next series of genome-wide association scans enabled the identification of multiple new DNA variants involved in the heritability of hair and eye colour. For example, a large study of 192,986 European individuals from 10 populations identified 50 new loci for eye colour [28]. The study revealed signals with genome-wide significance for 12,192 SNPs from 52 genomic regions in the discovery set of 157,485 individuals. By combining discovery and replication sets, the study finally identified 124 independent associations from 61 genomic regions and concluded that the known variants explain 53.2% of eye colour variation. Notably, the study also investigated Asian cohorts and found consistency in the genetic architecture of eye colour in populations from Europe and Asia [28].

Human skin colour is highly variable at continental and intercontinental levels, complicating research on the genetic architecture and heritability of this trait [29]. The rs1426654

in *SLC24A5*, discovered thanks to a Zebrafish study, plays an important role in skin colour differences at the continental level, explaining more than 30% of skin colour differences between African and European populations [30]. GWAS on skin colour conducted on various population samples discovered multiple genes and gene variants involved in skin colour variation at the intercontinental level [25,31–34]. Notably, the studies of African populations showed large differences in skin colour, revealing the high complexity of the genetic architecture of skin colour in Africa and the significance of genes unknown to European studies [35,36].

A meta-analysis that involved almost 300,000 genomes from individuals of European ancestry included in two different cohorts (23andMe, UK Biobank) discovered 124 loci relevant to human hair colour, mostly novel associations, including genes with strong effect, such as *SLC45A1*, *DSTYK*, *FOSL2*, *LHX2*, *EDNRB*, *SHC4*, *KRT31*, and *BCAS1*. The study was highly successful at explaining up to 34.6% (red hair) of the heritability of human hair colour, despite an imperfect phenotyping regime involving self-reported hair colour in adulthood [37]. Another study based on UK Biobank resources examined 343,234 genomes from participants reporting British descent and these were, thus, more homogeneous. This study assessed that all identified variants explain 90% of the SNP heritability of red hair colour but, surprisingly, it found that a DNA variant located 97 kb from the 5′ end of the *MC1R* gene may be more important for explaining red hair colour than the polymorphism within the *MC1R* exon, and it identified an additional eight loci that contribute to the genetics of red hair colour. This research also revealed 213 variants important to the determination of blond hair colour, accounting for 73% of SNP heritability. In addition, a set of 56 DNA variants was found to be important for brown hair colour and was assessed to account for 47% of SNP heritability of this hair category [38].

2.2. Hair Features

Along with hair colour, other features describing the properties of human hair can be useful to define the physical appearance of an individual. Research shows that genetics plays a key role in the determination of hair features. However, the level of heritability may differ for various hair traits. Very high heritability (85–95%) was estimated for hair shape [39]. Heritability of around 70% was reported for monobrow and beard thickness [40]. Studies are contradictory in terms of the heritability level of hair loss (~40–80%) [40–43] and hair greying (~30–90%) [40,44], but the accuracy of the heritability measurement may be affected by the definition of heritability, the study design and the method of analysis used [45]. It is worth noting that heritability values calculated from the entire SNPs analysed in GWA studies tend to be underestimated compared to estimates of pedigree heritability, because the former do not include phenotypic variation due to rare variants that are not correctly determined by the SNPs genotyped on microarrays or common variants with small effect sizes that are not correctly identified if the sample size is not large enough. In turn, pedigree heritability may be biased by the common environmental factors to which families are typically exposed [43,45]. Notably, heritability estimates may vary due to changes in allele frequencies in populations caused by different evolutionary mechanisms and environmental contributions that change with the age of individuals [46]. The genetic basis of hair loss and hair shape are the most investigated so far. Androgenetic alopecia, known in men as male pattern baldness (MPB), is the most common type of progressive loss of hair from the scalp and is particularly frequent among men in Europe. Over the last >10 years, several GWA studies on hair loss have been carried out, with the vast majority of research conducted on Europeans [43,47–53]. These studies revealed multiple genes that are associated with the risk of MPB, with two loci showing the strongest effect of association, Xq12 (*AR*/*EDA2R*) and 20p11 (*PAX1*/*FOXA2*). Of these two loci, only 20p11 is known to act in Asians, indicating the significance of population heterogeneity [54,55]. A long list of additional loci, representing smaller effect sizes, identified through GWAS and/or candidate gene approaches, is available in the literature (e.g., *HDAC9*, *WNT10*, *TARDBP*, *EBF1*, *SUCNR1*, *AUTS2*, *FGF5*, *IRF4*, *C1orf127*, *RUNX1*, and *TWIST2*). Studies published in

2017–2018 led to significant advances in research on the genetics of hair loss. Four large GWA scans have been conducted on individuals of European descent. The first three of those studies, which investigated 20,000–50,000 genomes each, detected altogether more than 300 GWA signals, including 253 novel MPB associations [42,52,53]. The largest study, which investigated 200,000 genomes, allowed the identification of >600 genome-wide associations, explaining altogether 25% of the phenotypic variation observed in alopecia [43]. These large-scale studies not only discovered many new loci involved in alopecia, but also highlighted the implicated molecular pathways and discovered the genetic links of alopecia with different traits/conditions, including bone mineral density, puberty, metabolic traits, and Parkinson's disease. However, a significant part of MPB heritability remains missing. Recent studies showed that the use of advanced statistical methods and the incorporation of functional genomics data prior to association tests may improve the efficiency of SNP detection in GWAS and these approaches were proved to be successful in MPB research by increasing the number of SNP hits by an additional ~4% [56,57].

For hair morphology (shape), four GWA studies have been published so far, two of which were carried out on Europeans, with one study on Latin Americans and one on East Asians [40,58–60]. Hair shape, usually defined as straight vs. wavy vs. curly, is a highly distinctive feature of human appearance. As with hair loss, genetic heterogeneity between populations is observed with different mechanisms and genes underlying straight hair formation in Europeans and East Asians. The *TCHH* gene is known to act only in Europeans, while *EDAR* is the main contributor to straight hair in East Asians [58,59,61]. However, the proportion of known heritability attributed to both genes in respective populations was found to be small (<10%). The *TCHH* gene was discovered in the first GWA study published in 2009, which was conducted on three cohorts with a total of more than 4800 individuals of European descent [58]. *TCHH* was the only gene in this study that reached genome-wide significance, but suggestive associations were also disclosed for several additional loci, including the *FRAS1* and *WNT10* genes. The role of the *EDAR* gene in hair straightness and thickness in Asians was discovered through candidate gene analyses [61,62] and confirmed in a later GWA study conducted in 2016 on ~2900 Chinese people, with no additional genes reaching GWA significance in this study [59]. A GWA study conducted on >6000 Latin Americans discovered a novel association for *PRSS53* [40]. The latest meta-analysis of European GWA studies, exploring a total of more than 16,000 samples, allowed the identification of 12 hair shape genes, including eight novel association signals (*ERRFI1/SLC45A1*, *PEX14*, *PADI3*, *TGFA*, *LGR4*, *HOXC13*, *KRTAP*, and *PTK6*) [60]. The study showed that a model consisting of 14 SNPs across novel and literature loci, together with sex, explains 10% of the total hair shape variability. Further research pointed to the role of gene–gene interactions in hair shape determination as one of the factors underlying missing heritability [63]. The *RPTN* gene has been implicated in straight hair formation in Europeans and East Asians, but throughout interactions with different previously known head hair shape genes.

Only a few studies have addressed the genetic basis of other hair traits. In recent years, the first genes responsible for the thickness of the eyebrows, e.g., *EDAR*, *FOXL2*, *LIMS1*, *TMEM174*, SOX2, and *FOXD1* [40,64,65], monobrows, e.g., *PAX3* and 5q13.2 [40,51], beard thickness, e.g., *EDAR*, *LNX1*, *PREP*, and *FOXP2* [40], and hair greying, e.g., *IRF4*, and *KIF1A* [40,66], have been identified through whole-genome or whole-exome analyses.

2.3. Human Height

Human height is heritable in approximately 80%, but only the single genes associated with this trait were known before the era of GWA studies. The study of human height genetics is a good example of the effectiveness of explaining the heritability of complex traits through the GWAS approach. An important advantage of stature studies is undoubtedly the ease of measuring this phenotype and thus the homogeneity of the phenotypic data. The genome-wide association scans for human stature identified gene variants with a small effect size, clearly confirming the high polygenicity of this trait. The first three GWA

studies of human height, which collectively included more than 50,000 samples, detected only 54 loci with a statistically significant association with stature [67–69]. Most of these loci have not been previously linked to human height and, in many cases, the known biological function did not make them candidates for the regulation of human stature. The genes discovered explained only about 5% of the variation in height, which was very discouraging, especially for predicting this distinctive feature. A huge meta-analysis that considered GWAS data for more than 180,000 genomes made only a small advance in explaining the heritability of human growth, enabling the discovery of 180 loci. The work demonstrated the importance of allelic heterogeneity in explaining the complex genetic architecture of human stature [70]. At the same time, it has been argued that testing each SNP individually for an association with a trait, which is typical for GWAS investigations, leads to missing many real associations, especially when the effect sizes of individual SNPs on a trait are small. By fitting all SNPs simultaneously, Yang et al. provided an unbiased estimate of the variance explained by the SNPs in total, and showed that common genetic variants are able to explain as much as 45% of the variance in human height [71]. However, consistently increasing the number of genomes analysed with high-density DNA microarrays has proven to be an effective method for elucidating the still-missing genetic variation responsible for human height. The large meta-analysis that included 700,000 European genomes (250,000 previously investigated [72] and 450,000 from the UK Biobank) identified 3290 near-independent SNPs associated with human stature which were found to explain 24.6% of variance of this trait [73]. Still, unpublished data suggest that the large proportion of missing heritability may be hiding in rare genetic variants ($\leq$0.01) that can be detected via the whole genome sequencing of a sufficient number of genomes [74]. Notably, Zoledziewska et al. showed that human height can be under pressure from natural selection, presenting data showing that known height-decreasing alleles were found at higher frequency in Sardinians than would be expected to be caused by genetic drift [75]. Research on the genetic architecture of human stature, on a smaller scale, is also being conducted in populations in Asia and Africa. A meta-analysis of 93,926 individuals from East Asia identified 98 loci, including 17 novel for human height [76]. A GWA study based on 191,787 Japanese genomes disclosed 573 height-associated variants and assessed that 64 rare (<0.01) and low-frequency (<0.05) variants explain 1.7% of the height variance. The study revealed genes not previously associated with stature [77]. Eighty-three low-frequency variants affecting human height have also been reported in [78].

2.4. Facial Morphology

The human face represents a set of correlated complex phenotypes that are highly variable at inter- and intra-population levels and define what is apparently the most differentiating human trait [79]. The high similarity of the faces of monozygotic twins clearly indicates that most of this variability is genetically determined. Despite this, research into the heritability of facial features has caused quite a few problems, probably due to the three-dimensional nature of human faces. Only a recent face heritability study performed on 952 British twins using an advanced phenotyping and landmarking system confirmed the high heritability (>65%) of many facial traits [80]. Indeed, contrary to some physical phenotype traits, collecting phenotypic data for faces can be challenging. A self-reported categorisation is less useful, and measurement ideally requires the involvement of methods that are able to capture the three-dimensionality of faces. The approaches used to collect facial appearance data for studying the genetics of craniofacial variation that can be found in the literature are standard 2D photographs, magnetic resonance imaging (MRI) and 3D scanning. The latter has quickly gained a dominant position in craniofacial genetics research. It should be noted that the phenotypic assessment of facial variability from 3D images is not an easy task and makes large-scale studies and comparisons between different studies difficult. Initially, the process relied on a labour-intensive process of the manual determination of landmarks, and later, several automated landmarking methods applicable to 3D images have facilitated research on the association of facial phenotype

with genotype [80–83]. In one of the first works on the genetics of natural craniofacial variation, 11 DNA variants previously associated with cleft-lip phenotypes were tested in two European cohorts with the phenotypes captured using 2D photos or magnetic resonance images [84]. A DNA variant near the *GREM1* gene was associated with nose width, and another near the *CCDC26* gene was associated with bizygomatic distance. The first GWAS scan that was aimed at investigating normal facial variation identified only a single intronic DNA variant in the *PAX3* gene, which showed association with nasion position. This first study was conducted on a relatively small group of 2185 adolescents. The study used a 3D laser scanning method to collect phenotypic data. The 22 identified landmarks were then used to generate 54 3D and 2D distances featuring different facial characteristics. Additionally, following a previous method, a principal component analysis enabled the identification of 14 independent groups of correlated coordinates [85]. These parameters were used in association testing, which identified four associations, but only *PAX3* was replicated in an independent cohort of 1622 participants [86]. A larger GWAS analysis of almost 10,000 individuals of European origin from several cohorts used 3D MRI scans and 2D photos, and identified five genes involved in facial variation. *PAX3*, *PRDM16*, and *TP63* have previously been linked to craniofacial development, while *C5orf50* and *COL17A1* were new findings [87]. The strongest signal was again obtained for *PAX3*, which soon gained further confirmation in an independent study of about 6000 Latin Americans investigated in the large CANDELA project [40]. It is worth noting that rare variants in *PAX3*, the most replicated gene for natural variation in facial appearance, cause Waardenburg syndrome, which involves some facial dysmorphism, including a broad nasal bridge. The phenotyping regime in Adhikari's study involved a simple approach based on standard 2D photographs, and the study also implicated *DCHS2*, *RUNX2*, *GLI3*, and *PAX1* in nose morphology and *EDAR* in chin protrusion [40]. Another GWAS study, which included 3D images of 3118 individuals of European ancestry that were used to derive 20 facial distance measurements, identified several genomic regions and implicated *MAFB*, *PAX9*, *MIPOL1*, *ALX3*, *HDAC8*, and *PAX1* in normal facial variation, including the measures of eye, nose, and facial breadth. The study also provided additional evidence for the association between *PRDM16* and *C5orf50* and facial features [88]. Crouch et al. investigated the hypothesis that the DNA variants responsible for large effects on facial morphology exist in the human genome, and focused on individuals displaying extreme facial characteristics to find them. The study included 3D images of 1832 individuals from the general population as a discovery set and 1567 3D scans of twins from the TwinsUK databank, plus 33 of East Asians for replication. The original 3D scans were used to manually mark each face with 14 well-defined landmarks, allowing a mesh of 50,000–150,000 surface points in 3D space to be transformed into a set of 29,658 surface points for each face. This approach enabled the identification of three SNPs in *PCDH15*, *MBTPS1*, and *TMEM163*, genes that have previously been associated with various pathological phenotypes involving craniofacial dysmorphias [89]. The study by Claes et al. (2018) involved 2329 individuals at the discovery stage and an additional 1719 at the replication stage, and found associations for 15 loci with facial features, including four new genes, nine consistently confirmed, and two linked with pleiotropic facial phenotypic features. The study used an innovative, data-driven facial phenotyping approach based on structural correlations between about 10,000 3D quasi-landmarks, which enabled the hierarchical (global-to-local) clustering of the human face into segments [90]. This approach also yielded good results for a meta-analysis, which included 8246 European individuals and enabled the identification of 203 loci associated with normal facial variation [91] and for a study of facial features in East Africans, which investigated 2595 3D facial images collected on Tanzanian children [92]. The latter cohort was previously investigated, with two genes, *SCHIP1* and *PDE8A*, identified that were associated with measures of human facial size [83]. GWA studies investigating human facial morphology in non-European cohorts are rare. Worth noting is a GWAS conducted on an exploratory panel of Uyghurs that identified six loci important for the

genetic architecture of the human face, four of which were replicated in independent cohorts of Uyghur or southern Han Chinese [93].

3. DNA-Based Predictive Tools for Forensic Applications

Several factors determine the accuracy of DNA-based predictive methods, including high heritability of a trait, the identification of appropriate predictors, and the selection of the best mathematical approach to model development. The forensic community very early recognised the investigative potential of extracting phenotypes from DNA data. The practical importance of a simple amelogenin genetic sex test [94], and also of the inference of biogeographical ancestry [95,96], made it clear that a description of the phenotypic characteristics of a person of undetermined identity can provide important investigative leads. The variation of the *MC1R* gene was soon proposed as an indicator of red hair colour [97], while the predictive potential of the *OCA2* variation was proposed for the inference of eye colour [14]. The availability of GWAS data has made it possible to develop tools for predicting human appearance traits more effectively. The research carried out has made it possible to develop predictive tools with varying performances and practicalities of application for different physical characteristics (Table 1).

Table 1. Examples of various approaches proposed for genetic prediction of physical traits.

Physical Trait	Statistical Model	Number of Predictors in the Model	Prediction Accuracy Parameters	Ref.
Eye colour	Multinomial logistic regression (IrisPlex) [1]	6 SNPs	AUC_{brown} = 0.93 [2] $AUC_{intermediate}$ = 0.72 AUC_{blue} = 0.91	[98]
	Likelihood ratio	4 SNPs	$LR_{light-dark}$ depends on genotypes	[99]
	Multiple linear regression	3 SNPs	R^2 = 0.764	[100]
	No statistical model, classification based on genotypes	6 SNPs	Overall classification success rate (blue–green–brown): 98.94%	[101]
	Likelihood ratio	6 SNPs	$LR_{light-dark}$ depends on genotypes $AUC_{light-dark}$ = 0.925	[102]
	Bayesian naïve classifier (Snipper)	23 SNPs	Classification success rate: blue = 98.27%, green-hazel = 97.81% brown = 96.67%.	[103]
	Multiple response classification tree	4 SNPs	Classification success rate: blue = 89% intermediate = 46% brown = 94%	[104]
	No statistical model, prediction based on genotypes	5 SNPs	Overall classification success rate (blue–green–brown): 97.64%	[105]
Hair colour	Multinomial logistic regression + prediction guide (HIrisPlex) [1]	22 SNPs	Classification success rate: AUC_{blond} = 0.81 AUC_{brown} = 0.82 AUC_{black} = 0.87 AUC_{red} = 0.93	[106]

Table 1. *Cont.*

Physical Trait	Statistical Model	Number of Predictors in the Model	Prediction Accuracy Parameters	Ref.
Hair colour	Bayesian naïve classifier (Snipper)	12 SNPs	Classification success rate: blond = 92.3% brown = 76.7% black = 74.6% red = 85% Sex-related prediction accuracy differences noted	[107]
	Multinomial logistic regression	270 SNPs	AUC_{blond} = 0.74 AUC_{brown} = 0.68 AUC_{black} = 0.86 AUC_{red} = 0.86	[37]
Skin colour	Multiple linear regression, including interaction	3 SNPs	R^2 = 0.496	[100]
	No statistical model, classification based on genotypes	5 SNPs	Overall classification success rate (dark–medium–light): 62% 38% of results inconclusive	[105]
	Bayesian naïve classifier (Snipper)	10 SNPs	AUC_{white} = 0.999$AUC_{intermediate}$ = 0.803 AUC_{black} = 0.966	[108]
	Multinomial logistic regression (HIrisPlex-S) [1]	36 SNPs	AUC_{light} = 0.97 AUC_{dark} = 0.83 $AUC_{dark-black}$ = 0.96 or $AUC_{very-pale}$ = 0.74 AUC_{pale} = 0.72 $AUC_{intermediate}$ = 0.73 AUC_{dark} = 0.87 $AUC_{dark-black}$ = 0.97	[109]
	Multiple linear regression	9 SNPs	R^2 = 0.65	[110]
Freckles	Binomial logistic regression	34 SNPs + sex	$AUC_{freckled}$ = 0.809	[111]
	Multinomial logistic regression	20 SNPs + sex	$AUC_{non-freckled}$ = 0.754 $AUC_{freckled}$ = 0.657 $AUC_{heavily-freckled}$ = 0.792	[112]
Hair loss	Binomial logistic regression	20 SNPs	AUC_{bald} = 0.66 AUC_{bald} = 0.76 in men $\geq$ 50 years old	[113]
	Binomial logistic regression	14 SNPs	$AUC_{early-onset\ baldness}$ = 0.74	[114]
	Polygenic scores (weighted allele sums)	261 autosomal SNPs; 70 X chromosomal SNPs	$AUC_{severe\ baldness}$ = 0.748 (autosomal SNPs) $AUC_{severe\ baldness}$ = 0.621 (X chromosome SNPs) With autosomal and X SNPs + age included in the model: $AUC_{severe\ baldness}$ = 0.79; $AUC_{moderate\ baldness}$ = 0.70; $AUC_{slight\ baldness}$ = 0.61	[52]

Table 1. *Cont.*

Physical Trait	Statistical Model	Number of Predictors in the Model	Prediction Accuracy Parameters	Ref.
Hair shape	Binomial logistic regression	3 SNPs	$AUC_{straight}$ = 0.62	[115]
	Binomial and multinomial logistic regression	32 SNPs in binomial modelor33 SNPs in multinomial model	$AUC_{straight}$ = 0.66 in Europeans $AUC_{straight}$ = 0.79 in non-Europeans or $AUC_{straight}$ = 0.67 in Europeans AUC_{wavy} = 0.60 in Europeans AUC_{curly} = 0.60 in Europeans $AUC_{straight}$ = 0.80 in non-Europeans AUC_{wavy} = 0.61 in non-Europeans AUC_{curly} = 0.74 in non-Europeans	[116]
Hair greying	Binary and multi-class neural network	10 SNPs + age and sex in binary model or 12 SNPs + age and sex in multi-class model	$AUC_{greying}$ = 0.87 (mostly based on age) or $AUC_{no\ greying}$ = 0.86 $AUC_{mild\ greying}$ = 0.79 $AUC_{severe\ greying}$ = 0.86	[66]
Height	Polygenic scores (weighted allele sums)	54 SNPs	$AUC_{tall\ stature}$ = 0.65	[117]
	Polygenic scores (weighted allele sums)	180 SNPs	$AUC_{tall\ stature}$ = 0.75	[118]
	Polygenic scores (weighted allele sums)	689 SNPs	$AUC_{tall\ stature}$ = 0.79	[119]
	L_1-penalized regression (LASSO)	>20,000 SNPs	r = 0.64	[120]
Face	Partial least squares regression	Genomic ancestry (68 DNA variants) + sex + 24 SNPs	Genomic ancestry explains 9.6% of the total facial variation; sex independently from ancestry explains 12.9%; SNPs make a small contribution to improving facial distinctiveness	[82]
	Ridge regression	Genomic ancestry (1000 genomic Principal Components) + sex, BMI and age	Genomic ancestry and sex explain large proportion of the predictive accuracy of the model; age and BMI improve the accuracy of the model	[121]
	Simple quantitative method (principal component analysis and partial least square analysis used to extract new face traits)	277 SNPs	SSA statistic [3]: no difference between SNP-based prediction and random predictions in females; SNP-based predictions significantly better than random predictions in males	[93]

[1] SNaPshot and MPS forensically validated genetic tests for data collection available; [2] AUC—area under the ROC (receiver operating characteristic) curve, describes the general performance of the model, 1 means perfect prediction and 0.5 means random assignment; [3] SSA—a shape similarity statistic (shape space angle) developed to measure the angle between two shapes in the 3D face modelling data space.

3.1. Pigmentation Traits

In particular, the discovery of eye colour markers with large phenotypic effects has made it easy to develop pretty accurate genetic predictors of this trait. The best-known tool commonly used in the forensic field today is the IrisPlex predictive system, which includes both a genetic test for data acquisition and a mathematical algorithm for predicting the three categories of eye colour [98]. The algorithm was developed based on the systematic selection of markers made by Liu et al., who reported 24 variants from eight genes, enabling the prediction of blue and brown eye colour with a prediction accuracy expressed by an AUC of 0.91 and 0.93, respectively [122]. AUC, which stands for area under the ROC

(receiver operating characteristic) curve, describes the general performance of the model in such a way that 1 means perfect classification and 0.5 means random assignment to the phenotype categories. For forensic purposes, the number of markers from the originally identified 24 was restricted to the six with the largest effect [98,122]. The six crucial predictors included *HERC2* rs12913832, *OCA2* rs1800407, *SLC24A4* rs12896399, *SLC45A2* rs16891982, *TYR* rs1393350, and *IRF4* rs12203592. The original IrisPlex method implements a multinomial logistic regression algorithm and a simple single base extension method based on SNaPshot minisequencing, which allows the PCR amplification and genotyping of several SNPs in a multiplex reaction. Importantly, the products of primer extension are analysed using capillary electrophoresis platforms, which are commonly used in human identification testing laboratories. Other tools based on other mathematical solutions were soon developed but, essentially, each of these algorithms relied on exploiting information in the *HERC2-OCA2* gene complex. In general, these works were limited to the development of predictive algorithms using various sets of samples and mathematical approaches, but did not present specific tools for the collection of genetic data [99–104]. Notably, IrisPlex and other forensic methods of eye colour prediction can accurately predict blue and brown iris colours, but have difficulty with the prediction of intermediate eye colours [3]. Moreover, in some populations, the effect of sex was noted on prediction results [123–125]. The IrisPlex tool for the genotyping and prediction of eye colour evolved to HIrisPlex [106] and finally to the HIrisPlex-S tool [109], which were developed based on the same strategy as IrisPlex. The algorithm for hair colour prediction implemented in HIrisPlex was developed based on the investigation of a Polish population sample, which enabled the selection of 22 crucial SNPs from 11 genes for hair colour. The study showed a high level of accuracy for red and black hair colour prediction (AUC ~ 0.9) and a lower prediction accuracy for blond and brown hair colour (AUC ~ 0.8) [126]. The skin colour predictor was proposed by Walsh et al. after a systematic study of skin colour candidate variants in a sample of 2025 individuals from 31 worldwide populations. The algorithm predicted skin colour with very high accuracy, with an AUC = 0.97 for light skin colour, 0.83 dark, and 0.96 for dark-black skin colour [127]. Notably, it has been demonstrated that the original SNaPshot protocol can be replaced by the targeted massive parallel sequencing (MPS) method [128], and the HIrisPlex-S method was also adopted in a tool combining pigmentation prediction capability with ancestry inference developed by the VISAGE consortium [129]. Other studies also investigated the possibility of hair and skin colour prediction in the forensic field [100,105,107,108,110]. The Snipper Application suite deserves more attention because it provides an online tool that allows the performance of predictive calculations based on data generated by any genotyping method. The tool was originally developed for the statistical interpretation of data in ancestry inference studies, but a number of new functionalities have subsequently been added to enable the prediction of pigmentation and even age [130]. A more complete prediction of pigmentation will be provided by the developed algorithms for freckle prediction [111,112]. It is worth noting that the use of extended DNA variant sets for prediction has begun to be explored, which may lead to the development of next-generation prediction tools. For example, the previously described association work of Hysi et al. was extended to predictive modelling. Hair colour prediction was compared in two independent cohorts using prediction models based on the 258 associated SNPs and the original HIrisPlex method, and these new models outperformed the previous HIrisPlex model [37]. Further development of pigmentation predictors may also require the use of sex information, and age will naturally be needed for the final interpretation of the data [37,123]. This issue is also addressed later in the article, as sex in particular can be important for predicting other appearance traits.

3.2. Hair Loss

Numerous association studies conducted for MPB raised questions about the predictive ability of the discovered genetic variants. In 2015, a compact regression model was developed based on analysis of five SNPs from five genomic regions (Xq12, 20p11,

EBF1, *TARDBP*, and *HDAC9*), trained and validated on >600 samples from six European populations [113]. The model was shown to enable the prediction of hair loss in Europeans at an acceptable level, but only in two extreme phenotype categories, i.e., young men with significant alopecia vs. older men without symptoms of alopecia with AUC of 0.76. In the same study, Marcińska et al. also pointed to the potential role of allelic heterogeneity in determining scalp hair loss. Expanding the number of DNA variants in both crucial regions, i.e., Xq12 and 20p11, improved the accuracy of prediction, suggesting that there might be more functional variants in these loci. The extended 20-SNP regression model predicted hair loss with an AUC of 0.66 in all samples of all age categories and had the highest AUC value for the age category of ≥50 years old (AUC = 0.76; sensitivity = 67.7%; specificity = 90%), where the sensitivity refers to the ability of the model to correctly classify individuals with the particular phenotype (here baldness), while the specificity refers to the ability of the model to correctly classify individuals without this phenotype [113].

Liu et al. conducted a parallel study on the prediction of MPB in >2700 Europeans and developed a 14-SNP model that was found to predict early-onset MPB cases with a cross-validated AUC of 0.74 [114]. The accuracy of hair loss prediction status in elderly and middle-aged individuals was lower, with an AUC of 0.69–0.71. In 2017, Hagenaars and colleagues developed a polygenic predictor based on the genome-wide data generated for a large cohort of 40,000 individuals and showed that it can discriminate individuals with no signs of hair loss from those with severe baldness, with an AUC = 0.78, sensitivity = 0.74, and specificity = 0.69 [52].

3.3. Hair Shape and Other Hair Features

The first preliminary model for head hair shape was developed as a follow-up to the first GWA study conducted on hair characteristics [58], and included an analysis of three SNPs in three genes (*TCHH*, *WNT10A*, *FRAS1*), and was trained on data generated for 528 samples from Polish individuals [115]. The model was reported to predict straight hair with high accuracy but low specificity (cross-validated AUC = 0.622, sensitivity = 93.2%, specificity = 15.4%). The application of the model on an independent test set consisting of samples from six European populations and using a 65% probability threshold allowed for higher sensitivity (81.4%) and improved specificity (50.0%) of prediction, but at the same time with a very high rate of inconclusive results (66.9%). In 2018, a large-scale prediction study for hair shape prediction was conducted with more than 9600 samples used for predictor selection and model development and more than 2400 samples used for prediction model validation, collected from both European and non-European populations [116]. The binomial logistic regression model was developed to predict hair shape, defined as straight vs. non-straight, based on 32 informative SNPs from 26 loci. The model was reported to explain ~12% of hair shape variation and can predict straight vs. non-straight hair in European populations with an accuracy of AUC of 0.66, a sensitivity of 84.1% and a specificity of 34.2%. It was shown that the same set of SNP markers can predict hair shape with significantly different accuracies in Europeans and non-Europeans. For non-European samples, the AUC value was 0.79, sensitivity = 82.9%, and specificity = 49.8%. The higher prediction accuracy obtained for non-European populations compared to Europeans is due to the effect of the *EDAR* gene, which has a significant effect on the determination of straight hair in non-European populations, primarily East Asian. In addition to the binomial model, a multinomial logistic regression model was developed to allow for a higher resolution of hair shape prediction, considering three categories—straight, wavy and curly—based on an analysis of 33 SNP positions. There are few or no prediction studies of the remaining hair features. In 2016, Adhikari et al. predicted different hair traits using the GWAS data generated for Latin Americans and reported the highest accuracy of prediction for beard thickness and the lowest for hair greying, with ~18% and ~7% of the phenotypic variation explained by the associated SNPs, respectively [40]. Interestingly, for both of these traits, a large effect of age and sex on prediction was observed, explaining the additional ~11% and ~20% of the phenotypic variation, respectively, for beard thickness and greying. Age was

found to be a main predictor of hair greying in a study conducted in 2020, explaining around 48% of the variation observed in hair greying in a cohort of 849 people from Poland [66]. A binary neural network model for greying vs. no greying prediction was developed in this study based on information relating to age, sex, and 10 SNPs selected using whole-exome sequencing data analysis (e.g., *KIF1A* rs59733750, *SEMA4D* rs45483393) and literature resources (*IRF4* rs12203592, *FGF5* rs7680591). The model achieved a high accuracy of prediction with a cross-validated AUC = 0.87 (sensitivity = 0.73; specificity = 0.85) but most of the prediction information was driven by age itself, while SNPs were found to explain merely ~7% of the variation in hair greying. As mentioned earlier, age is a very important factor in predicting hair loss. Sex and age were also shown to slightly improve the accuracy of prediction of hair shape [116].

This implies that there is a need to determine the sex and age of an individual from the analysed biological sample. Information on a person's sex is usually available in criminal investigations due to the inclusion of marker for the amelogenin gene located on the X and Y chromosome in standard STR DNA profiling, as previously mentioned, whereas age can be estimated via epigenetic analysis [131].

3.4. Human Stature

Attempts at forensic human height prediction have not been particularly numerous and have been limited to the development of predictive algorithms that are not equipped with data collection tools. The reasons are related to the limitations of DNA analysis technology and stem from the need to analyse too many DNA variants. While the 5% heritability explained by the 54 DNA variants identified by the initial GWAS scans for human height was unlikely to predict the full range of human height, Aulchenko et al. tested whether it would allow the reliable prediction of extreme height. However, this turned out to be possible with only limited accuracy. Tall stature prediction was possible at AUC of 0.65, thus only moderately improving the accuracy resulting from a random hit (AUC = 0.5) [117]. Using the 180 height markers identified in the Lango Allen et al. paper improved the prediction of tall stature to AUC of 0.75 [118]. The study suggested the importance of allelic heterogeneity for the prediction of human stature. Further increasing the number of predictors to 697 reported in the paper by [72] enabled the prediction of tall stature with an AUC of 0.79 [119]. The possibilities of human height prediction have also been explored outside the forensic mainstream using a non-standard approach that has nevertheless yielded very promising results, enabling the prediction of the full range of human height at a good level of accuracy [120]. Based on the results obtained, the authors suggested changing the approach to phenotype prediction, pointing out the benefits of also including as predictors polymorphisms that do not show an association with a given trait, but only on the basis of the improved prediction accuracy obtained after their inclusion in the prediction model [132].

3.5. The Human Face

Drawing a forensic sketch based on the instructions of a witness in a criminal case is a tool that has been used for years to identify the perpetrator of a crime. People recognise each other through the high variability of facial features. Therefore, having a good understanding of the genetics of human facial variation and being able to predict this complex phenotype is a very exciting prospect for forensic DNA intelligence. The small amount of explained heritability for craniofacial traits does not bring good prospects for the prediction of human facial phenotypes. Nevertheless, attempts have been made to develop models that would allow the prediction of facial appearance. The proposed methods are based on the indirect prediction of facial phenotypes, with ancestry and sex prediction DNA data playing a key role in this regard. The method by Claes et al. implements a bootstrapped response-based imputation modelling that makes use of information on genomic ancestry and sex first to create a sketch called a base-face. At the second stage, the information in 24 SNPs associated with facial variation is used to improve the prediction outcome [82]. A similar

strategy was proposed by Lippert et al., who used the whole genome sequencing data to gain information about the sex and ancestry proportions of the individual [121]. The data on genetic face predictors did not improve facial appearance predictions, but the study showed a positive effect on the prediction of age and body mass index. The genetic prediction of facial features was also explored by Qiao at el., who developed a quantitative model based on multiple SNP loci and tried to simulate 3D face models. The study suggests that epistasis is part of the genetic architecture of facial features and concludes that the model developed should be treated as an exploratory basis for future, more advanced predictive models [93].

4. Appearance Prediction in the Era of Big Data

4.1. Appearance Trait Predicition as a Supervised Learning Task

The prediction of human externally visible characteristics using DNA markers can be treated as a supervised learning problem in which the considered appearance trait corresponds to a response (target) variable, whereas genetic markers correspond to explanatory variables (also known as features or predictors). The supervised learning models are fitted using training data, which consist of observations for which the value of the target variable is known. Depending on the type of the target variable, three tasks can be distinguished: regression (for a quantitative trait, e.g., human height), binary classification (for a binary trait, e.g., the presence of freckles), and multi-class classification (for a categorical trait, e.g., eye colour).

The specificity of the problem and the greatest challenge lies in the large number of potential features (genetic markers), which may significantly exceed the number of observations in the training data. Due to this, the use of traditional models and estimation methods (such as the maximum likelihood method in logistic regression) is not feasible. The simplest solution is to use some initial filtering method to reduce the total number of markers. However, simple filters only assess the marginal dependence between the variable and the trait; they may exclude variables that are potentially useful for the model, for example, variables that contribute by interacting with already selected ones. Therefore, there is a need to apply the estimation methods as well as feature selection approaches specially tailored to high-dimensional settings. This is one of the greatest challenges in designing learning models for appearance trait prediction.

Finally, it is important to note that traditional genome-wide association studies focus on detecting the genetic variants associated with the trait with high statistical confidence, which, in particular, includes controlling the probability of at least one rejection via multiple-testing procedures. When the prediction is the main task, the paradigm shift is needed, because focusing on the accuracy of the model becomes the main objective [133]. This approach requires the careful selection of variables. On one hand, unlike in GWAS, it is allowed to include a certain number of non-significant variables in the model, since the excessive pruning of SNPs, which may result in the discarding of some significant variables, can negatively affect prediction accuracy [132]. On the other hand, including too many spurious variables may cause the overfitting of the model and decrease its accuracy [134].

4.2. Linear Easily Interpretable Models

Despite its simplicity, the linear model and its generalisations are powerful tools for appearance trait prediction. The theory [135] and empirical evidence [136,137] suggest that in many cases the dependence between the trait and genetic markers can be captured using linear models. Several studies indicate that they frequently work on par or even better than more complex models, such as ensemble methods or neural networks [120,132,135–137], as they are not liable to overfitting. A distinct advantage of the linear models is their interpretability; the parameter value indicates how the given variable influences the dependent variable for fixed values of the remaining variables. Within the linear models, there are many methods of parameter estimation, among which the regularised (also known as penalised) maximum likelihood methods play the most prominent role in modern genetic

data analysis. First, for the regularisation methods, there are theoretical guarantees that the solution of the related optimisation problem exists and is unique, even for a high-dimensional setting. Second, some forms of the regularisation, such as lasso, ensure the sparsity of the vector of estimated coefficients, meaning that a majority of coefficients will be zero. Under some unfortunately stringent conditions, this majority will correspond to non-significant variables in the model. Thus, the selected regularisation techniques can be seen as methods of simultaneous parameter estimation and feature selection. Below, we discuss the three most important generalised linear models (linear regression, logistic regression, and multinomial regression) and the methods of parameter estimation within them.

In the case of the quantitative trait, it is natural to consider the *linear regression model*, which assumes that for an i-th observation, we have $y_i = \beta_0 + x_i^T\beta + \epsilon_i$, where y_i is the value of the target variable, β_0 is an intercept, $\beta = (\beta_1, \ldots, \beta_p)^T$ is the coefficients vector, ϵ_i is noise, and $x_i = (x_{i,1}, \ldots, x_{i,p})^T$ is a vector of features. Coordinates $x_{i,1}, \ldots, x_{i,p}$ denote the genetic markers for the i-th observation. They can be coded numerically as 0, 1, or 2, where 0 indicates the homozygosity of the major allele, 1 the heterozygosity and 2 the homozygosity of the minor allele. In the penalised least squares method, we solve:

$$\hat{\beta}_0, \hat{\beta} = \arg \min_{b_0 \in R, b \in R^p} \sum_{i=1}^{n} (y_i - b_0 - x_i^T b)^2 + \lambda \text{pen}(b),$$

where $\lambda > 0$ is the regularisation parameter that controls the penalty strength and $\text{pen}(b)$ is the penalty. For example, in the lasso method, $\text{pen}(b) = ||b||_1 = \sum_{j=1}^{p} |b_j|$, we discuss other choices below. In the case of a binary trait, the logistic regression model is usually used in which the posterior probability is modelled as:

$$P(y_i = 1|x_i) = \frac{\exp(\beta_0 + x_i^T\beta)}{1 + \exp(\beta_0 + x_i^T\beta)}$$

and parameters are estimated using the penalised maximum likelihood method:

$$\hat{\beta}_0, \hat{\beta} = \arg \max_{b_0 \in R, b \in R^p} \sum_{i=1}^{n} [y_i \log(\sigma(b_0 + x_i^T b)) + (1 - y_i) \log(1 - \sigma(b_0 + x_i^T b))] + \lambda \text{pen}(b),$$

where $\sigma(s) = \exp(s)/(1 + \exp(s))$ is the sigmoid logistic function. The multinomial logistic regression (MLR) extends the logistic model when the number of categories of the dependent variable $K > 2$. This is the most commonly used model, as usually the considered trait has multiple categories (eye colour, skin colour, hair type, etc.). The posterior probability for the k-th category is:

$$P(y_i = k|x_i) = \frac{\exp(\beta_{0k} + x_i^T\beta_k)}{1 + \sum_{k=1}^{K-1} \exp(\beta_{0k} + x_i^T\beta_k)},$$

for $k = 1, \ldots, K-1$, where β_k is a coefficients vector corresponding to the k-th category and $P(y_i = K|x_i) = 1 - \sum_{k=1}^{K-1} P(y_i = k|x_i)$. In this model, we have $K \times p$ parameters, which are estimated using the penalised maximum likelihood method. The interaction terms $x_{i,j} \times x_{i,k}$ can be included in the above models, at the cost of a significant increase in the number of parameters. In addition to linear models, additive models are an important class of models in which, instead of the linear combination $\beta_0 + x_i^T\beta$, the combination of M non-linear base functions $\beta_0 + \sum_{m=1}^{M} \beta_m h_m(x_i)$ is used. In this group, the notable approach is the MARS method (multivariate adaptive regression splines; see Section 9 in [134]) in which the functions h_m are constructed as products of so-called hinge functions in a forward stage-wise manner. Importantly, the functions h_m in MARS can capture non-linear dependencies as well as interactions between variables. Note that the considered

model is linear in predictors $h_m(x_i)$ and is an important example of the transformation of predictors method.

Regarding regularisation in the above models, the lasso penalty $\text{pen}(b) = ||b||_1$ is the most popular choice, which was successfully used in appearance trait prediction, e.g., in prediction of human height [120] or eye colour [137]. The lasso method selects features with non-zero estimated coefficients, and the number selected depends on parameter $\lambda > 0$. A small value of λ will result in a larger number of features included in the model, whereas for a larger λ, we obtain a more parsimonious model. The optimal value of λ is chosen using cross-validation or by minimising the prediction error with a validation set. An alternative to the lasso is ridge penalty $\text{pen}(b) = ||b||_2$ which, instead of performing feature selection, only shrinks the estimated parameters towards zero. The ridge penalty facilitates a reduction in the variance of the estimators, especially when the variables are highly correlated, and thus may yield an even higher accuracy for the prediction than the lasso method.

Although the lasso method has many excellent properties and high predictive power, in recent years, several modifications have been proposed in statistical and machine learning literature. For example, it has been noticed that the lasso method produces biased estimators for truly significant variables with large coefficients, and this bias does not necessarily disappear for a large sample size. To overcome this drawback, non-convex penalties, such as SCAD (smoothly clipped absolute deviation) [138] or MCP (minimax concave penalty) [139] have been proposed and effective algorithms for solving the related optimisation problems have been developed [140]. Another important line of research is focused on controlling the false discovery rate (FDR) (the expected fraction of non-significant variables that are selected for the model) instead of the much stronger control of probability that at least one non-significant variable is selected (familywise error rate). Unfortunately, the standard lasso does not control the FDR, which means that, among the selected variables, we can expect a significant portion of spurious variables. The problem is exacerbated by the fact that there is no known way of testing the significance of a specific feature based on its estimated lasso coefficient that would allow the application of one of multiple testing approaches, such as the Benjamini-Hochberg procedure [141], to control the FDR.

A notable alternative approach is the knockoff filter method [142]. It can be seen as a refinement of randomisation methods [143,144] that, by permuting the values of a studied predictor (which renders the resulting artificial predictor non-significant), creates a benchmark situation in which its usefulness can be checked. The basic idea in [142] is to construct extra variables called 'knockoff' variables, which are noisy copies of original ones but which have a certain similar correlation structure, as they allow for FDR control when standard variable selection methods (such as lasso) are applied. Namely, the lasso method is run using both the original variables and knockoff variables (thus there are $2 \times p$ variables in total). The original variable is deemed useful when its pertaining estimated coefficient is significantly larger than that of the corresponding knockoff.

The nonconvex penalties, as well as the randomisation methods, seem to be worthwhile alternatives to the lasso method for predicting human traits. The above methods are implemented, e.g., in R software, see packages glmnet (lasso and ridge), ncvreg (MCP, SCAD), knockoff (knockoff filter), and earth (MARS method).

4.3. Complex Black-Box Models

The black-box model is a class of predictive models that are able to recover complex dependencies between explanatory variables and the dependent variable, including interaction terms, and which can potentially achieve higher accuracy then linear models. The main limitations are the high computational complexity, the difficulty in interpreting the model, and the necessity of parameter tuning. In this group, ensemble methods and neural networks play the leading role. The former are usually based on decision trees [145] and overcome two limitations of single trees: their instability and tendency to overfitting. The simplest approach is bagging (bootstrap aggregating) [146] in which each tree in the

ensemble is trained using a bootstrap sample, i.e., a sample drawn with replacements from the original training data. In order to classify a new instance, each decision tree provides the classification for the input data. The majority vote classification is then chosen as the final prediction. In the case of regression, the predictions from individual tress are averaged. Another important class of models are random subspace methods (RSM), in which each base classifier is learnt using the randomly selected subset of variables [147,148]. One of the most successful and commonly used methods is random forest (RF), which can be seen as a combination of bagging and RSM. The RF uses a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features of size m, where m is a hyper-parameter. Making m smaller helps to avoid the danger of overfitting. Nowadays, the most powerful class of ensemble methods are gradient boosting (GB) algorithms (Section 10 in [134]). In GB, the subsequent models $F_1(x), \ldots, F_M(x)$ are learned sequentially, and the last model $F_M(x)$ serves as a final model. The main advantage of GB algorithms is that they are able to optimise different loss functions, depending on the considered task. The classifier in step $m+1$ (usually a decision tree) is learnt using current training data, in which the residuals from the previous model are treated as the current target variable (where the squared loss is considered, and the residuals are $y_i - F_m(x_i)$). The residuals are related to the so-called functional gradient of the loss function and, therefore, GB methods can be seen as gradient descent algorithms, which take steps in the direction of the steepest descent and converge to the minimum of the loss function. The common property of all boosting algorithms is that the current model zooms in on samples where its predecessor failed. Usually, some regularisation techniques are used in boosting algorithms to prevent overfitting. There are many versions of gradient boosting algorithms, among which XGB (extreme gradient boosting) is considered to be one of the most powerful variants [149]. The ensemble methods are controlled by different parameters, whose optimal choice may significantly improve the performance: the number of trees, the size of the random subspace (in RF and RSM), as well as the regularisation and pruning parameters.

The ensemble methods described above (RF and XGB) are often used to assess the importance of the features. The simplest approach is based on a permutation scheme and is very similar to the randomisation feature selection described above. The first method (called mean decrease accuracy) involves fitting two ensemble models (e.g., RF or XGB): the first is based on the original training data and the second is based on training data in which the values of the j-th variable are randomly permuted. The variable importance measure for the j-th variable is defined as the difference in accuracies corresponding to these two models. A large value of the difference indicates the significance of the variable. The second measure (called mean decrease impurity) is based on observing how well the given variable separates the classes. The Boruta algorithm [150], based on the above two measures, contains a testing procedure that allows the rejecting of the noisy variables. Other more sophisticated variable importance measures are also advocated for, e.g., the MCFS method [151], in which one of its major advantages is that the predictive power of each tree in the ensemble is taken into account in the measure definition.

The second important group of black-box models is artificial neural networks (ANN) [152]. The latest advances in computational and optimisation methods have made it possible to train networks with very complex architectures corresponding to large families of functions, such as convolution networks (in image recognition) and recurrent networks (in text analysis). The deep networks used today may consist of hundreds of hidden layers and can model very complex dependencies [153]. In appearance trait prediction, the feed-forward neural network is usually used. In such networks, the input signal (the vector of features for the i-th observation) is transmitted from the input layer to the output layer, which yields the prediction of the response. The hidden layers consist of artificial neurons in which the linear combination of the signals from the previous layers is computed and the signal is passed through the activation function as the input for the following layers. The models are trained using gradient algorithms (the ADAM algorithm [154] is now the state-of-the-art method) and the back-propagation algorithm is used to effectively compute

the gradient of the considered risk function [153]. A number of parameters need to be tuned in ANN, such as the number of layers, the number of neurons in each layer, and the value of the learning rate. Other spectacular advances with ANN, such as variational autoencoders (VAE), which enable latent feature analysis (see [155]), are of potential interest in appearance trait prediction. For the methods described here, see R packages randomForest, xgboost, rmcfs, Boruta, and tensorflow.

4.4. Feature Selection

Feature selection is an essential element when building predictive models, as it prevents overfitting and allows discovering the dependency structure between variables and, in particular, recovering the features that affect the target variable. In the models described above, feature selection is usually embedded in learning algorithms. For example, in linear models as well as neural networks, selection is performed via regularisation, whereas in tree-based methods, the relevant features are selected when building the tree. However, including too many potential features may significantly increase the computational cost of fitting the model. Thus, very often in practice, there is a need to apply some fast preliminary filtering method. In the machine learning community, methods based on information theory have gained the most popularity in recent years [156]. They are fast, model free, and are able to detect non-linear dependencies and interactions between variables, as well as take into account redundancies. The basic quantity used in such methods is mutual information (MI):

$$I(Y, X_k) = \sum_{x,y} P(X_k = x, Y = y) \log \frac{P(X_k = x, Y = y)}{P(X_k = x)P(Y = y)}$$

which is a non-parametric measure of dependence between some feature X_k and target variable Y. Moreover, analogously defined, the conditional mutual information $I(Y, X_k|Z)$ quantifies the dependence strength between X_k and Y given the possibly multivariate variable Z. It is commonly used in feature selection of a new predictor X_k when Z consists of predictors already chosen. Another important quantity used in genome-wise interaction studies (GWIS) is interaction information (II):

$$II(Y, X_j, X_k) = I((X_k, X_j), Y) - I(Y, X_k) - I(Y, X_j)$$

which measures the interaction strength between X_k and X_j for the prediction of Y. The positive value of II indicates a synergistic interaction, whereas a negative value indicates redundancy. II has been successfully used in many genetic studies to detect epistasis [157,158], and also in the context of appearance trait prediction, such as human pigmentation [159]. It has been shown that the methods based on II are able to detect interactions that remain undetected by the logistic regression model [160].

The existing filters based on MI are forward sequential procedures that, in each step, add a candidate feature X_k to the set of already selected features S. The quality of a candidate feature can be assessed using various criteria, and the representative one is CIFE (conditional infomax feature extraction) [156,161]. It adds candidate X_k, being the maximizer of $I(Y, X_k) + \sum_{j \in S} II(Y, X_k, X_j)$. The CIFE takes into account the marginal dependence between a candidate feature and the target variable, as well as interactions between the candidate feature and the previously selected features. Methods taking into account higher order interactions are also considered [162]. In practice, it is important to decide at which step to stop the procedure of adding new candidate variables, with the possible solution based on the approximate distribution of the criterion function when a candidate feature is not significant, as proposed in [163].

5. The Need for a High-Throughput, Low-Input DNA Sequencing Method in Forensic Science

The limitations of DNA sequencing technologies used in the forensic field are increasingly problematic because they are hindering the implementation of new methods that can improve law enforcement and justice, and which are therefore important for the safety of society. The lack of a suitable method for generating large amounts of SNP data from degraded DNA, validated for use in forensics, was considered to be a barrier to the forensic implementation of investigative genetic genealogy, an approach that was very successful at solving a number of criminal cases [1,164]. Such a method also seems to be essential for developing next-generation tools for the DNA-based prediction of appearance traits, which requires information derived from hundreds or even thousands of SNPs. It may be argued that the optimal method for all the applications developed for forensic DNA intelligence would be whole-genome sequencing (WGS). Notably, WGS that uses high-throughput methods (massively parallel sequencing) has revolutionised the studies of ancient DNA and enabled a better understanding of human evolutionary history. Similarities between forensic genetics and palaeogenetics, especially in terms of the specificity of research material with a high content of inhibitors and small amounts of highly fragmented DNA, and the enormous success of palaeogenetics in the analysis of such samples, prompts a closer look at the methods developed in this field. Several technological advancements were crucial for the effective analysis of ancient DNA, including the very efficient extraction of short ancient DNA fragments, the implementation of the uracil-DNA glycosylase (UDG) protocol for the selective removal of damaged sections of ancient DNA, improved protocols for library preparation, and, finally, progress has also been made in high-throughput DNA sequencing [165,166]. The major advantage for ancient DNA research brought about by high-throughput sequencing technology is the ability to sequence very short DNA fragments. Research material analysed in forensic DNA laboratories is not as degraded as ancient samples, and current DNA extraction methods are efficient and effective at removing inhibitors. Therefore, the transfer of DNA analysis protocols from palaeogenomics to forensic genomics should perhaps primarily focus on library preparation methods that work well with low-input DNA. Standard library preparation protocols are optimised for large amounts of DNA and perform poorly in the case of samples containing degraded DNA. However, a number of modified protocols have been proposed to reduce the requirement for DNA inputs to be at subnanogram quantities.

One category of protocols involves library construction based on double-stranded DNA. A first protocol was described by Meyer and Kircher in 2010, and this was laborious and had limitations that resulted in the losses of ancient DNA sequences due to incompatible adapter combinations and three purification steps prior to amplification [167]. Double-stranded library preparation protocols involve the blunt-end repair of the degraded DNA fragments, the non-directional blunt-end ligation of two adapters and the fill-in of the nicks formatted between adapters and the DNA fragment [168]. A more advanced alternative of double stranded library preparation method is the protocol proposed by Carøe et al., named blunt-end-single-tube. As the name suggests, the protocol is carried out in a single tube and relies on heat denaturation instead of purification between the subsequent steps of end-repair, the ligation of double-stranded adapters to the 5′ ends, and adapter fill-in [169].

The second approach for library preparation from samples containing low amounts of degraded DNA is particularly interesting, as it implies a process of library construction based on single-stranded DNA, allowing the use of DNA that was preserved in a single-stranded state and which is considered to be more efficient compared to double-stranded approaches. The original protocol for single-stranded library preparation, although it recovered more endogenous DNA, was very expensive and laborious [170]. However, the protocol evolved to a simplified version proposed in Gansauge et al., 2017 [171]. This is a method that involves the dephosphorylation of the template DNA, the splinted ligation of a biotinylated adapter to the 3′ end, bonding to streptavidin beads, annealing an extension

primer to allow the synthesis of a second strand, and the ligation of the 5′ end of a double-stranded adapter to the 3′ end of the newly synthesized strand. The authors also proposed an automated version of this protocol [172]. An interesting modification of the single-strand library preparation method was recently proposed by Kapp et al. (2021). The advantage of the method, named the Santa Cruz Reaction, relies on simplicity and cost effectiveness. The method converts single-stranded DNA into sequencing libraries using a single enzymatic reaction, enabling the simultaneous directional splinted ligation of Illumina's P5 and P7 adapters [173]. Technological improvements in ancient DNA analysis have resulted in significant progress in sequencing efficiency. Whole-genome data from ancient hominin material were generated with an average sequence coverage of only 1.3-fold in 2010 [174] and 30-fold in 2012 [175]. The usefulness of these protocols was also confirmed in clinical research of problematic biological material, including formalin-fixed paraffin embedded tissues [176]. The future will show whether the protocols developed in palaeogenomics can be easily transferred to forensic genomics. This would undoubtedly be extremely helpful for the further development of forensic DNA intelligence methods.

6. Concluding Remarks

Research on the genetic architecture of natural variation in the human physical phenotype is growing in scale and involves different human populations. The genetic prediction of physical appearance traits occupies an important place in forensic research, although the available tools are limited to the least complex traits, mainly pigmentation. Notably, there are examples of using predictive methods that have been developed by the forensic community in ancient DNA research, and which have been carried out in the field of molecular anthropology [177–179] and in the identification of historical figures [180–182], which is further evidence that molecular anthropology and forensic genetics have a lot in common. Some DNA-based predictive tools developed by the forensic community have been implemented in commercial kits. The most famous ForenSeq kit allows the analysis of HIrisPlex SNPs and therefore the prediction of eye and hair colour [183]. The HIrisPlex-S variants are also available in the Ion AmpliSeq™ PhenoTrivium Panel [184]. Predicting reliable sketches in forensic science is highly desirable at the investigation stage. For this reason, there are reports of police using private companies offering services in this regard, particularly for facial appearance prediction. For example, the Snapshot Forensic DNA Phenotyping System offered by Parabon NanoLabs claims to facilitate the accurate prediction of genetic ancestry, eye colour, hair colour, skin colour, freckling, and face shape [185]. Further research offers the opportunity to better understand the evolutionary and genetic basis of human appearance traits. The prospect of future studies on the heritability of complex traits and the exploration of the importance of rare DNA variants, as well as epistatic interactions of the second and higher orders, seems interesting. The explanation of heritability will consequently enable a more reliable prediction of physical phenotype. Undoubtedly, however, the application of next-generation predictive methods, which must rely on much larger sets of predictors and more sophisticated statistical and machine learning algorithms, will require improvements in the technology of DNA polymorphism analysis used in the forensic field. Proper interpretation of the data requires knowledge of age, which is best determined via DNA methylation analysis. However, DNA methylation analysis requires the largest amounts of DNA, so in studying biological traces for intelligence purposes, it would be beneficial to develop more sensitive age prediction methods. The application of novel predictive approaches will also require answers to important ethical questions arising from the use of high-throughput DNA analysis methods.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kling, D.; Phillips, C.; Kennett, D.; Tillmar, A. Investigative genetic genealogy: Current methods, knowledge and practice. *Forensic Sci. Int. Genet.* **2021**, *52*, 102474. [CrossRef]
2. Phillips, C. Forensic genetic analysis of bio-geographical ancestry. *Forensic Sci. Int. Genet.* **2015**, *18*, 49–65. [CrossRef] [PubMed]
3. Kayser, M. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. *Forensic Sci. Int. Genet.* **2015**, *18*, 33–48. [CrossRef]
4. Lee, H.Y.; Lee, S.D.; Shin, K.J. Forensic DNA methylation profiling from evidence material for investigative leads. *BMB Rep.* **2016**, *49*, 359–369. [CrossRef] [PubMed]
5. Polderman, T.J.; Benyamin, B.; de Leeuw, C.A.; Sullivan, P.F.; van Bochoven, A.; Visscher, P.M.; Posthuma, D. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* **2015**, *47*, 702–709. [CrossRef]
6. Botstein, D.; Risch, N. Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **2003**, *33*, 228–237. [CrossRef]
7. Visscher, P.M.; Brown, M.A.; McCarthy, M.I.; Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **2012**, *90*, 7–24. [CrossRef] [PubMed]
8. Uffelmann, E.; Huang, Q.Q.; Munung, N.S.; DeVries, J.; Okada, Y.; Martin, A.R.; Martin, H.C.; Lappalainen, T.; Posthuma, D. Genome-wide association studies. *Nat. Rev. Methods Primers* **2021**, *1*, 59. [CrossRef]
9. Clark, P.; Stark, A.E.; Walsh, R.J.; Jardine, R.; Martin, N.G. A twin study of skin reflectance. *Ann. Hum. Biol.* **1981**, *8*, 529–541. [CrossRef] [PubMed]
10. Lin, B.D.; Mbarek, H.; Willemsen, G.; Dolan, C.V.; Fedko, I.O.; Abdellaoui, A.; De Geus, E.J.; Boomsma, D.I.; Hottenga, J.-J. Heritability and Genome-Wide Association Studies for Hair Color in a Dutch Twin Family Based Sample. *Genes* **2015**, *6*, 559–576. [CrossRef]
11. Bito, L.Z.; Matheny, A.; Cruickshanks, K.J.; Nondahl, D.M.; Carino, O.B. Eye Color Changes Past Early Childhood: The Louisville Twin Study. *Arch. Ophthalmol.* **1997**, *115*, 659–663. [CrossRef]
12. Eiberg, H.; Mohr, J. Assignment of genes coding for brown eye colour (BEY2) and brown hair colour (HCL3) on chromosome 15q. *Eur. J. Hum. Genet.* **1996**, *4*, 237–241. [CrossRef] [PubMed]
13. Rebbeck, T.R.; Kanetsky, P.A.; Walker, A.H.; Holmes, R.; Halpern, A.C.; Schuchter, L.M.; Elder, D.E.; Guerry, D. P gene as an inherited biomarker of human eye color. *Cancer Epidemiol. Biomark. Prev.* **2002**, *11*, 782–784.
14. Frudakis, T.; Thomas, M.; Gaskin, Z.; Venkateswarlu, K.; Chandra, K.S.; Ginjupalli, S.; Gunturi, S.; Natrajan, S.; Ponnuswamy, V.K.; Ponnuswamy, K.N. Sequences associated with human iris pigmentation. *Genetics* **2003**, *165*, 2071–2083. [CrossRef] [PubMed]
15. Duffy, D.L.; Montgomery, G.W.; Chen, W.; Zhao, Z.Z.; Le, L.; James, M.R.; Hayward, N.K.; Martin, N.G.; Sturm, R.A. A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. *Am. J. Hum. Genet.* **2007**, *80*, 241–252. [CrossRef] [PubMed]
16. Branicki, W.; Brudnik, U.; Kupiec, T.; Wolańska-Nowak, P.; Szczerbińska, A.; Wojas-Pelc, A. Association of polymorphic sites in the OCA2 gene with eye colour using the tree scanning method. *Ann. Hum. Genet.* **2008**, *72*, 184–192. [CrossRef]
17. Sulem, P.; Gudbjartsson, D.F.; Stacey, S.N.; Helgason, A.; Rafnar, T.; Magnusson, K.P.; Manolescu, A.; Karason, A.; Palsson, A.; Thorleifsson, G.; et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat. Genet.* **2007**, *39*, 1443–1452. [CrossRef]
18. Sturm, R.A.; Duffy, D.L.; Zhao, Z.Z.; Leite, F.P.; Stark, M.S.; Hayward, N.K.; Martin, N.G.; Montgomery, G.W. 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*, 424–431. [CrossRef]
19. Eiberg, H.; Troelsen, J.; Nielsen, M.; Mikkelsen, A.; Mengel-From, J.; Kjaer, K.W.; Hansen, L. Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum. Genet.* **2008**, *123*, 177–187. [CrossRef]
20. Visser, M.; Kayser, M.; Grosveld, F.; Palstra, R.J. Genetic variation in regulatory DNA elements: The case of OCA2 transcriptional regulation. *Pigment Cell Melanoma Res.* **2014**, *27*, 169–177. [CrossRef] [PubMed]
21. Valverde, P.; Healy, E.; Jackson, I.; Rees, J.L.; Thody, A.J. Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nat. Genet.* **1995**, *11*, 328–330. [CrossRef] [PubMed]
22. Rees, J.L. Genetics of hair and skin color. *Annu. Rev. Genet.* **2003**, *37*, 67–90. [CrossRef] [PubMed]
23. Flanagan, N.; Healy, E.; Ray, A.; Philips, S.; Todd, C.; Jackson, I.J.; Birch-Machin, M.A.; Rees, J.L. Pleiotropic effects of the melanocortin 1 receptor (MC1R) gene on human pigmentation. *Hum. Mol. Genet.* **2000**, *9*, 2531–2537. [CrossRef] [PubMed]
24. Bastiaens, M.; ter Huurne, J.; Gruis, N.; Bergman, W.; Westendorp, R.; Vermeer, B.J.; Bouwes Bavinck, J.N. The melanocortin-1-receptor gene is the major freckle gene. *Hum. Mol. Genet.* **2001**, *10*, 1701–1708. [CrossRef]
25. Han, J.; Kraft, P.; Nan, H.; Guo, Q.; Chen, C.; Qureshi, A.; Hankinson, S.E.; Hu, F.B.; Duffy, D.L.; Zhao, Z.Z.; et al. A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genet.* **2008**, *4*, e1000074. [CrossRef]
26. Manolio, T.A.; Collins, F.S.; Cox, N.J.; Goldstein, D.B.; Hindorff, L.A.; Hunter, D.J.; McCarthy, M.I.; Ramos, E.M.; Cardon, L.R.; Chakravarti, A.; et al. Finding the missing heritability of complex diseases. *Nature* **2009**, *461*, 747–753. [CrossRef] [PubMed]
27. Nelson, R.M.; Pettersson, M.E.; Carlborg, Ö. A century after Fisher: Time for a new paradigm in quantitative genetics. *Trends Genet.* **2013**, *29*, 669–676. [CrossRef]

28. Simcoe, M.; Valdes, A.; Liu, F.; Furlotte, N.A.; Evans, D.M.; Hemani, G.; Ring, S.M.; Smith, G.D.; Duffy, D.L.; Zhu, 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*, eabd1239. [CrossRef] [PubMed]
29. Norton, H.L. The color of normal: How a Eurocentric focus erases pigmentation complexity. *Am. J. Hum. Biol.* **2021**, *33*, e23554. [CrossRef]
30. Lamason, R.L.; Mohideen, M.A.; Mest, J.R.; Wong, A.C.; Norton, H.L.; Aros, M.C.; Jurynec, M.J.; Mao, X.; Humphreville, V.R.; Humbert, J.E.; et al. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **2005**, *310*, 1782–1786. [CrossRef]
31. Liu, F.; Visser, M.; Duffy, D.L.; Hysi, P.G.; Jacobs, L.C.; Lao, O.; Zhong, K.; Walsh, S.; Chaitanya, L.; Wollstein, A.; et al. Genetics of skin color variation in Europeans: Genome-wide association studies with functional follow-up. *Hum. Genet.* **2015**, *134*, 823–835. [CrossRef]
32. Sulem, P.; Gudbjartsson, D.F.; Stacey, S.N.; Helgason, A.; Rafnar, T.; Jakobsdottir, M.; Steinberg, S.; Gudjonsson, S.A.; Palsson, A.; Thorleifsson, G.; et al. Two newly identified genetic determinants of pigmentation in Europeans. *Nat. Genet.* **2008**, *40*, 835–837. [CrossRef] [PubMed]
33. Edwards, M.; Bigham, A.; Tan, J.; Li, S.; Gozdzik, A.; Ross, K.; Jin, L.; Parra, E.J. Association of the OCA2 polymorphism His615Arg with melanin content in east Asian populations: Further evidence of convergent evolution of skin pigmentation. *PLoS Genet.* **2010**, *6*, e1000867. [CrossRef]
34. Stokowski, R.P.; Pant, P.V.; Dadd, T.; Fereday, A.; Hinds, D.A.; Jarman, C.; Filsell, W.; Ginger, R.S.; Green, M.R.; van der Ouderaa, F.J.; et al. A genomewide association study of skin pigmentation in a South Asian population. *Am. J. Hum. Genet.* **2007**, *81*, 1119–1132. [CrossRef]
35. Crawford, N.G.; Kelly, D.E.; Hansen, M.E.B.; Beltrame, M.H.; Fan, S.; Bowman, S.L.; Jewett, E.; Ranciaro, A.; Thompson, S.; Lo, Y.; et al. Loci associated with skin pigmentation identified in African populations. *Science* **2017**, *358*, eaan8433. [CrossRef] [PubMed]
36. Martin, A.R.; Lin, M.; Granka, J.M.; Myrick, J.W.; Liu, X.; Sockell, A.; Atkinson, E.G.; Werely, C.J.; Möller, M.; Sandhu, M.S.; et al. An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. *Cell* **2017**, *171*, 1340–1353. [CrossRef]
37. Hysi, P.G.; Valdes, A.M.; Liu, F.; Furlotte, N.A.; Evans, D.M.; Bataille, V.; Visconti, A.; Hemani, G.; McMahon, G.; Ring, S.M.; et al. Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability. *Nat. Genet.* **2018**, *50*, 652–656. [CrossRef]
38. Morgan, M.D.; Pairo-Castineira, E.; Rawlik, K.; Canela-Xandri, O.; Rees, J.; Sims, D.; Tenesa, A.; Jackson, I.J. Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. *Nat. Commun.* **2018**, *9*, 5271. [CrossRef] [PubMed]
39. Medland, S.E.; Zhu, G.; Martin, N.G. Estimating the heritability of hair curliness in twins of European ancestry. *Twin Res. Hum. Genet.* **2009**, *12*, 514–518. [CrossRef] [PubMed]
40. Adhikari, K.; Fontanil, T.; Cal, S.; Mendoza-Revilla, J.; Fuentes-Guajardo, M.; Chacón-Duque, J.C.; Al-Saadi, F.; Johansson, J.A.; Quinto-Sanchez, M.; Acuña-Alonzo, V.; et al. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. *Nat. Commun.* **2016**, *7*, 10815. [CrossRef]
41. Nyholt, D.R.; Gillespie, N.A.; Heath, A.C.; Martin, N.G. Genetic basis of male pattern baldness. *J. Investig. Dermatol.* **2003**, *121*, 1561–1564. [CrossRef] [PubMed]
42. Pirastu, N.; Joshi, P.K.; de Vries, P.S.; Cornelis, M.C.; McKeigue, P.M.; Keum, N.; Franceschini, N.; Colombo, M.; Giovannucci, E.L.; Spiliopoulou, A.; et al. GWAS for male-pattern baldness identifies 71 susceptibility loci explaining 38% of the risk. *Nat. Commun.* **2017**, *8*, 1584, Erratum in: *Nat Commun.* **2018**, *9*, 2536. [CrossRef] [PubMed]
43. Yap, C.X.; Sidorenko, J.; Wu, Y.; Kemper, K.E.; Yang, J.; Wray, N.R.; Robinson, M.R.; Visscher, P.M. Dissection of genetic variation and evidence for pleiotropy in male pattern baldness. *Nat. Commun.* **2018**, *9*, 5407. [CrossRef] [PubMed]
44. Gunn, D.A.; Rexbye, H.; Griffiths, C.E.; Murray, P.G.; Fereday, A.; Catt, S.D.; Tomlin, C.C.; Strongitharm, B.H.; Perrett, D.I.; Catt, M. Why some women look young for their age. *PLoS ONE* **2009**, *4*, e8021. [CrossRef]
45. Weissbrod, O.; Flint, J.; Rosset, S. Estimating SNP-Based Heritability and Genetic Correlation in Case-Control Studies Directly and with Summary Statistics. *Am. J. Hum. Genet.* **2018**, *103*, 89–99. [CrossRef] [PubMed]
46. Visscher, P.M.; Hill, W.G.; Wray, N.R. Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.* **2008**, *9*, 255–266. [CrossRef] [PubMed]
47. Richards, J.B.; Yuan, X.; Geller, F.; Waterworth, D.; Bataille, V.; Glass, D.; Song, K.; Waeber, G.; Vollenweider, P.; Aben, K.K.; et al. Male-pattern baldness susceptibility locus at 20p11. *Nat. Genet.* **2008**, *40*, 1282–1284. [CrossRef]
48. Hillmer, A.M.; Brockschmidt, F.F.; Hanneken, S.; Eigelshoven, S.; Steffens, M.; Flaquer, A.; Herms, S.; Becker, T.; Kortüm, A.K.; Nyholt, D.R.; et al. Susceptibility variants for male-pattern baldness on chromosome 20p11. *Nat. Genet.* **2008**, *40*, 1279–1281. [CrossRef] [PubMed]
49. Brockschmidt, F.F.; Heilmann, S.; Ellis, J.A.; Eigelshoven, S.; Hanneken, S.; Herold, C.; Moebus, S.; Albblas, M.A.; Lippke, B.; Kluck, N.; et al. Susceptibility variants on chromosome 7p21.1 suggest HDAC9 as a new candidate gene for male-pattern baldness. *Br. J. Dermatol.* **2011**, *165*, 1293–1302. [CrossRef]
50. Li, R.; Brockschmidt, F.F.; Kiefer, A.K.; Stefansson, H.; Nyholt, D.R.; Song, K.; Vermeulen, S.H.; Kanoni, S.; Glass, D.; Medland, S.E.; et al. Six novel susceptibility Loci for early-onset androgenetic alopecia and their unexpected association with common diseases. *PLoS Genet.* **2012**, *8*, e1002746. [CrossRef]

51. Pickrell, J.K.; Berisa, T.; Liu, J.Z.; Ségurel, L.; Tung, J.Y.; Hinds, D.A. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **2016**, *48*, 709–717. [CrossRef]
52. Hagenaars, S.P.; Hill, W.D.; Harris, S.E.; Ritchie, S.J.; Davies, G.; Liewald, D.C.; Gale, C.R.; Porteous, D.J.; Deary, I.J.; Marioni, R.E. Genetic prediction of male pattern baldness. *PLoS Genet.* **2017**, *13*, e1006594. [CrossRef] [PubMed]
53. Heilmann-Heimbach, S.; Herold, C.; Hochfeld, L.M.; Hillmer, A.M.; Nyholt, D.R.; Hecker, J.; Javed, A.; Chew, E.G.; Pechlivanis, S.; Drichel, D.; et al. Meta-analysis identifies novel risk loci and yields systematic insights into the biology of male-pattern baldness. *Nat. Commun.* **2017**, *8*, 14694. [CrossRef]
54. Liang, B.; Yang, C.; Zuo, X.; Li, Y.; Ding, Y.; Sheng, Y.; Zhou, F.; Cheng, H.; Zheng, X.; Chen, G.; et al. Genetic variants at 20p11 confer risk to androgenetic alopecia in the Chinese Han population. *PLoS ONE* **2013**, *8*, e71771. [CrossRef]
55. Zhuo, F.L.; Xu, W.; Wang, L.; Wu, Y.; Xu, Z.L.; Zhao, J.Y. Androgen receptor gene polymorphisms and risk for androgenetic alopecia: A meta-analysis. *Clin. Exp. Dermatol.* **2012**, *37*, 104–111. [CrossRef]
56. Loh, P.R.; Kichaev, G.; Gazal, S.; Schoech, A.P.; Price, A.L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **2018**, *50*, 906–908. [CrossRef]
57. Kichaev, G.; Bhatia, G.; Loh, P.R.; Gazal, S.; Burch, K.; Freund, M.K.; Schoech, A.; Pasaniuc, B.; Price, A.L. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am. J. Hum. Genet.* **2019**, *104*, 65–75. [CrossRef] [PubMed]
58. Medland, S.E.; Nyholt, D.R.; Painter, J.N.; McEvoy, B.P.; McRae, A.F.; Zhu, G.; Gordon, S.D.; Ferreira, M.A.; Wright, M.J.; Henders, A.K.; et al. Common variants in the trichohyalin gene are associated with straight hair in Europeans. *Am. J. Hum. Genet.* **2009**, *85*, 750–755. [CrossRef] [PubMed]
59. Wu, S.; Tan, J.; Yang, Y.; Peng, Q.; Zhang, M.; Li, J.; Lu, D.; Liu, Y.; Lou, H.; Feng, Q.; et al. Genome-wide scans reveal variants at EDAR predominantly affecting hair straightness in Han Chinese and Uyghur populations. *Hum. Genet.* **2016**, *135*, 1279–1286. [CrossRef]
60. Liu, F.; Chen, Y.; Zhu, G.; Hysi, P.G.; Wu, S.; Adhikari, K.; Breslin, K.; Pospiech, E.; Hamer, M.A.; Peng, F.; et al. Meta-analysis of genome-wide association studies identifies 8 novel loci involved in shape variation of human head hair. *Hum. Mol. Genet.* **2018**, *27*, 559–575. [CrossRef]
61. Fujimoto, A.; Kimura, R.; Ohashi, J.; Omi, K.; Yuliwulandari, R.; Batubara, L.; Mustofa, M.S.; Samakkarn, U.; Settheetham-Ishida, W.; Ishida, T.; et al. A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum. Mol. Genet.* **2008**, *17*, 835–843. [CrossRef]
62. Tan, J.; Yang, Y.; Tang, K.; Sabeti, P.C.; Jin, L.; Wang, S. The adaptive variant EDARV370A is associated with straight hair in East Asians. *Hum. Genet.* **2013**, *132*, 1187–1191. [CrossRef]
63. Pośpiech, E.; Lee, S.D.; Kukla-Bartoszek, M.; Karłowska-Pik, J.; Woźniak, A.; Boroń, M.; Zubańska, M.; Bronikowska, A.; Hong, S.R.; Lee, J.H.; et al. Variation in the RPTN gene may facilitate straight hair formation in Europeans and East Asians. *J. Dermatol. Sci.* **2018**, *91*, 331–334. [CrossRef]
64. Endo, C.; Johnson, T.A.; Morino, R.; Nakazono, K.; Kamitsuji, S.; Akita, M.; Kawajiri, M.; Yamasaki, T.; Kami, A.; Hoshi, Y.; et al. Genome-wide association study in Japanese females identifies fifteen novel skin-related trait associations. *Sci. Rep.* **2018**, *8*, 8974. [CrossRef] [PubMed]
65. Wu, S.; Zhang, M.; Yang, X.; Peng, F.; Zhang, J.; Tan, J.; Yang, Y.; Wang, L.; Hu, Y.; Peng, Q.; et al. Genome-wide association studies and CRISPR/Cas9-mediated gene editing identify regulatory variants influencing eyebrow thickness in humans. *PLoS Genet.* **2018**, *14*, e1007640. [CrossRef] [PubMed]
66. Pośpiech, E.; Kukla-Bartoszek, M.; Karłowska-Pik, J.; Zieliński, P.; Woźniak, A.; Boroń, M.; Dąbrowski, M.; Zubańska, M.; Jarosz, A.; Grzybowski, T.; et al. Exploring the possibility of predicting human head hair greying from DNA using whole-exome and targeted NGS data. *BMC Genom.* **2020**, *21*, 538. [CrossRef]
67. Gudbjartsson, D.F.; Walters, G.B.; Thorleifsson, G.; Stefansson, H.; Halldorsson, B.V.; Zusmanovich, P.; Sulem, P.; Thorlacius, S.; Gylfason, A.; Steinberg, S.; et al. Many sequence variants affecting diversity of adult human height. *Nat. Genet.* **2008**, *40*, 609–615. [CrossRef]
68. Lettre, G.; Jackson, A.U.; Gieger, C.; Schumacher, F.R.; Berndt, S.I.; Sanna, S.; Eyheramendy, S.; Voight, B.F.; Butler, J.L.; Guiducci, C.; et al. Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat. Genet.* **2008**, *40*, 584–591. [CrossRef] [PubMed]
69. Weedon, M.N.; Lango, H.; Lindgren, C.M.; Wallace, C.; Evans, D.M.; Mangino, M.; Freathy, R.M.; Perry, J.R.; Stevens, S.; Hall, A.S.; et al. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat. Genet.* **2008**, *40*, 575–583. [CrossRef]
70. Lango Allen, H.; Estrada, K.; Lettre, G.; Berndt, S.I.; Weedon, M.N.; Rivadeneira, F.; Willer, C.J.; Jackson, A.U.; Vedantam, S.; Raychaudhuri, S.; et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **2010**, *467*, 832–838. [CrossRef]
71. Yang, J.; Benyamin, B.; McEvoy, B.P.; Gordon, S.; Henders, A.K.; Nyholt, D.R.; Madden, P.A.; Heath, A.C.; Martin, N.G.; Montgomery, G.W.; et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **2010**, *42*, 565–569. [CrossRef]
72. Wood, A.R.; Esko, T.; Yang, J.; Vedantam, S.; Pers, T.H.; Gustafsson, S.; Chu, A.Y.; Estrada, K.; Luan, J.; Kutalik, Z.; et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **2014**, *46*, 1173–1186. [CrossRef]

73. Yengo, L.; Sidorenko, J.; Kemper, K.E.; Zheng, Z.; Wood, A.R.; Weedon, M.N.; Frayling, T.M.; Hirschhorn, J.; Yang, J.; Visscher, P.M. Meta-analysis of genome-wide association studies for height and body mass index in ~700,000 individuals of European ancestry. *Hum. Mol. Genet.* **2018**, *27*, 3641–3649. [CrossRef] [PubMed]
74. Kaiser, J. Growth spurt for height genetics. *Science* **2020**, *370*, 645. [CrossRef] [PubMed]
75. Zoledziewska, M.; Sidore, C.; Chiang, C.W.K.; Sanna, S.; Mulas, A.; Steri, M.; Busonero, F.; Marcus, J.H.; Marongiu, M.; Maschio, A.; et al. Height-reducing variants and selection for short stature in Sardinia. *Nat. Genet.* **2015**, *47*, 1352–1356. [CrossRef] [PubMed]
76. He, M.; Xu, M.; Zhang, B.; Liang, J.; Chen, P.; Lee, J.Y.; Johnson, T.A.; Li, H.; Yang, X.; Dai, J.; et al. Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. *Hum. Mol. Genet.* **2015**, *24*, 1791–1800. [CrossRef]
77. Akiyama, M.; Ishigaki, K.; Sakaue, S.; Momozawa, Y.; Horikoshi, M.; Hirata, M.; Matsuda, K.; Ikegawa, S.; Takahashi, A.; Kanai, M.; et al. Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nat. Commun.* **2019**, *10*, 4393. [CrossRef]
78. Marouli, E.; Graff, M.; Medina-Gomez, C.; Lo, K.S.; Wood, A.R.; Kjaer, T.R.; Fine, R.S.; Lu, Y.; Schurmann, C.; Highland, H.M.; et al. Rare and low-frequency coding variants alter human adult height. *Nature* **2017**, *542*, 186–190. [CrossRef]
79. Roosenboom, J.; Hens, G.; Mattern, B.C.; Shriver, M.D.; Claes, P. Exploring the Underlying Genetics of Craniofacial Morphology through Various Sources of Knowledge. *Biomed. Res. Int.* **2016**, *2016*, 3054578. [CrossRef] [PubMed]
80. Tsagkrasoulis, D.; Hysi, P.; Spector, T.; Montana, G. Heritability maps of human face morphology through large-scale automated three-dimensional phenotyping. *Sci. Rep.* **2017**, *7*, 45885. [CrossRef]
81. Guo, J.; Mei, X.; Tang, K. Automatic landmark annotation and dense correspondence registration for 3D human facial images. *BMC Bioinform.* **2013**, *14*, 232. [CrossRef]
82. Claes, P.; Liberton, D.K.; Daniels, K.; Rosana, K.M.; Quillen, E.E.; Pearson, L.N.; McEvoy, B.; Bauchet, M.; Zaidi, A.A.; Yao, W.; et al. Modeling 3D facial shape from DNA. *PLoS Genet.* **2014**, *10*, e1004224. [CrossRef] [PubMed]
83. Cole, J.B.; Manyama, M.; Kimwaga, E.; Mathayo, J.; Larson, J.R.; Liberton, D.K.; Lukowiak, K.; Ferrara, T.M.; Riccardi, S.L.; Li, M.; et al. Genomewide Association Study of African Children Identifies Association of SCHIP1 and PDE8A with Facial Size and Shape. *PLoS Genet.* **2016**, *12*, e1006174. [CrossRef]
84. Boehringer, S.; van der Lijn, F.; Liu, F.; Günther, M.; Sinigerova, S.; Nowak, S.; Ludwig, K.U.; Herberz, R.; Klein, S.; Hofman, A.; et al. Genetic determination of human facial morphology: Links between cleft-lips and normal variation. *Eur. J. Hum. Genet.* **2011**, *19*, 1192–1197. [CrossRef] [PubMed]
85. Toma, A.M.; Zhurov, A.I.; Playle, R.; Marshall, D.; Rosin, P.L.; Richmond, S. The assessment of facial variation in 4747 British school children. *Eur. J. Orthod.* **2012**, *34*, 655–664. [CrossRef] [PubMed]
86. Paternoster, L.; Zhurov, A.I.; Toma, A.M.; Kemp, J.P.; St Pourcain, B.; Timpson, N.J.; McMahon, G.; McArdle, W.; Ring, S.M.; Smith, G.D.; et al. Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. *Am. J. Hum. Genet.* **2012**, *90*, 478–485. [CrossRef]
87. Liu, F.; van der Lijn, F.; Schurmann, C.; Zhu, G.; Chakravarty, M.M.; Hysi, P.G.; Wollstein, A.; Lao, O.; de Bruijne, M.; Ikram, M.A.; et al. A genome-wide association study identifies five loci influencing facial morphology in Europeans. *PLoS Genet.* **2012**, *8*, e1002932. [CrossRef] [PubMed]
88. Shaffer, J.R.; Orlova, E.; Lee, M.K.; Leslie, E.J.; Raffensperger, Z.D.; Heike, C.L.; Cunningham, M.L.; Hecht, J.T.; Kau, C.H.; Nidey, N.L.; et al. Genome-Wide Association Study Reveals Multiple Loci Influencing Normal Human Facial Morphology. *PLoS Genet.* **2016**, *12*, e1006149. [CrossRef] [PubMed]
89. Crouch, D.J.M.; Winney, B.; Koppen, W.P.; Christmas, W.J.; Hutnik, K.; Day, T.; Meena, D.; Boumertit, A.; Hysi, P.; Nessa, A.; et al. Genetics of the human face: Identification of large-effect single gene variants. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E676–E685. [CrossRef]
90. Claes, P.; Roosenboom, J.; White, J.D.; Swigut, T.; Sero, D.; Li, J.; Lee, M.K.; Zaidi, A.; Mattern, B.C.; Liebowitz, C.; et al. Genome-wide mapping of global-to-local genetic effects on human facial shape. *Nat. Genet.* **2018**, *50*, 414–423. [CrossRef] [PubMed]
91. White, J.D.; Indencleef, K.; Naqvi, S.; Eller, R.J.; Hoskens, H.; Roosenboom, J.; Lee, M.K.; Li, J.; Mohammed, J.; Richmond, S.; et al. Insights into the genetic architecture of the human face. *Nat. Genet.* **2021**, *53*, 45–53. [CrossRef] [PubMed]
92. Liu, C.; Lee, M.K.; Naqvi, S.; Hoskens, H.; Liu, D.; White, J.D.; Indencleef, K.; Matthews, H.; Eller, R.J.; Li, J.; et al. Genome scans of facial features in East Africans and cross-population comparisons reveal novel associations. *PLoS Genet.* **2021**, *17*, e1009695. [CrossRef]
93. Qiao, L.; Yang, Y.; Fu, P.; Hu, S.; Zhou, H.; Peng, S.; Tan, J.; Lu, Y.; Lou, H.; Lu, D.; et al. Genome-wide variants of Eurasian facial shape differentiation and a prospective model of DNA based face prediction. *J. Genet. Genom.* **2018**, *45*, 419–432. [CrossRef]
94. Sullivan, K.M.; Mannucci, A.; Kimpton, C.P.; Gill, P. A rapid and quantitative DNA sex test: Fluorescence-based PCR analysis of X-Y homologous gene amelogenin. *Biotechniques* **1993**, *15*, 636–638, 640–641. [PubMed]
95. Evett, I.W.; Pinchin, R.; Buffery, C. An investigation of the feasibility of inferring ethnic origin from DNA profiles. *J. Forensic Sci. Soc.* **1992**, *32*, 301–306. [CrossRef]
96. Shriver, M.D.; Smith, M.W.; Jin, L.; Marcini, A.; Akey, J.M.; Deka, R.; Ferrell, R.E. Ethnic-affiliation estimation by use of population-specific DNA markers. *Am. J. Hum. Genet.* **1997**, *60*, 957–964.
97. Grimes, E.A.; Noake, P.J.; Dixon, L.; Urquhart, A. Sequence polymorphism in the human melanocortin 1 receptor gene as an indicator of the red hair phenotype. *Forensic Sci. Int.* **2001**, *122*, 124–129. [CrossRef]

98. Walsh, S.; Liu, F.; Ballantyne, K.N.; van Oven, M.; Lao, O.; Kayser, M. IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Sci. Int. Genet.* **2011**, *5*, 170–180. [CrossRef]
99. Mengel-From, J.; Børsting, C.; Sanchez, J.J.; Eiberg, H.; Morling, N. Human eye colour and HERC2, OCA2 and MATP. *Forensic Sci. Int. Genet.* **2010**, *4*, 323–328. [CrossRef]
100. Valenzuela, R.K.; Henderson, M.S.; Walsh, M.H.; Garrison, N.A.; Kelch, J.T.; Cohen-Barak, O.; Erickson, D.T.; John Meaney, F.; Bruce Walsh, J.; Cheng, K.C.; et al. Predicting phenotype from genotype: Normal pigmentation. *J. Forensic Sci.* **2010**, *55*, 315–322. [CrossRef] [PubMed]
101. Spichenok, O.; Budimlija, Z.M.; Mitchell, A.A.; Jenny, A.; Kovacevic, L.; Marjanovic, D.; Caragine, T.; Prinz, M.; Wurmbach, E. Prediction of eye and skin color in diverse populations using seven SNPs. *Forensic Sci. Int. Genet.* **2011**, *5*, 472–478. [CrossRef] [PubMed]
102. Pośpiech, E.; Draus-Barini, J.; Kupiec, T.; Wojas-Pelc, A.; Branicki, W. Prediction of eye color from genetic data using Bayesian approach. *J. Forensic Sci.* **2012**, *57*, 880–886. [CrossRef]
103. Ruiz, Y.; Phillips, C.; Gomez-Tato, A.; Alvarez-Dios, J.; de Cal, M.C.; Cruz, R.; Maroñas, O.; Söchtig, J.; Fondevila, M.; Rodriguez-Cid, M.J.; et al. Further development of forensic eye color predictive tests. *Forensic Sci. Int. Genet.* **2013**, *7*, 28–40. [CrossRef]
104. Allwood, J.S.; Harbison, S. SNP model development for the prediction of eye colour in New Zealand. *Forensic Sci. Int. Genet.* **2013**, *7*, 444–452. [CrossRef] [PubMed]
105. Hart, K.L.; Kimura, S.L.; Mushailov, V.; Budimlija, Z.M.; Prinz, M.; Wurmbach, E. Improved eye- and skin-color prediction based on 8 SNPs. *Croat. Med. J.* **2013**, *54*, 248–256. [CrossRef]
106. Walsh, S.; Liu, F.; Wollstein, A.; Kovatsi, L.; Ralf, A.; Kosiniak-Kamysz, A.; Branicki, W.; Kayser, M. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Sci. Int. Genet.* **2013**, *7*, 98–115. [CrossRef]
107. Söchtig, J.; Phillips, C.; Maroñas, O.; Gómez-Tato, A.; Cruz, R.; Alvarez-Dios, J.; de Cal, M.Á.; Ruiz, Y.; Reich, K.; Fondevila, M.; et al. Exploration of SNP variants affecting hair colour prediction in Europeans. *Int. J. Legal Med.* **2015**, *129*, 963–975. [CrossRef] [PubMed]
108. Maroñas, O.; Phillips, C.; Söchtig, J.; Gomez-Tato, A.; Cruz, R.; Alvarez-Dios, J.; de Cal, M.C.; Ruiz, Y.; Fondevila, M.; Carracedo, Á.; et al. Development of a forensic skin colour predictive test. *Forensic Sci. Int. Genet.* **2014**, *13*, 34–44. [CrossRef]
109. Chaitanya, L.; Breslin, K.; Zuñiga, S.; Wirken, L.; Pośpiech, E.; Kukla-Bartoszek, M.; Sijen, T.; Knijff, P.; Liu, F.; Branicki, W.; 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. [CrossRef]
110. Andersen, J.D.; Meyer, O.S.; Simão, F.; Jannuzzi, J.; Carvalho, E.; Andersen, M.M.; Pereira, V.; Børsting, C.; Morling, N.; Gusmão, L. Skin pigmentation and genetic variants in an admixed Brazilian population of primarily European ancestry. *Int. J. Legal Med.* **2020**, *134*, 1569–1579. [CrossRef]
111. Hernando, B.; Ibañez, M.V.; Deserio-Cuesta, J.A.; Soria-Navarro, R.; Vilar-Sastre, I.; Martinez-Cadenas, C. Genetic determinants of freckle occurrence in the Spanish population: Towards ephelides prediction from human DNA samples. *Forensic Sci. Int. Genet.* **2018**, *33*, 38–47. [CrossRef]
112. Kukla-Bartoszek, M.; Pośpiech, E.; Woźniak, A.; Boroń, M.; Karłowska-Pik, J.; Teisseyre, P.; Zubańska, M.; Bronikowska, A.; Grzybowski, T.; Płoski, R.; et al. DNA-based predictive models for the presence of freckles. *Forensic Sci. Int. Genet.* **2019**, *42*, 252–259. [CrossRef]
113. Marcińska, M.; Pośpiech, E.; Abidi, S.; Andersen, J.D.; van den Berge, M.; Carracedo, Á.; Eduardoff, M.; Marczakiewicz-Lustig, A.; Morling, N.; Sijen, T.; et al. Evaluation of DNA variants associated with androgenetic alopecia and their potential to predict male pattern baldness. *PLoS ONE* **2015**, *10*, e0127852. [CrossRef] [PubMed]
114. Liu, F.; Hamer, M.A.; Heilmann, S.; Herold, C.; Moebus, S.; Hofman, A.; Uitterlinden, A.G.; Nöthen, M.M.; van Duijn, C.M.; Nijsten, T.E.; et al. Prediction of male-pattern baldness from genotypes. *Eur. J. Hum. Genet.* **2016**, *24*, 895–902. [CrossRef] [PubMed]
115. Pośpiech, E.; Karłowska-Pik, J.; Marcińska, M.; Abidi, S.; Andersen, J.D.; Berge, M.V.D.; Carracedo, Á.; Eduardoff, M.; Freire-Aradas, A.; Morling, N.; et al. Evaluation of the predictive capacity of DNA variants associated with straight hair in Europeans. *Forensic Sci. Int. Genet.* **2015**, *19*, 280–288. [CrossRef] [PubMed]
116. Pośpiech, E.; Chen, Y.; Kukla-Bartoszek, M.; Breslin, K.; Aliferi, A.; Andersen, J.D.; Ballard, D.; Chaitanya, L.; Freire-Aradas, A.; van der Gaag, K.J.; 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. [CrossRef]
117. Aulchenko, Y.S.; Struchalin, M.V.; Belonogova, N.M.; Axenovich, T.I.; Weedon, M.N.; Hofman, A.; Uitterlinden, A.G.; Kayser, M.; Oostra, B.A.; van Duijn, C.M.; et al. Predicting human height by Victorian and genomic methods. *Eur. J. Hum. Genet.* **2009**, *17*, 1070–1075. [CrossRef] [PubMed]
118. Liu, F.; Hendriks, A.E.; Ralf, A.; Boot, A.M.; Benyi, E.; Sävendahl, L.; Oostra, B.A.; van Duijn, C.; Hofman, A.; Rivadeneira, F.; et al. Common DNA variants predict tall stature in Europeans. *Hum. Genet.* **2014**, *133*, 587–597. [CrossRef]
119. Liu, F.; Zhong, K.; Jing, X.; Uitterlinden, A.G.; Hendriks, A.E.J.; Drop, S.L.S.; Kayser, M. Update on the predictability of tall stature from DNA markers in Europeans. *Forensic Sci. Int. Genet.* **2019**, *42*, 8–13. [CrossRef]
120. Lello, L.; Avery, S.G.; Tellier, L.; Vazquez, A.I.; de Los Campos, G.; Hsu, S.D.H. Accurate Genomic Prediction of Human Height. *Genetics* **2018**, *210*, 477–497. [CrossRef]

121. Lippert, C.; Sabatini, R.; Maher, M.C.; Kang, E.Y.; Lee, S.; Arikan, O.; Harley, A.; Bernal, A.; Garst, P.; Lavrenko, V.; et al. Identification of individuals by trait prediction using whole-genome sequencing data. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 10166–10171. [CrossRef]
122. Liu, F.; van Duijn, K.; Vingerling, J.R.; Hofman, A.; Uitterlinden, A.G.; Janssens, A.C.; Kayser, M. Eye color and the prediction of complex phenotypes from genotypes. *Curr. Biol.* **2009**, *19*, R192–R193. [CrossRef] [PubMed]
123. Martinez-Cadenas, C.; Peña-Chilet, M.; Ibarrola-Villava, M.; Ribas, G. Gender is a major factor explaining discrepancies in eye colour prediction based on HERC2/OCA2 genotype and the IrisPlex model. *Forensic Sci. Int. Genet.* **2013**, *7*, 453–460. [CrossRef]
124. Pietroni, C.; Andersen, J.D.; Johansen, P.; Andersen, M.M.; Harder, S.; Paulsen, R.; Børsting, C.; Morling, N. The effect of gender on eye colour variation in European populations and an evaluation of the IrisPlex prediction model. *Forensic Sci. Int. Genet.* **2014**, *1*, 1–6. [CrossRef] [PubMed]
125. Pośpiech, E.; Karłowska-Pik, J.; Ziemkiewicz, B.; Kukla, M.; Skowron, M.; Wojas-Pelc, A.; Branicki, W. Further evidence for population specific differences in the effect of DNA markers and gender on eye colour prediction in forensics. *Int. J. Legal. Med.* **2016**, *130*, 923–934. [CrossRef]
126. Branicki, W.; Liu, F.; van Duijn, K.; Draus-Barini, J.; Pośpiech, E.; Walsh, S.; Kupiec, T.; Wojas-Pelc, A.; Kayser, M. Model-based prediction of human hair color using DNA variants. *Hum. Genet.* **2011**, *129*, 443–454. [CrossRef] [PubMed]
127. Walsh, S.; Chaitanya, L.; Breslin, K.; Muralidharan, C.; Bronikowska, A.; Pospiech, E.; Koller, J.; Kovatsi, L.; Wollstein, A.; Branicki, W.; et al. Global skin colour prediction from DNA. *Hum. Genet.* **2017**, *136*, 847–863. [CrossRef]
128. Breslin, K.; Wills, B.; Ralf, A.; Ventayol Garcia, M.; Kukla-Bartoszek, M.; Pospiech, E.; Freire-Aradas, A.; Xavier, C.; Ingold, S.; de La Puente, M.; 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. [CrossRef] [PubMed]
129. Xavier, C.; de la Puente, M.; Mosquera-Miguel, A.; Freire-Aradas, A.; Kalamara, V.; Vidaki, A.; Gross, T.; Revoir, A.; Pośpiech, E.; Kartasińska, E.; et al. Development and validation of the VISAGE AmpliSeq basic tool to predict appearance and ancestry from DNA. *Forensic Sci. Int. Genet.* **2020**, *48*, 102336. [CrossRef]
130. Phillips, C.; Salas, A.; Sánchez, J.J.; Fondevila, M.; Gómez-Tato, A.; Alvarez-Dios, J.; Calaza, M.; de Cal, M.C.; Ballard, D.; Lareu, M.V.; et al. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci. Int. Genet.* **2007**, *1*, 273–280. [CrossRef] [PubMed]
131. Noroozi, R.; Ghafouri-Fard, S.; Pisarek, A.; Rudnicka, J.; Spólnicka, M.; Branicki, W.; Taheri, M.; Pośpiech, E. DNA methylation-based age clocks: From age prediction to age reversion. *Ageing Res. Rev.* **2021**, *68*, 101314. [CrossRef]
132. De Los Campos, G.; Vazquez, A.I.; Hsu, S.; Lello, L. Complex-Trait Prediction in the Era of Big Data. *Trends Genet.* **2018**, *34*, 746–754. [CrossRef]
133. Shmueli, G. To Explain or to Predict? *Stat. Sci.* **2010**, *25*, 289–310. [CrossRef]
134. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, NY, USA, 2009.
135. Hill, G.H.; Goddard, M.E.; Vissher, P.M. Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. *PLoS Genet.* **2008**, *4*, 1–10. [CrossRef]
136. Katsara, M.A.; Branicki, W.; Walsh, S.; Kayser, M.; Nothnagel, M. Evaluation of supervised machine-learning methods for predicting appearance traits from DNA. *Forensic Sci. Int. Genet.* **2021**, *53*, 1–9. [CrossRef] [PubMed]
137. Kukla-Bartoszek, M.; Teisseyre, P.; Pośpiech, E.; Karłowska-Pik, J.; Zieliński, P.; Woźniak, A.; Boroń, M.; Dąbrowski, M.; Zubańska, M.; Jarosz, A.; et al. Searching for improvements in predicting human eye colour from DNA. *Int. J. Legal Med.* **2021**, *135*, 2175–2187. [CrossRef] [PubMed]
138. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Stat. Assoc.* **2001**, *96*, 1348–1361. [CrossRef]
139. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2001**, *38*, 894–942. [CrossRef]
140. Breheny, P.; Huang, J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **2011**, *5*, 232–253. [CrossRef]
141. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and multiple approach to multiple testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289–300.
142. Barber, R.F.; Candes, E.J. Controlling the False Discovery Rate by knockoffs. *Ann. Stat.* **2015**, *43*, 2055–2085. [CrossRef]
143. Tsamardinos, I.; Borboudakis, G. Permutation testing improves Bayesian Network learning. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Barcelona, Spain, 19–23 September 2010; pp. 322–337.
144. Tansey, W.; Veitch, V.; Zhang, H.; Rabadan, R.; Blei, D. The Holdout Randomisation Test for Feature Selection in Black Box Models. *J. Comput. Graph. Stat.* **2021**, 1–37. Available online: https://arxiv.org/abs/1811.00645 (accessed on 8 December 2021).
145. Bishop, C. *Pattern Recognition and Machine Learning*, 1st ed.; Springer: New York, NY, USA, 2006.
146. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
147. Ho, T.K. The Random Subspace Method for Constructing Decision Forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 832–844.
148. Mielniczuk, J.; Teisseyre, P. Using Random Subspace Method for Prediction and Variable Importance Assessment in Regression. *Comput. Stat. Data Anal.* **2014**, *71*, 725–742. [CrossRef]
149. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.

150. Kursa, M.; Rudnicki, W. Feature Selection with the Boruta Package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]
151. Dramiński, M.; Rada-Iglesias, A.; Enroth, S.; Wadelius, C.; Koronacki, J.; Komorowski, J. Monte Carlo feature selection for supervised classification. *Bioinformatics* **2008**, *24*, 110–117. [CrossRef] [PubMed]
152. Rosenblatt, F. The Perceptron—A Perceiving and Recognizing Automaton, Report. 1957. Available online: https://blogs.umass.edu/brain-wars/files/2016/03/rosenblatt-1957.pdf (accessed on 8 December 2021).
153. Goodfellow, J.; Bengio, Y.; Courville, A. *Deep Learning*, 1st ed.; MIT Press: Cambridge, MA, USA, 2016.
154. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–13.
155. Kingsma, D.; Welling, M. Autoencoding variational Bayes. In Proceedings of the 2nd Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014; pp. 1–14.
156. Brown, G.; Pocock, A.; Zhao, M.-J.; Luján, M. Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. *J. Mach. Learn. Res.* **2012**, *13*, 27–66.
157. Moore, J.H.; Hu, T. Epistatic analysis using information theory. *Methods Mol. Biol.* **2015**, *1253*, 257–268. [PubMed]
158. Cordell, H.J. Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* **2002**, *11*, 2463–2468. [CrossRef]
159. Pośpiech, E.; Wojas-Pelc, A.; Walsh, S.; Liu, F.; Maeda, H.; Ishikawa, T.; Skowron, M.; Kayser, M.; Branicki, W. The common occurrence of epistasis in the determination of human pigmentation and its impact on DNA-based pigmentation phenotype prediction. *Forensic Sci. Int. Genet.* **2014**, *11*, 64–72. [CrossRef] [PubMed]
160. Mielniczuk, J.; Teisseyre, P. Deeper Look at Two Concepts of Measuring Gene-Gene Interactions: Logistic Regression and Interaction Information Revisited. *Genet. Epidemiol.* **2018**, *42*, 187–200. [CrossRef] [PubMed]
161. Lin, D.; Tang, X. Conditional infomax learning: An integrated framework for feature extraction and fusion. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 68–82.
162. Vinh, N.X.; Zhou, S.; Chan, J.; Bailey, J. Can high-order dependencies improve mutual information based feature selection? *Pattern Recognit.* **2016**, *53*, 46–58. [CrossRef]
163. Mielniczuk, J.; Teisseyre, P. Stopping rules for mutual information-based feature selection. *Neurocomputing* **2019**, *358*, 255–274. [CrossRef]
164. Tillmar, A.; Sjölund, P.; Lundqvist, B.; Klippmark, T.; Älgenäs, C.; Green, H. Whole-genome sequencing of human remains to enable genealogy DNA database searches—A case report. *Forensic Sci. Int. Genet.* **2020**, *46*, 102233. [CrossRef] [PubMed]
165. Hofman, C.A.; Warinner, C. Ancient DNA 101: An introductory guide in the era of high-throughput sequencing. *SAA Rec.* **2019**, *19*, 18–25.
166. Hofreiter, M.; Sneberger, J.; Pospisek, M.; Vanek, D. Progress in forensic bone DNA analysis: Lessons learned from ancient DNA. *Forensic Sci. Int. Genet.* **2021**, *54*, 102538. [CrossRef]
167. Meyer, M.; Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, *2010*, pdb.prot5448. [CrossRef]
168. Psonis, N.; Vassou, D.; Kafetzopoulos, D. Testing a series of modifications on genomic library preparation methods for ancient or degraded DNA. *Anal. Biochem.* **2021**, *623*, 114193. [CrossRef]
169. Carøe, C.; Gopalakrishnan, S.; Vinner, L.; Mak, S.S.T.; Sinding, M.-H.S.; Samaniego, J.A.; Wales, N.; Sicheritz-Pontén, T.; Gilbert, M.T.P. Single-tube library preparation for degraded DNA. *Methods Ecol. Evol.* **2018**, *9*, 410–419. [CrossRef]
170. Gansauge, M.T.; Meyer, M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat. Protoc.* **2013**, *8*, 737–748. [CrossRef]
171. Gansauge, M.T.; Gerber, T.; Glocke, I.; Korlevic, P.; Lippik, L.; Nagel, S.; Riehl, L.M.; Schmidt, A.; Meyer, M. Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Res.* **2017**, *45*, e79. [CrossRef]
172. Gansauge, M.T.; Aximu-Petri, A.; Nagel, S.; Meyer, M. Manual and automated preparation of single-stranded DNA libraries for the sequencing of DNA from ancient biological remains and other sources of highly degraded DNA. *Nat. Protoc.* **2020**, *15*, 2279–2300. [CrossRef] [PubMed]
173. Kapp, J.D.; Green, R.E.; Shapiro, B. A Fast and Efficient Single-stranded Genomic Library Preparation Method Optimized for Ancient DNA. *J. Hered.* **2021**, *112*, 241–249. [CrossRef] [PubMed]
174. Green, R.E.; Krause, J.; Briggs, A.W.; Maricic, T.; Stenzel, U.; Kircher, M.; Patterson, N.; Li, H.; Zhai, W.; Fritz, M.H.; et al. A draft sequence of the Neandertal genome. *Science* **2010**, *328*, 710–722. [CrossRef] [PubMed]
175. Meyer, M.; Kircher, M.; Gansauge, M.T.; Li, H.; Racimo, F.; Mallick, S.; Schraiber, J.G.; Jay, F.; Prüfer, K.; de Filippo, C.; et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **2012**, *338*, 222–226. [CrossRef] [PubMed]
176. Stiller, M.; Sucker, A.; Griewank, K.; Aust, D.; Baretton, G.B.; Schadendorf, D.; Horn, S. Single-strand DNA library preparation improves sequencing of formalin-fixed and paraffin-embedded (FFPE) cancer DNA. *Oncotarget* **2016**, *7*, 59115–59128. [CrossRef]
177. Rasmussen, M.; Li, Y.; Lindgreen, S.; Pedersen, J.S.; Albrechtsen, A.; Moltke, I.; Metspalu, M.; Metspalu, E.; Kivisild, T.; Gupta, R.; et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **2010**, *463*, 757–762. [CrossRef]
178. Keller, A.; Graefen, A.; Ball, M.; Matzas, M.; Boisguerin, V.; Maixner, F.; Leidinger, P.; Backes, C.; Khairat, R.; Forster, M.; et al. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat. Commun.* **2012**, *3*, 698. [CrossRef] [PubMed]

179. Olalde, I.; Allentoft, M.E.; Sánchez-Quinto, F.; Santpere, G.; Chiang, C.W.; DeGiorgio, M.; Prado-Martinez, J.; Rodríguez, J.A.; Rasmussen, S.; Quilez, J.; et al. Derived immune and ancestral pigmentation alleles in a 7000-year-old Mesolithic European. *Nature* **2014**, *507*, 225–228. [CrossRef]
180. Bogdanowicz, W.; Allen, M.; Branicki, W.; Lembring, M.; Gajewska, M.; Kupiec, T. Genetic identification of putative remains of the famous astronomer Nicolaus Copernicus. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 12279–12282. [CrossRef]
181. Kupiec, T.; Branicki, W. Genetic examination of the putative skull of Jan Kochanowski reveals its female sex. *Croat. Med. J.* **2011**, *52*, 403–409. [CrossRef] [PubMed]
182. King, T.E.; Fortes, G.G.; Balaresque, P.; Thomas, M.G.; Balding, D.; Maisano Delser, P.; Neumann, R.; Parson, W.; Knapp, M.; Walsh, S.; et al. Identification of the remains of King Richard III. *Nat. Commun.* **2014**, *5*, 5631. [CrossRef] [PubMed]
183. Salvo, N.M.; Janssen, K.; Kirsebom, M.K.; Meyer, O.S.; Berg, T.; Olsen, G.H. Predicting eye and hair colour in a Norwegian population using Verogen's ForenSeq™ DNA signature prep kit. *Forensic Sci. Int. Genet.* **2021**, *56*, 102620. [CrossRef]
184. 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* **2020**, *11*, 1398. [CrossRef] [PubMed]
185. Parabon NanoLabs. Available online: https://snapshot.parabon-nanolabs.com (accessed on 8 December 2021).