Genetic Approaches to Appearance and Ancestry

Improving Forensic DNA Analysis

Lakshmi Chaitanya Reddy

ISBN/EAN: 9789462333987

The work presented in this thesis was performed at the department of Genetic Identification (formerly: Forensic Molecular Biology), Erasmus MC University Medical Center, Rotterdam, The Netherlands. This work was financially supported by the Erasmus MC and by a grant from the Netherlands Genomics Initiative (NGI)/ Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands (FGCN).

Layout and Cover design: Lakshmi Chaitanya Reddy, Stijn Kooij
Cover design: image adapted from http://stock.adobe.com

Printed by: Gildeprint - Enschede

Genetic Approaches to Appearance and Ancestry

Improving forensic DNA analysis

Genetische benaderingen voor vormgeving en voorouders

Verbetering van forensische DNA-analyse

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus
Prof.dr. H.A.P. Pols
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
woensdag 07 december 2016 om 11:30 uur
door

Lakshmi Chaitanya Reddy
geboren te Bangalore, India

Erasmus University Rotterdam

Erasmus

Promotiecommissie

Promotor:	Prof.dr. M.H. Kayser
Overige leden:	Prof.dr. A.G. Uitterlinden
	Prof.dr. R. DeCorte
	Prof.dr. P. de Knijff

Table of Contents

Chapter 1

General Introduction

Forensic DNA analysis has made conspicuous progress as an integral tool for contemporary forensic investigations since the discovery of DNA fingerprinting by Sir Alec Jeffreys in 1985 [1]. Forensic DNA analysis can identify individuals, exonerate the innocents, xxmay also help reconstructing the crime scene and help solve cases. Forensic DNA analysis has emerged as an indispensable and robust human identification tool to aid in criminal offences, kinship analysis, missing person investigations and mass disaster victim identification (DVI). There is a continuous demand for innovative tools and technologies in forensic DNA analysis to improve existing applications and to instigate the development of new applications aiming to solve more cases rapidly. Over the past few years the field of forensic genetics has witnessed rapid scientific and technological advances, with repertory of genetic markers, enabling efficient extraction of large amounts of genetic information from challenging biological samples, with increasing resolution and sensitivity of detection [2].

Forensic DNA analysis is based on a variety of techniques focusing on the substantial amount of variation in the sequences or polymorphisms in the human genome between individuals. Combination of several factors such as recombination, mutation, inbreeding, selection, genetic drift and migration affect the human genetic variation [3]. Short tandem repeats (STRs) and single nucleotide polymorphisms (SNPs) are the common types of genetic markers used. Typically in the forensic field, autosomal DNA markers are used but in certain cases STRs and SNPs from the non-recombining part of the Y chromosome and from the non-recombining mitochondrial DNA are also applied.

Standard forensic DNA analysis is based on the comparison of the DNA profiles obtained from genotyping the STR markers from evidentiary samples collected at the crime scene and reference samples taken directly from known individuals (or from items used by them). When the evidentiary DNA profile does not match with the perpetrator's reference DNA profile or with any in a forensic DNA database, this standard comparison approach cannot help in solving criminal cases [4]. Similarly, when STRs are used in missing person identification or mass disaster victim identification, the DNA profiles from these individuals are compared to DNA profiles generated from their direct reference samples or ante-mortem samples, or to those of known or assumed putative relatives, if ante-mortem samples are unavailable. In the absence of such reference samples or known relatives, standard DNA profiling is not informative and hence such individuals may remain unknown [4]. Recent advancements in the field of forensic genetics presents new possibilities for extracting

information from DNA, which otherwise is not possible from STRs, and is directed towards a source of intelligence. This has led to the development of an alternative approach, i.e., Forensic DNA Phenotyping (FDP) to overcome the limitations of standard DNA profiling. Unlike standard DNA profiling, FDP is not based on the comparative approach and thus provides useful information in cases where standard DNA profiling fails because the person to be identified is not already known with his/her DNA profile to the investigating authorities [4]. FDP refers to the prediction of externally visible characteristics (EVC), such as eye, hair, and skin colour, and prediction of bio-geographical ancestry (BGA) from the biological samples of unknown DNA donors found at crime scenes or obtained from the biological remains of unknown missing persons. The outcomes of FDP about EVCs and BGA, serving as biological witnesses [5] is used to narrow down the pool of potential suspects, which typically is large in cases where the investigators have no clue who the sample donor may be. This would provide leads and impetus to police investigations when the comparative approach of DNA identification fails to generate a match and there is no possibility further to trace the unknown perpetrators. Single nucleotide polymorphisms (SNPs) are currently the marker of choice for DNA-based intelligence tools. Typing a battery of EVC and ancestry informative SNPs would generate a composite sketch of the unknown sample donor that would aid in finding them by focussed police investigation. Given that these data are corroborated statistically from reference datasets, they provide a significant advantage over eyewitness testimony, which is known to be unreliable and comes without statistical evidence whether it is true or not [6].

Single Nucleotide Polymorphisms

SNPs are single base substitutions (Figure 1) that account for 90% of all the sequence variation in the human genome [7] and occur with a high frequency of about 1 in 100 to 300 bases [8]. Their low mutation rate (few mutations every 100,000,000 generations per locus [9,10] and short amplicon size makes them valuable genetic markers in forensic analyses, especially in highly degraded or limited DNA samples.

Employing SNPs has ascertained that the affluence of genetic information can be garnered from challenging biological samples and degraded DNA, especially when it is not possible from the analysis of the battery of STR loci [2,11,12]. This is because the DNA fragment targeted for SNP analysis can be much shorter due to the single base change than

with STRs where the polymorphism is produced by the number of tandem repeats resulting in longer stretches of polymorphic DNA relative to a SNP. DNA degradation over relatively short periods of time, as being most relevant in the forensic context, typically results in DNA fragmentation so that the length of the targeted DNA fragment is a crucial element for a successful analysis. Given that enough markers are used to compensate for the reduced variation per each bi-allelic locus relative to the multi-allelic STRs, SNPs are also useful for kinship analysis, which can be troubled by the mutation rate of STRs that is about 100,000 higher than that of SNPs [9]. Their particular suitability for successfully analysing fragmented DNA makes them useful DNA markers in missing person identification and mass disaster victim identification cases involving evidence material harbouring degraded DNA [2]. Because of their low mutation rate, SNPs allow bio-geographic ancestry to be traced from DNA, and their potential functional involvement in phenotypes either by changing amino acids of the coded protein or serving as regulator of gene expression allows their use to predict EVCs from DNA.

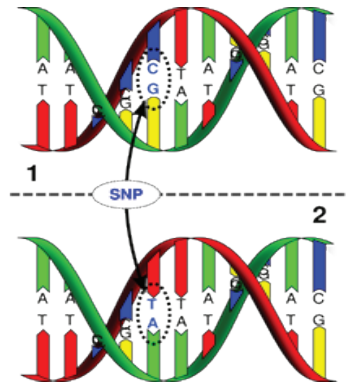


Figure 1: SNP - DNA strand 1 differs from DNA strand 2 at a single base-pair location (a C/T polymorphism) *(Taken from http://www.dnabaser.com/articles/SNP/SNP-single-nucleotide-polymorphism.html.)*

Prediction of externally visible characteristics from SNPs: Eye, hair and skin colour

Pigmentation is the most conspicuous, variable and discerning of human traits, noticed in the colouration of the eye (iris), hair and skin. This has created a penchant for identifying the pigmentation genes and their polymorphic variants that represent the different phenotypic traits existing between individuals within and between populations. Human pigmentation is the result of several complex genetic and biochemical pathways [13]. Melanin, the pigment

present in the basal layer of the iris, hair bulb and skin epidermis, determines the eye, hair and skin colour [14]. Melanin is synthesized within melanosomes in the melanocytes. Two main types of melanin are present in humans - eumelanin and pheomelanin. Eumelanin, produced in eumelanosomes, is a brown-black pigment responsible for the dark colour, and pheomelanin, produced in pheomelanosomes, is a red-yellow pigment [13]. The overall pigmentation is governed by the amount and type of melanin (ratio of eumelanin to pheomelanin) and the shape and distribution of the melanosomes [13,14].

SNPs in a number of pigmentation genes were found to be associated with the eye, hair and skin colour phenotypes [15-17]. The human genome project and the HapMap projects delivered scientific information about the human DNA sequence and its haplotype block structure. This information was used to develop commercial SNP microarrays allowing the parallel analysis of tens (initially) to hundreds to thousands of SNPs up to one million, which are used in genome wide association studies (GWAS) to identify trait associated SNPs. This information enabled the selection of highly associated phenotypic SNPs and testing their trait predictive effect in prediction studies subsequent to GWAS. Past few years witnessed the progress of research in better understanding the genetic basis of eye, hair and skin colour via GWASs [15,17-20], as well as the DNA predictability of these three pigmentation traits via dedicated prediction studies [5,21-25]. At this moment, human pigmentation traits are the EVCs that most accurately are predictable from DNA, allowing practical FDP applications (Figure 2). For all other EVCs, the genetic knowledge currently is not complete enough to consider practical FDP although progress in better understanding the genetic basis of several EVCs is constantly going on [26].

Figure 2: DNA can be extracted from skeletal remains and bodily fluids; and the inherent information can be used to predict eye, hair and skin colour of an individual.
(Taken from Chapter 17 of Primorac, D., & Schanfield, M., 2014 [27])

Eye Colour

Human eye (iris) colour represents the first EVC (after sex) that could be predicted from DNA with useful accuracy. Human eye colour can be described on the basis of a continuous spectrum from the lightest shades of blue to the darkest shades of brown or black with intermediate colours such as green, yellow or hazel as well mixed-colour eyes consisting of areas in different colour [28]. In the eyes, melanosomes are found only in the melanocytes in the iris. The type and the amount of melanosomes and melanin within the melanocytes of the iris determine eye colour [29]. Large number of melanosomes and higher amounts of melanin are characteristics of the dominant brown irises; in contrast, blue irises have fewer melanosomes and no or low amounts of melanin. Notably, humans carry no blue pigment and blue eyes are because of the Tyndall effect, a phenomenon due to which the sky is blue [30]. Tyndall effect is the scattering of light by particles in a liquid suspension. The light entering the eye is scattered and reflected back into the space and as a result of the Tyndall effect, a blue hue is generated [30]. Hence the blue eye colour is dependent on the quality and quantity of available in the surroundings. Intermediate amounts of melanin and varying proportions of eu- versus pheomelanin are responsible for the other eye colours such as green, hazel or various colour shades [28].

Eye colour is a highly polymorphic phenotype restricted to the people of (at least partial) European descent and those of neighbouring regions. Brown eye colour phenotype is considered the ancestral human state following the Out of Africa hypothesis of human origin and dispersal [31,32] and is seen everywhere in the world [5,33,34]. Mutations in pigmentation genes causing light eye colour are assumed to have occurred in European pre-history at a currently unknown time, at least well before 5000 years ago [20]. Given the absence of evidence on environmental adaptation for eye colour (in contrast to skin colour, see below) it is assumed European eye colour variation evolved via sexual selection via mate choice preference of the initially rare light eye colour types [18,28].

Intensive research has been done to understand the genetics behind human eye colour via candidate gene studies, linkage analysis and GWAS [18, 32,35-38] followed by predictive studies [39]. Several SNPs within different genes have been identified that contribute to human eye colour variation. A single SNP, rs12913832 in the *HERC2* gene on chromosome 15 provides a large proportion of the blue and brown eye colour predictive information

[36,39,40]. However, one SNP is not sufficient to provide accurate information on eye colour categories, typically categorized as blue, brown, and intermediate (including everything not classified as blue or brown). Other SNPs in several other genes, such as *SLC24A4, SLC45A2 (MATP), TYRP1, TYR* and *IRF4*, have been identified to contribute to the eye colour variation [18,35,37,38,41]. Based on these findings, several prediction models have been proposed for DNA-based eye colour prediction [5,21,24,25,42]. However, the first DNA-based eye colour prediction tool for forensic applications was the highly sensitive and robust IrisPlex system, comprising of six most highly informative SNPs: rs12913832 (*HERC2*), rs1800407 (*OCA2*), rs12896399 (*SLC24A4*), rs16891982 (*SLC45A2*), rs1393350 (*TYR*) and rs12203592 (*IRF4*) in a single multiplex genotyping SNaPshot® assay [5]. Figure 3 (taken from Walsh et al., 2011) depicts the genetic variation in eye colour based on the genotypes of these six SNPs. In addition to the multiplex genotyping assay, the IrisPlex system involves a statistical prediction model to estimate the categorical eye colour probabilities of brown, blue and intermediate colours initially introduced via an interactive and user-friendly Excel sheet [5,21].

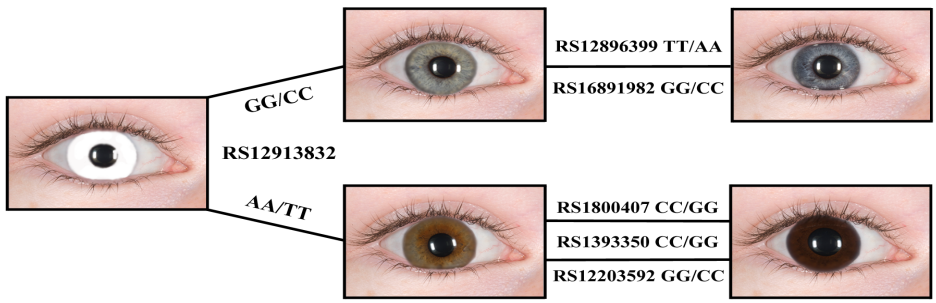


Figure 3: Hypothesised scenario for genetic determination of brown and blue eye colours showing the impact of the most influential SNP genotypes from the 6-SNP model.
(Taken from Walsh et al., 2011 [5])

Whenever a new method is introduced into the forensic community, the standardisations of the method and quality assurance are of key importance for their applicability in routine casework in laboratories worldwide. The Scientific Working Groups on DNA analysis (SWGDAM) developmental validation guidelines emphasize on the accuracy, precision, reproducibility and concordance of a novel procedure or a technique prior to its routine implementation in all the forensic laboratories. The IrisPlex system had

been successfully validated complying with the strict guidelines of SWGDAM [21]. To further test the reliability of the prediction model, the IrisPlex system was evaluated on a broad dataset of >3,800 DNA samples from seven sites in seven countries across Europe, as part of the European Eye study (EUREYE) [22]. Hence, it has been proved that the IrisPlex system can accurately predict blue and brown eye colours with an accuracy of >94% [22], whereas its potential to predict non-blue and non-brown eye colour is much lower. However, there is a necessity to test the ease of implementation of the IrisPlex system and assess the performance of the system across forensic laboratories with varying levels of experience, which was done as part of this thesis work. Notably, at this moment genetic understanding of eye colour and consequently DNA-based eye colour prediction is mostly restricted to eye colour categories but not yet to the full spectrum of human eye colouration. A GWAS study on continuous (not categorical) eye colour variation, which found three new genes [41] not previously identified when using categorical eye colour, demonstrates that many more genes wait for being identified to genetically understand the complete spectrum of human eye colour variation.

Hair Colour

Melanin that gives rise to human hair colour is produced in the follicular melanocytes in the hair bulb, from where they are transferred to the surrounding keratinocytes that finally differentiate and migrate forming the pigmented hair shafts [43]. The varying shades of hair colour can be described on a continuous spectrum from red, blond, brown to black colours. High levels of pheomelanin characterize red hair colour, while the highest ratios of eumelanin to pheomelanin are indicative of black hair colour. Blond and light-brown colours are a result of intermediate ratios of eumelanin and pheomelanin [44].

Variation in human hair colour, similar to the iris colour, has been geographically restricted to the people of (at least partial) European descent and surrounding regions of Middle East and parts of Western Asia [33], which is assumed to have arisen via initial mutations and subsequent sexual selection of light hair colour [34]. The only known exception is blond hair in New Guinea and nearby islands such as the Solomons, which has been shown to have evolved independently from blond hair in Europeans as the causing mutation is not found in blond Europeans, albeit different mutations in the same gene [45]. Extensive studies on the genes associated with hair colour variation over the last few years

concluded that *MC1R* gene on chromosome 16 is associated with human hair pigmentation [18,38,46-48], representing the first gene identified to be involved in human hair colour. *MC1R* gene has high association with red hair colour and lower association with other non-red hair colours. Recent scientific research has identified several other genes (*OCA2, HERC2, SLC45A2, SLC24A4, TYRP1, ASIP, TYR* and *KITLG*) associated with non-red hair colours [16,18,38,49]. It currently is suggested that while red hair is caused by a single gene (*MC1R*), all other hair colours are polygenic traits determined by various genes of which not all are known as of yet. SNPs within the *HERC2* gene, like with eye colour, is the most contributing DNA marker to predict non-red hair colour. SNPs within the different genes have been highly associated with the different hair colour phenotypes - *HERC2* SNPs with black and dark brown hair colour phenotype; *OCA2* and *TYR* SNPs with brown hair. It has been found that the SNPs within *OCA2* are most significantly associated with brown hair, *SLC45A2* and *IRF4* SNPs with black hair, *SLC24A4* SNPs with blond and dark blond hair and ASIP SNPs with red, blond-red and dark blond hair colours [16,18,38,49].

Based on the previous findings, and similar to IrisPlex, a DNA tool was developed to predict four categories of hair colour (blond, brown, red and black), shades of the hair colour (light and dark), together with categorical eye colour [5,21,49]. The HIrisPlex system comprises of 24 SNPs from 11 genes - Y152OCH, N29insA, rs1805006, rs11547464, rs1805007, rs1805008, rs1805009, rs1805005, rs2228479, rs1110400 and rs885479 from the *MC1R*, rs1042602 (*TYR*), rs4959270 (*EXOC2*), rs28777 (*SLC45A2 (MATP*)), rs683 (*TYRP1*), rs2402130 (*SLC24A4*), rs12821256 (*KITLG*), rs2378249 (*PIGU/ASIP*), rs12913832 (*HERC2*), rs1800407 (*OCA2*), rs16891982 (*SLC45A2 (MATP*)) and rs12203592 (I*RF4*). Of these 24 variants, 6 of these IrisPlex markers, rs12913832, rs1800407, rs12896399, rs16891982, rs1393350 and rs12203592 are used for eye colour prediction, of which all but two (rs12896399 and rs1393350) are used for hair colour prediction. [5,21,22,39]. An interactive and user-friendly macro based excel tool is initially introduced for obtaining the prediction probabilities of the four hair colour categories (blond, brown, red and black), hair colour shades (light and dark) and the three eye colour categories (brown, black and intermediate) [5,21]. Additionally, a prediction guide is also made available to aid in interpretation of the individual hair colour and the shades accurately [5,21].

Figure 4: Hypothesised scenario of the effect of each HIrisPlex SNPs minor allele input on the model for hair colour prediction as a homozygous genotype (the minor allele input is 2). The highest effect in terms of probability for a certain hair colour category is noted and the SNP is named near that category within the figure *(Taken from Walsh et al., 2013 [23]).*

Figure 4 depicts the different SNPs contributing to the different hair colour categories. To aid in police investigations for predicting the eye and hair colour of unknown perpetrators and missing persons from DNA, the HIrisPlex system must be implemented in accredited forensic laboratories. Hence, it is necessary to perform developmental validation of the HIrisPlex system adhering to the strict guidelines of SWGDAM, which was done as part of this thesis work.

Skin Colour

Human skin colour is a complex trait and the skin colour varies tremendously over a broad range from very pale to very dark. Skin colour variation is determined by the amount and type of melanin produced in the melanocytes, and the shape and the distribution of the melanocytes in the basal layer of the epidermis [14,15,44]. Darker skin colours have higher levels of melanin enriched with eumelanin and single units of larger and more pigmented melanosomes [50]. Lighter skin has higher levels of light brown eumelanin and yellow/red

pheomelanin with smaller and less pigmented melanosomes, packaged in groups [51].

Variation in skin colour is globally distributed, in contrast to the eye and hair colour variation. It is highly correlated to geographic patterns, explained by natural selection via adaptations to environmental changes, such as latitude varying levels of ultraviolet radiations (UVR) [52] as a result of human migration out of Africa into northern areas. Skin colour is darker in areas near the equator and becomes lighter with increasing distance from the equator. UVR levels are high near the equator, and darker skin is needed to protect against the negative effects of sun radiations such as UV-induced DNA damage. In areas further from the equator, due to low levels of UVR lighter skin is beneficial to avoid overprotection from the sun and to allow the needed levels of vitamin D to be synthesised from pro-vitamin D for which UV radiation is needed to become more independent from vitamin D rich diet [52].

Compared to eye and hair colour, the knowledge about genetics underlying the skin colour variation is currently more limited. Skin colour is largely varying between populations and lesser within populations, which causes problems for GWAS that can only be carried out in genetically homogeneous populations to limited false positive discoveries. Therefore, GWAS on skin colour were conducted within Europeans [18,38] or within Asians [53], but both groups harbour less skin colour variation, respectively, than is observed between continental groups. Therefore, we currently understand less of the genetic basis of skin colour than of eye and hair colour. Several SNPs associated with eye and hair colour variation also appeared to be associated with the variation in skin colour. A 7-SNP set was described for predicting the non-light and non-dark skin colour [25]. The SNP rs6119471 (*ASIP*) was highly variable in dark skinned populations, whereas, the remaining six SNPs rs1426654 (*SLC24A5*), rs12913832 (*HERC2*), rs16891982 (*SLC45A2*), rs12203592 (*IRF4*), rs1545397 (*OCA2*) and rs885479 (*MC1R*) were observed to be highly variable in populations with light skin colour [25]. A comprehensive set with a panel of 59 SNPs previously associated with eye, hair and skin colour was published in 2014 [54]. Out of these 59 SNPs, 29 were identified to be strongly associated with the skin colour variation and were able to differentiate individuals with white skin colour from those with intermediate/black skin colour. Amongst the 29 SNPs, 10 best skin colour predictive SNPs (rs10777129, rs1426654, rs16891982, rs13289, rs3829241, rs6058017, rs6119471, rs2402130, rs1408799 and rs1448484) were identified with AUC values of 0.999 for white, 0.803 for intermediate and 0.966 for black skin colour [54]. However, no dedicated genotyping assay had been developed from this knowledge.

Similar to the IrisPlex and HIrisPlex, a forensically validated DNA intelligence-based tool for predicting the categorical skin colour would be beneficial to the forensic community, which was developed as part of this thesis work.

Prediction of maternal bio-geographical ancestry and identification of maternal lineages from mtDNA SNPs

Bio-geographic ancestry can be explained as associating the ancestral origin of a person with a geographic location [55]. Ascertaining the bio-geographical origin of an unknown suspect has immense value in forensic applications. Inference of ancestry information from DNA is gaining importance in genealogy and the recent field of personal ancestry testing, in addition to population history studies [56]. Different patterns of genetic variation across diverse population groups have been explained by population studies. Several DNA polymorphisms, especially SNPs have been divulged for discerning the distinct global populations for ancestry prediction [57–59]. The potentiality of such ancestry informative markers to deduce the ancestry affiliations in discrete and admixed populations has already been illustrated [60–62].

The intriguing properties of mitochondrial DNA (mtDNA) have made its application substantial in forensics for maternal ancestry prediction and lineage identification, which is the major focus of this thesis. MtDNA is useful in forensic investigations of remains of a missing person recovered from a mass disaster. These remains are often highly fragmented or are recovered in tiny amounts, such as a single tooth or a sliver of a bone [63]. Biological material from known, maternal relatives, even quite distant, can be used as a reference for direct comparison to the recovered remains. Since the introduction of mtDNA analysis in the forensic field, several legal cases involving different evidentiary materials and population studies have been conducted by use of mtDNA [64–67].

MtDNA is a small, circular and double stranded molecule. MtDNA is present in higher abundance in the human genome as compared to the nuclear DNA. The circular shape of the mtDNA provides higher resistance against degradation, thereby making it easier to recover the mtDNA in aged specimens, relative to nuclear DNA. Hence, mtDNA is preferred in situations when the nuclear DNA is degraded in the available material found at the crime scene, such as hairs, bones, teeth and skeletal remains [68,69]. The higher mutation rate of mtDNA as compared to nuclear DNA and the resulting significant sequence variability

between maternally unrelated individuals has bolstered its value as an imperative tool in forensic investigations. Further, in forensic investigations the maternal inheritance pattern of mtDNA and the lack of recombination permitted the use of maternal relatives as references for unknown samples in the absence of direct references.

The mitochondrial genome (mtgenome), with ~16570 bp, is divided into a large coding region and ~1100 bp non-coding control region (also called D-loop or the displacement loop) (Figure 5). Within the control region are three hypervariable regions (HVS) in the (HVSI: 16024-16365, HVSII: 73-340 and HVSIII: 438-574) and these are mainly analysed in several areas of research, such as population genetics, evolutionary studies, medical genetics and forensic analysis. Complete mtDNA was the first significant part of the whole human genome to be sequenced and published in 1981 [70] and was termed as the Cambridge reference sequence (CRS). A revised version of CRS was reported in 1999 as rCRS [71]. Routine mtDNA analysis involves detection of SNP variation in the hypervariable regions of approximately 610 bp by Sanger sequencing and the resulting sequences are evaluated for deviations from the rCRS. MtDNA harbours unique set of mutations that accumulate over time and they carry information about maternal ancestry and one's haplotype.

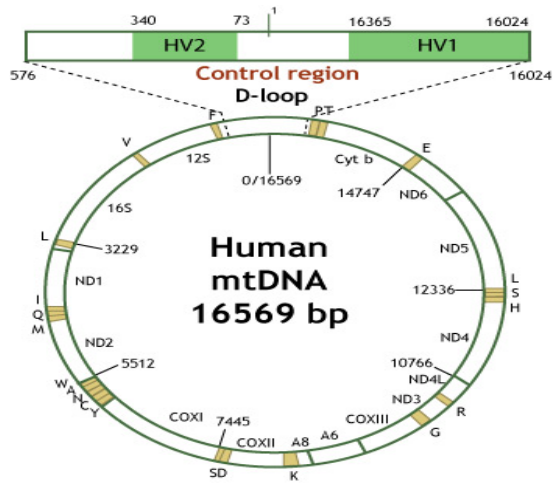


Figure 5: Mitochondrial DNA *(Taken from http://projects.nfstc.org/pdi/Subject09/pdi_s09_m01_01_b.htm)*

Intensive research into the worldwide distribution of mtDNA genetic variation yielded a large number of ancestry informative SNPs within the mitochondrial genome, linking specific clades of the mtDNA phylogeny with certain geographic areas [72,73]. MtDNA was used for biogeographical ancestry inference from five different ethnic groups as early as 1994 [74]. All the individuals with the same sequence variations were grouped into haplogroups that were characterized by mutations at particular nucleotide positions [75–77]. Haplogroups are often geographically related and the most common HVI/HVII type (the sub-haplogroup H1) is shared by 7% of the Caucasian population [72, 78-82]. The lack of recombination and the reduced effective population size empowered mtDNA haplogroups to become restricted to specific geographic areas [83], such as haplogroup L to Africa, V to Europe and the Middle East, or P and Q to Oceania [6,83].

MtDNA is ten times more prone to mutation than nuclear DNA, which enhances the degree of variation and thus its application for maternal ancestry inference. Traditionally, Sanger sequencing has been the method of choice to sequence the entire control region as the hypervariable regions exhibit the highest degree of sequence divergence in the mitochondrial genome [84–88]. Despite the fact that the mtDNA control region sequences are informative for assigning several mtDNA haplogroups, not all haplogroups present suitable diagnostic variants in HVS region that allow an unequivocal assignment. Thus to improve the discrimination power, it is also imperative to consider the SNPs in the comparatively larger and more stable coding region. A vast number of such SNPs, which are used to establish a haplogroup status, have been published [79, 89–91]. Several multiplex SNP assays have been optimised and applied in forensic laboratories [72,91,92]. Using the available knowledge, a global mtDNA SNP kit, divided into three hierarchically organized multiplex assays, was developed to target mitochondrial DNA SNPs allowing to identify maternal haplogroups that are informative for maternal bio-geographic ancestry. These assays targeted 36 SNP variants present in the coding region of the mtDNA that together differentiate 43 matrilineal haplo-/paragroups typical for Africa, Western Eurasia, Eastern Eurasia and Native America [55]. Additionally, three more hierarchically organized SNP multiplexes for genotyping 26 SNPs defining all major mtDNA haplogroups for Oceania (including Australia, Near Oceania and Remote Oceania) were developed [93]. Figures 6a and 6b depict the phylogenetic tree of the used mtDNA SNPs and the major geographic regions of the detected haplogroups from van Oven et al. [55] and Ballantyne et al. [93]. For routine implementation into the forensic

community, these multiplex assays must undergo the developmental validation following the SWGDAM guidelines, which was done as part of this thesis work. It is to be noted that mtDNA represents only the maternal ancestry of an individual that is only a part of an individual's overall genetic makeup.

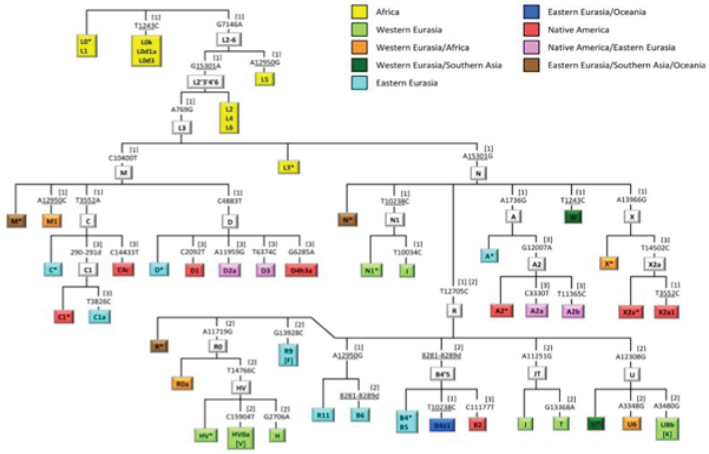


Figure 6a: Overall phylogenetic scheme of targeted mtSNPs with geographic haplogroup classification. The combined use of the three multiplex assays allows any person's mtDNA to be classified into one of the colour-labelled haplogroups. Colours correspond to the geographic origin of the haplogroups as indicated. SNP position numbers are relative to the revised Cambridge Reference Sequence (rCRS). Deletion mutations are denoted by the suffix 'd'. Recurrent SNPs are underlined. The numbers 1, 2 or 3 in square brackets shown for each SNP refer to the respective multiplex assay in which the SNP is included. Note: haplogroups F, K and V are encompassed within R9, U8b and HV0a, respectively, as indicated because this does not follow logically from the nomenclature *(Taken from van Oven et al., 2011 [55]).*

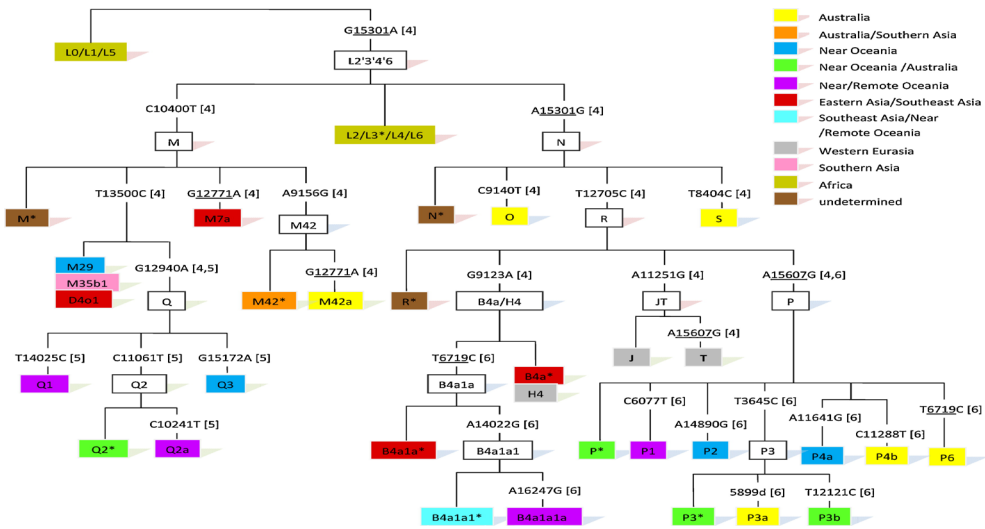


Figure 6b: MtDNA haplogroup classification with their diagnostic SNPs and their geographic assignment as provided by the three Oceanian multiplexes. The combined use of the three multiplexes allows the assignment of an individual to one of the colour-coded haplogroups, with the colours corresponding to the assumed geographic region of origin as indicated. SNP positions are relative to the rCRS, with the number in square brackets referring to the multiplex required for SNP typing. Recurrent SNPs are underlined *(Taken from Ballantyne et al., 2012 [93]).*

In addition to maternal ancestry with mtDNA, paternal ancestry with Y and bi-parental ancestry with autosomal ancestry informative markers are necessary to improve the level of detail of ancestry inference of an unknown person from DNA and for inferring bio-geographic ancestry of admixed individuals. Autosomal SNPs showing major allelic differences between populations have been discovered and are increasingly used in discerning continental categories and subpopulations [59,94,95]. Being bi-parentally inherited loci and covering a greater proportion of genome history, more comprehensive information about an individual's ancestry is provided by the autosomal markers as compared to the uni-parentally inherited markers on the non-recombining parts of the mtgenome and in Y-chromosome. Autosomal SNPs are increasingly used due to their capability to apprehend the genetic diversity contributed by all the ancestors, providing evidence about the recent genetic events [56]. SNPs on the non-recombining part of the Y-chromosome are extensively used for providing information about the paternal lineage in different fields such as anthropology, genealogy and forensics [6,96,97].

Together with the prediction of bio-geographic ancestry, mtDNA is used for identifying the maternal lineage an unknown individual belongs to. The analysis of mtDNA results in a lower power of discrimination, compared to the analysis of multiple nuclear markers, due to the lack of recombination so that with mtDNA groups of maternally related individuals are identified (i.e. maternal lineages) but no single individuals. Informative SNPs on the mtDNA genome are sequenced and compared, for a match, to a reference database that contains several haplotypes identified in specific populations [98]. However, lineage-based analyses provide information about only one line of descent out of the many that contributed to the current genetic profile of an individual [98]. The accuracy of lineage-based analyses mainly depends on the size and quality of the reference database. Further, the level of geographic resolution depends on the haplotype of the sample and the populations included in the reference database. With the advancements in lineage-based mtDNA research and its increasing use in missing persons, identification, DVI and also in exclusion of individuals as contributors of forensic samples, it is essential to standardise the forensic mtDNA analysis. The DNA Commission of the International Society for Forensic Genetics created guidelines for mtDNA typing in a forensic science laboratory. They detailed the standard amplification and sequencing methodology, the nomenclature of the mtDNA variants, criteria for inclusion or exclusion in case of mismatches between the analysed samples, as well as general

principles of heteroplasmy or indel interpretation [99-101]. Whenever there is a match between an evidence sample and a reference sample, it is an indication that the profile might originate from the same individual. However, in such a case it is important to estimate the frequency of occurrence or the rarity of a certain profile in a population [86,87], before final conclusion. Some freely available public databases, such as mtDB – Human mitochondrial genome database [102], provide information on mitochondrial haplotypes and help in estimating the frequency of occurrence of mtDNA sequences.

One of the mtDNA databases, popular among the forensic community due to the high quality of the stored mtDNA sequences following strict database curating, is the European DNA Profiling Group (EDNAP) MtDNA population database, in short EMPOP (http://www.empop.org). EMPOP is a collaborative effort of several DNA laboratories under the lead of the DNA Laboratory of the Institute of Legal Medicine (GMI) in Innsbruck, Austria. All the data or the mtDNA sequences in the EMPOP are collected with strict guidelines adhering to the forensic requirements [85,103] and majority of them are generated using Sanger sequencing of the control region. Previously, heavy criticism has been raised against some mtDNA databases for high error rates [104-106], clerical errors and mistakes in analysis [68,69]. Such errors introduce phantom polymorphisms that interfere with estimating the frequency of a particular mtDNA haplotype. In order to solve such issues, phylogenetic analysis may be applied based on the available knowledge about evolutionary patterns in human mtgenome [69,107]. All the sequences that are included in the EMPOP database are critically verified for any erroneous data [102,108]. For instance, the EMPOP database employs quasi-median networks to test the quality of the mtDNA data. The quasi-median network graphically represents the genetic structure of the lineages in a data set [102,108].

The size of the database and the coverage, i.e. the distribution and geographic representation of the samples, are crucial for estimating the rarity of the mtDNA profile and calculating the random match probability. In order to attain a good representation of the mtDNA diversity large databases, including individuals from different geographic regions, are necessary. There are chances of wrong statistical calculations in case of small haplotype coverage in a database that represents small fraction of population diversity. This becomes relevant in scenarios when mtDNA haplotype of a suspect is quite common in the pertinent geographical region, but lacks its representation in the database [107], so rare profiles are not represented at all. Hence, there is a need for consistent expansion of the EMPOP database,

which was achieved via work carried out as part of this thesis.

Efforts are also needed to generate full mtgenome data with high quality and meeting the standards necessary for application to forensic cases. Genotyping of the coding region and control region separately is time-consuming and highly laborious. Full potential of the mtDNA data can be exploited once high quality mtgenome data is available on a large scale. With the introduction of next generation sequencing (NGS) or massively parallel sequencing (MPS) technology, whole mtgenome sequencing has become viable. The advent of NGS or MPS technologies has revolutionized genomic research, and a striking impact has been made on several fields of life sciences, including forensics. MPS has the potential to further decipher the complexity of human genome by unravelling the rare genetic variations that could not be revealed with current technology so far [109]. It is now possible to produce huge volumes of genomic data at moderately low costs by sequencing amplified DNA templates massively in a parallel fashion. Sanger sequencing allows much longer sequence reads than most MPS technologies; however, the parallel nature of MPS means that longer reads can be constructed from many contiguous short reads [110]. MPS exhibits the ability to alleviate the challenges faced by forensic laboratories in terms of degraded samples. Also, multiplexing becomes fairly easy with MPS approaches and the technology has made possible the sequencing of STRs [111], Y-STRs [112], SNPs [113] and mtgenome [114–122]. Recently, a MPS tool using short amplicons of average size of 203 bp for simultaneous analysis of >530 Y-chromosome SNPs, that target >430 worldwide Y-haplogroups for ultrahigh resolution of paternal lineage and ancestry inference, was introduced [123]. A similar tool for the maternal lineage and ancestry inference from as many as possible mtDNA SNPs, at best the entire mtgenome, is needed, which was achieved as part of this thesis work.

Genotyping methods used in this thesis work

SNaPshot Single Base Extension Minisequencing Assay

The SNaPshot (Life Technologies) single base extension (SBE) minisequencing assay is the method of choice in several forensic laboratories for high throughput SNP genotyping for developing identification and DNA intelligence based assays [5,29,63,84,113–115]. The SNaPshot assay is highly reliable, relatively cheap, highly robust and sensitive with high multiplexing capabilities [116–120]. The SNaPshot genotyping strategy is based on the single base extension of unlabelled primers with labelled dideoxyribonucleotide

triphosphates (ddNTPs) (Figure 7).

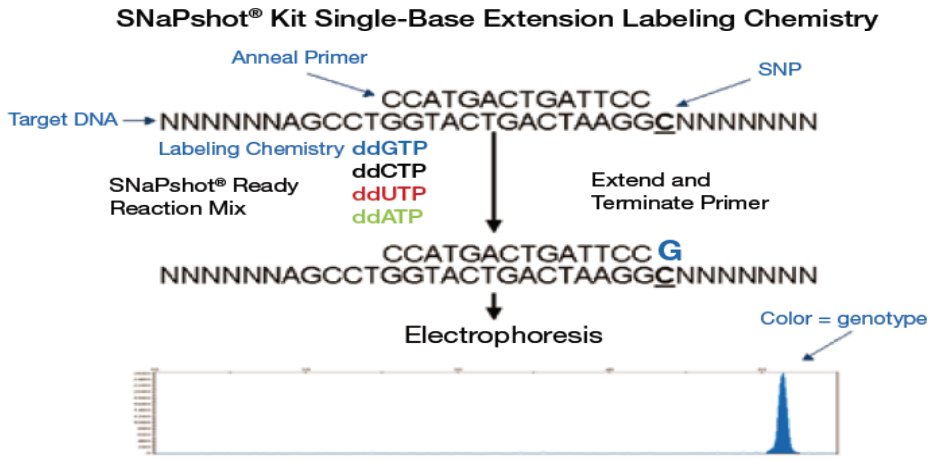


Figure 7: SNaPshot labeling chemistry relies on single-base extension and termination. The SNaPshot Multiplex Kit uses a single-tube reaction to interrogate SNPs at known locations. The chemistry is based on the dideoxy single-base extension of an unlabeled oligonucleotide primer (or primers). Each primer binds to a complementary template in the presence of fluorescently labeled ddNTPs and DNA polymerase. The polymerase extends the primer by one nucleotide, adding a single ddNTP to its 3′ end. The fluorescence colour readout reports which base was added *(Taken from http://lifesequencing.blogspot.nl/2012/09/aplicacoes-da-tecnica-de-snapshot.html).*

The first step is the PCR amplification of the target region, followed by purification of the PCR product by removing PCR primers and remaining deoxyribonucleotide triphosphates (dNTPs) using Exonuclease (Exo) and shrimp alkaline phosphatase (SAP) [121]. Then, these purified PCR products are added to SNaPshot® reagents, which results in annealing of a minisequencing oligonucleotide primer (extension primer) directly adjacent to the target SNP position. Then fluorescently labelled dideoxynucleotide triphosphates (ddNTPs), complementary to the target SNP site is added to the minisequencing primer by DNA polymerase. Non-binding tails are added to extension primers in order to spread out resulting extension products for easy visualisation using capillary electrophoresis. These single base extension primers bind adjacent to the target SNP site and the extension is terminated due to the absence of a 3'-hydroxyl group required for further chain extension by DNA polymerase. Finally, SNaPshot products are purified with SAP by removing unincorporated fluorescently labelled ddNTPs; then separated using capillary electrophoresis and the SNP extension product is visualized. Genotyping softwares such as GeneMapper (Applied Biosystems) are used for data analysis and one of its applications is SNP genotyping for the data generated

with the capillary electrophoresis. The raw data is processed using algorithms and the alleles (A, G, C and T) are assigned to the peaks accordingly.

Sanger Sequencing with dye terminators

DNA sequencing unravels the order of the four nucleotides in a DNA strand. The first method for sequencing DNA was discovered by Fred Sanger in 1975 [122]. It is a robust method for finding de novo mutations and the read length for each DNA template is about 1-1.2 kb [99]. This was the technology that was used to sequence the DNA from the Human Genome Project [123]. During Sanger sequencing, PCR primers, unlabelled dNTPs and fluorescently labelled ddNTPs are mixed in a single reaction. The key to Sanger sequencing is the use of fluorescently labelled ddNTPs that, similar to dNTPs, incorporate into the chain by forming a phosphodiester bond at the 5'end. DNA polymerase makes copies of the single-stranded DNA templates by adding nucleotides (dNTPs or ddNTPs) to the 3' end of the growing chain. When the ddNTPs are incorporated at the 3' end, chain elongation is terminated selectively at A, C, G, or T, as ddNTPs lack a hydroxyl group, which they need for binding with the next nucleotide. At the end of the reaction, several DNA fragments of different lengths (i.e. differing in size by one base) are generated. These fragments are then separated by length by capillary electrophoresis and thus revealing the order of the bases in the DNA sequence.

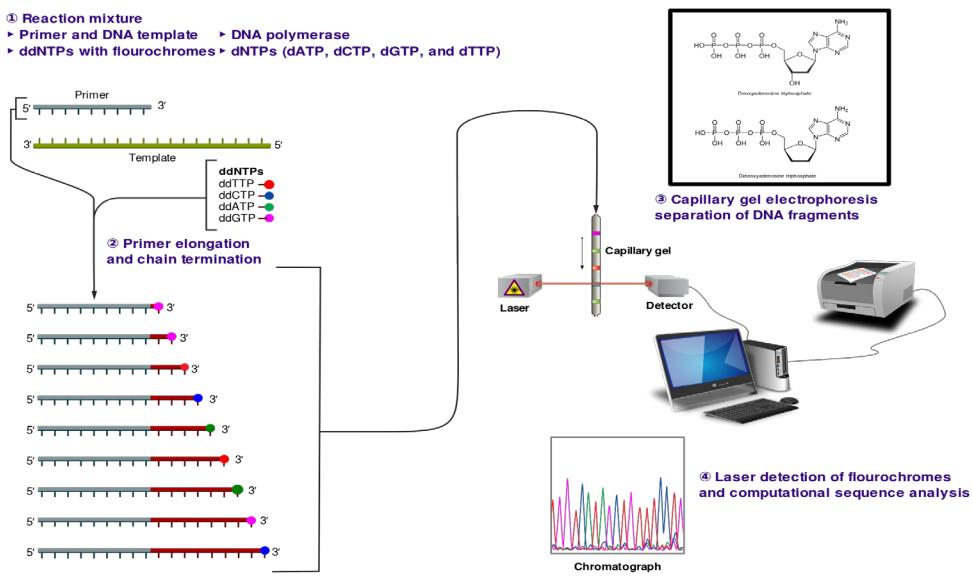


Figure 8: The Sanger method for DNA sequencing. *(Taken from https://en.wikipedia.org/wiki/Sanger_sequencing).*

Initially Sanger sequencing was carried out in four separate tubes, each tube containing appropriate amount of one of the ddNTPs that were not fluorescently labelled. However, in In the Big Dye Terminator chemistry, each of the four ddNTPs is fluorescently labelled with four different dyes. The sequencing is done in one reaction and is hence less time consuming and less labour intensive. Peaks of different colours are created once the emission of different fluorescent dyes is measured as the fragments pass the detector.

Massively Parallel Sequencing using the Ion Torrent Personal Genome Machine

Massively parallel sequencing (MPS) provides a myriad of opportunities to augment the research field into sectors not yet explored, especially in forensics that is often challenged with low copy number and/or low quality DNA samples. MPS increases the amount of information that can be obtained from a biological sample. With the advantages in terms of speed, efficiency and reliability, MPS has empowered several novel biological applications and scientific achievements [98]. MPS technology exhibits high throughput capacity, where millions or billions of sequences can be generated in parallel, using lower DNA template input, as compared to Sanger sequencing. Longer reads can be generated from aligning numerous contiguous short reads. Several MPS platforms, such as Roche 454, Illumina MiSeq, Illumina HiSeq, Life Technologies Ion Proton, are commercially available. For the purpose of this thesis, I will describe in brief, the MPS approach using the Life Technologies Personal Genome Machine (PGM). Unlike other sequencing platforms such as Roche 454 and Illumina that use optical signals, PGM uses the fact that whenever a dNTP is incorporated to a DNA template chain, a proton (H+) is released. Hence it is also referred to as semiconductor sequencing that detects the hydrogen ions that are released during the polymerization of DNA. Depending on whether a dNTP is added or not, a change in pH is detected. No voltage will be generated and no base will be called if the correct dNTP is not incorporated when the chip is sequentially flooded with dNTPs one after the other. A double voltage is detected when two identical dNTPs are incorporated [124].

Figure 8 illustrates the workflow of Ion Torrent PGM sequencing. The sequencing is carried out on a semiconductor chip. The chip has millions of wells and in order to sequence in massively parallel fashion, each well harbours a different DNA template. Beneath the wells is an ion sensitive layer and below it is a proprietary Ion sensor. The first step in the sequencing process is the construction of libraries. The sample of DNA is fragmented

into millions of fragments and ion sequence adapters are ligated to the 5' end of the DNA. The library fragments are then attached to beads and clonally amplified onto the proprietary Ion Sphere particles using emulsion PCR. Beads coated with complementary primers are mixed with dilute aqueous solution containing the fragments to be sequenced along with the necessary PCR. This solution is then mixed with oil forming an emulsion of micro droplets. The concentration of the beads and fragments is kept low enough such that each micro droplet contains only one of each fragment. Clonal amplification of each fragment is then performed within the micro droplets. The Ion Sphere particles coated with template are applied to the Ion chip. The Ion Sphere particles are then deposited in the chip wells by a short centrifugation step. The chip is placed on the PGM and the PGM touchscreen guides the user to set up the sequencing run. Once data is generated on the Ion PGM sequencer, it is automatically transferred to the required Torrent Server. The data is run through signal processing and base calling algorithms that produce the DNA sequences associated with individual reads. The data can then be downloaded into industry-standard data formats like SFF, FASTQ, or SAM/ BAM. The Ion Torrent data can then be imported into any number of NGS data analysis solutions.

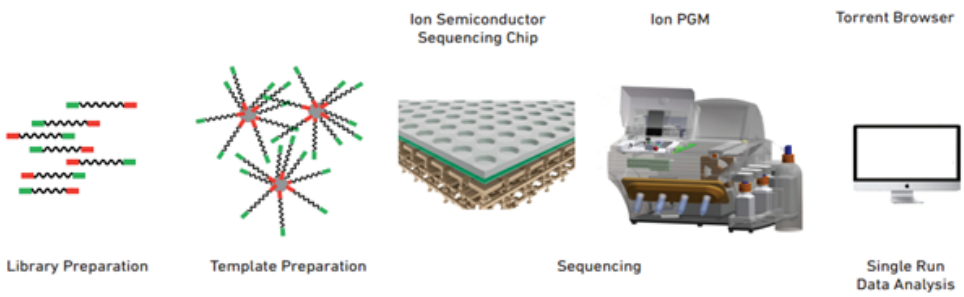


Figure 9: Schematic representation of the Ion Torrent sequencing workflow. A sequencing library is produced by generating DNA fragments flanked by the Ion Torrent sequencing adapters. These fragments are clonally amplified on the Ion Sphere particles by emulsion PCR. The Ion Sphere particles with the amplified template are then applied to the Ion Torrent chip and the chip is placed on the Ion PGM. The sequencing run is set up on the Ion PGM. Sequencing results are provided in standard file formats *(Taken from http://www.wright.edu/~oleg.paliy/NGS.html)*

Aims of the thesis

Traditionally, routine forensic casework is based on comparative grounds. DNA profiles obtained from crime-scenes are compared with those of potential suspects or DNA profiles deposited in forensic DNA databases. The principal limitation of such comparative approach is that trace donors unknown to the investigators with their DNA profiles cannot be identified. The recent advancements in Forensic DNA Phenotyping can provide investigative leads to help find previously unknown individuals by inferring externally visible characteristics and biogeographic ancestry prediction without comparative DNA testing. Some predictive DNA tests for appearance and ancestry were previously developed by our group, but their forensic validation, as prerequisite for casework application, was mostly missing thus far. For other appearance traits the genetic knowledge was previously not advanced enough to develop predictive DNA tests. Moreover, the genotyping technology capable of tolerating quality and quantity issues of trace DNA and at the same time allowing the parallel analysis of large number of DNA markers was previously not available, which limited previously developed forensic ancestry DNA tests.

The major aims of this thesis therefore are to:

- forensically validate DNA test systems for predicting appearance and ancestry that were recently developed by our group i.e., IrisPlex for eye color prediction, HIrisPlex for hair and eye color prediction, and a set of mtDNA multiplexes for maternal ancestry prediction;
- develop and forensically validate a new DNA test system for skin color prediction in combination with eye and hair color prediction from trace DNA;
- develop and forensically validate a new DNA test system for complete mitogenome analysis providing maximum-resolution maternal lineage and maternal ancestry inference from trace DNA.

Scope of this thesis

Traditionally, routine forensic casework is based on comparative grounds. DNA profiles obtained from crime-scenes are compared with those of potential suspects or DNA profiles deposited in forensic DNA databases. The principal limitation of such comparative approach is that persons unknown to the investigators with their DNA profiles cannot be identified. The recent advancements in Forensic DNA Phenotyping can provide investigative leads to help find previously unknown individuals by inferring externally visible characteristics and biogeographic ancestry prediction without comparative DNA testing. In this thesis, I describe the development, forensic developmental validation and end-user validation of FDP tools to predict human pigmentation traits and maternal bio-geographic ancestry in the forensic context. Moreover and beyond maternal ancestry testing, I present significant improvements in forensic mtDNA analysis for maternal lineage identification.

The IrisPlex assay for predicting the eye colour represents the first FDP system that had been developed and successfully validated under the strict SWGDAM guidelines. In **Chapter 2**, I describe an end-user validation of the IrisPlex system via a collaborative, inter-laboratory exercise, to test the reliability, consistency and the ease of implementation of the IrisPlex system across 21 forensic laboratories of the European DNA Profiling Group (EDNAP) with varying levels of experience, which was led by the author.

In **Chapter 3**, I focus on the HIrisPlex assay that was previously designed and developed to simultaneously predict hair and eye colour from DNA. **Chapter 3.1** details the developmental validation study of the HIrisPlex assay following the SWGDAM guidelines to demonstrate its robustness and reliability for forensic applications. **Chapter 3.2** illustrates the application of the validated HIrisPlex assay in DNA samples from victims of World War II and hence substantiates the suitability of HIrisPlex assay as a forensically-relevant, investigative tool particularly for DNA analysis of human remains.

In **Chapter 4**, I demonstrate the design and development of the new HIrisPlex-S system for predicting the skin, hair and eye colour from DNA as well as the forensic developmental validation of the newly developed HIrisPlex-S genotyping assay according to the SWGDAM guidelines as prerequisite for its use in forensic laboratories worldwide.

In **Chapter 5**, I focus on the other relevant aspect of DNA intelligence - the prediction of maternal bio-geographic ancestry using mitochondrial DNA (mtDNA). In **Chapter**

5.1, I evaluated the hierarchical system of five multiplex assays targeting 62 ancestry-informative mitochondrial SNPs for inferring matrilineal biogeographic ancestry origin on the continental level. In **Chapter 5.2**, I describe the generation of complete mtDNA control region sequences from randomly selected Dutch individuals sampled across the Netherlands, as a start of a Dutch mtDNA reference database for forensic use and for use in the worldwide EMPOP database. In **Chapter 5.3**, I highlight the establishment of an MPS tiling approach for simultaneous whole mtgenome sequencing with the Ion Torrent PGM and demonstrate its application on degraded DNA samples often confronted with in forensic DNA analysis.

Finally, in **Chapter 6**, I provide a general discussion of the findings described in **Chapters 2-4**, and give a perspective on future research.

References

[1] Jeffreys AJ, Wilson V, Thein SL. Hypervariable 'minisatellite' regions in human DNA. Nature. 1985 Mar ;314(6006):67-73.

[2] Budowle B, van Daal A. Forensically relevant SNP classes. Biotechniques. 2008 Apr;44(5):603-8, 610.

[3] Hedrick, P. W. 2011a. Genetics of populations, 4th Edn, Jones & Bartlett Publishers, Sudburry.

[4] Kayser M, Schneider PM. DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations. Forensic Sci Int Genet. 2009 Jun;3(3):154-61.

[5] Walsh S, Liu F, Ballantyne KN, 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 Jun;5(3):170-80.

[6] Kayser M, de Knijff P. Improving human forensics through advances in genetics, genomics and molecular biology. Nat Rev Genet. 2011 Mar;12(3):179-92.

[7] Collins FS, Brooks LD, Chakravarti A. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res. 1998 Dec;8(12):1229-31.

[8] Ke X, Taylor MS, Cardon LR. Singleton SNPs in the human genome and implications for genome-wide association studies. Eur J Hum Genet. 2008 Apr;16(4):506-15.

[9] John M. Butler 2012, Introduction, In Advanced Topics in Forensic DNA Typing: Methodology, Academic Press, San Diego.

[10] Campbell CD, Eichler EE. Properties and rates of germline mutations in humans. Trends Genet. 2013 Oct;29(10):575-84.

[11] Divne AM, Allen M. A DNA microarray system for forensic SNP analysis. Forensic Sci Int. 2005 Nov 25;154(2-3):111-21.

[12] Sobrino B, Brión M, Carracedo A. SNPs in forensic genetics: a review on SNP typing methodologies. Forensic Sci Int. 2005 Nov 25;154(2-3):181-94.

[13] Sturm RA, Teasdale RD, Box NF. Human pigmentation genes: identification, structure and consequences of polymorphic variation. Gene. 2001 Oct 17;277(1-2):49-62. Review.

[14] Parra EJ, Kittles RA, Shriver MD. Implications of correlations between skin colour and genetic ancestry for biomedical research. Nat Genet. 2004 Nov;36(11 Suppl):S54-60.

[15] Sturm RA. Molecular genetics of human pigmentation diversity. Hum Mol Genet. 2009 Apr: 18(R1):R9-17.

[16] Valenzuela RK, Henderson MS, Walsh MH, Garrison NA, Kelch JT, Cohen-Barak O, Erickson DT, John Meaney F, Bruce Walsh J, Cheng KC, Ito S, Wakamatsu K, Frudakis T, Thomas M, Brilliant MH. Predicting phenotype from genotype: normal pigmentation. J Forensic Sci. 2010 Mar 1;55(2):315-22.

[17] Sturm RA, Duffy DL. Human pigmentation genes under environmental selection. Genome Biology. 2012 Sept:13(9):248.

[18] Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, Manolescu A, Karason A, Palsson A, Thorleifsson G, Jakobsdottir M, Steinberg S, Pálsson S, Jonasson F, Sigurgeirsson B, Thorisdottir K, Ragnarsson R, Benediktsdottir KR, Aben KK, Kiemeney LA, Olafsson JH, Gulcher J, Kong A, Thorsteinsdottir U, Stefansson K. Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet. 2007 Dec;39(12):1443-52.

[19] Liu F, Wen B, Kayser M. Colourful DNA polymorphisms in humans. Semin Cell Dev Biol. 2013 Jun-Jul;24(6-7):562-75.

[20] Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterländer M, Hollfelder N, Potekhina ID, Schier W, Thomas MG, Burger J.Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. Proc Natl Acad Sci U S A. 2014 Apr 1;111(13):4832-7.

[21] Walsh S, Lindenbergh A, Zuniga SB, Sijen T, de Knijff P, Kayser M, Ballantyne KN. Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence. Forensic Sci Int Genet. 2011 Nov;5(5):464-71.

[22] Walsh S, Wollstein A, Liu F, Chakravarthy U, Rahu M, Seland JH, Soubrane G, Tomazzoli L, Topouzis F, Vingerling JR, Vioque J, Fletcher AE, Ballantyne KN, Kayser M. DNA-based eye colour prediction across Europe with the IrisPlex system. Forensic Sci Int Genet. 2012 May;6(3):330-40.

[23] 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 Jan;7(1):98-115.

[24] Hart KL, Kimura SL, Mushailov V, Budimlija ZM, Prinz M, Wurmbach E. Improved eye-and skin-colour prediction based on 8 SNPs.Croat Med J. 2013 Jun;54(3):248-56.

[25] Spichenok O, Budimlija ZM, Mitchell AA, Jenny A, Kovacevic L, Marjanovic D, Caragine T, Prinz M, Wurmbach E. Prediction of eye and skin colour in diverse populations using seven SNPs. Forensic Sci Int Genet. 2011 Nov;5(5):472-8.

[26] Kayser M. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes.Forensic Sci Int Genet. 2015 Sep;18:33-48.

[27] Primorac D, Schanfield M. Forensic DNA Applications: An Interdisciplinary Perspective: a new book in forensic science. Croat. Med. Journal. 2014 Aug;55(4):434-436.

[28] Sturm RA, Frudakis TN. Eye colour: portals into pigmentation genes and ancestry. Trends Genet. 2004 Aug;20(8):327-32.

[29] Sturm RA, Larsson M. Genetics of human iris colour and patterns. Pigment Cell Melanoma Res. 2009 Oct;22(5):544-62.

[30] http://www.sciencealert.com/this-is-how-blue-eyes-get-their-colour

[31] Branicki, W. and Kayser, M. 2015. Prediction of Human Pigmentation Traits from DNA Polymorphisms. eLS. 1–10.

[32] Kayser M, Liu F, Janssens CJW, Rivadeneira F, Lao O, Duijn K, Vermeulen M, Arp P, Jhamai MM, IJcken WFJ, et al. Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris colour gene. Am J Hum Genet. 2008;82:411–423.

[33] Beals, R.L. and H. Hoijer. 1965. An Introduction to Anthropology, 3rd edition, New York: MacMillan Co.

[34] Frost, P. European hair and eye colour - A case of frequency-dependent sexual selection? Evolution and Human Behavior. 2006 Mar: 27(2): 85-103.

[35] Frudakis T, Venkateswarlu K, Thomas MJ, Gaskin Z, Ginjupalli S, Gunturi S, Ponnuswamy V, Natarajan S, Nachimuthu PK. A classifier for the SNP-based inference of ancestry. J Forensic Sci. 2003 Jul;48(4):771-82.

[36] Sturm RA, Duffy DL, Zhao ZZ, Leite FPN, Stark MS, Hayward NK, Martin NG, Montgomery GW. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye colour. Am J Hum Genet. 2008;82:424–431.

[37] Duffy DL, Montgomery GW, Chen W, Zhao ZZ, Le L, James MR, Hayward NK, Martin NG, Sturm RA. A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-colour variation. Am J Hum Genet. 2007;80:241–252.

[38] Han Y, Gu S, Oota H, Osier M, Pakstis AJ, Speed WC, Kidd JR, Kidd KK. Evidence of positive selection on a class I ADH locus. Am J Hum Genet. 2007;80:441–456.

[39] Liu F, van Duijn K, Vingerling JR, Hofman A, Uitterlinden AG, Janssens AC, Kayser M. Eye colour and the prediction of complex phenotypes from genotypes. Curr Biol. 2009 Mar 10;19(5):R192-3.

[40] Eiberg H, Troelsen J, Nielsen M, Mikkelsen A, Mengel-From J, Kjaer KW, Hansen L. Blue eye colour in human 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.

[41] Liu F, Wollstein A, Hysi PG, Ankra-Badu GA, Spector TD, Park D, Zhu G, Larsson M, Duffy DL, Montgomery GW, Mackey DA, Walsh S, Lao O, Hofman A, Rivadeneira F, Vingerling JR, Uitterlinden AG, Martin NG, Hammond CJ, Kayser M. Digital quantification of human eye colour highlights genetic association of three new loci. PLoS Genet. 2010 May 6;6(5).

[42] Ruiz Y, Phillips C, Gomez-Tato A, Alvarez-Dios J, Casares de Cal M, Cruz R, Maroñas O, Söchtig J, Fondevila M, Rodriguez-Cid MJ, Carracedo A, Lareu MV. Further development of forensic eye colour predictive tests. Forensic Sci Int Genet. 2013 Jan;7(1):28-42.

[43] Slominski A, Wortsman J, Plonka PM, Schallreuter KU, Paus R, Tobin DJ. Hair follicle pigmentation. J Invest Dermatol. 2005 Jan;124(1):13-21. Review.

[44] Rees JL. Genetics of hair and skin colour. Annu Rev Genet. 2003;37:67-90. Review.

[45] Kenny EE, Timpson NJ, Sikora M, Yee MC, Moreno-Estrada A, Eng C, Huntsman S, Burchard EG, Stoneking M, Bustamante CD, Myles S. Melanesian blond hair is caused by an amino acid change in TYRP1. Science. 2012 May 4;336(6081):554.

[46] Valverde P, Healy E, Jackson I, Rees JL, Thody AJ. Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. Nat Genet. 1995 Nov;11(3):328-30.

[47] Harding RM, Healy E, Ray AJ, Ellis NS, Flanagan N, Todd C, Dixon C, Sajantila A, Jackson IJ, Birch-Machin MA, Rees JL. Evidence for variable selective pressures at *MC1R*. Am J Hum Genet.2000;66:1351–1361.

[48] Grimes EA, Noake PJ, 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.

[49] 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 colour using DNA variants.

Hum Genet. 2011 Apr;129(4):443-54.

[50] Alaluf S, Atkins D, Barrett K, Blount M, Carter N, Heath A. Ethnic variation in melanin content and composition in photoexposed and photoprotected human skin. Pigment Cell Res. 2002 Apr;15(2):112-8.

[51] Barsh G, Gunn T, He L, Schlossman S, Duke-Cohan J. Biochemical and genetic studies of pigment-type switching. Pigment Cell Res. 2000;13 Suppl 8:48-53.

52. Jablonski NG, Chaplin G. The evolution of human skin colouration. J Hum Evol. 2000 Jul;39(1):57-106.

[53] Stokowski RP, Pant PV, Dadd T, Fereday A, Hinds DA, Jarman C, Filsell W, Ginger RS, Green MR, van der Ouderaa FJ, Cox DR.A genomewide association study of skin pigmentation in a South Asian population. Am J Hum Genet. 2007 Dec;81(6):1119-32.

[54] Maroñas O, Phillips C, Söchtig J, Gomez-Tato A, Cruz R, Alvarez-Dios J, de Cal MC, Ruiz Y, Fondevila M, Carracedo Á, Lareu MV. Development of a forensic skin colour predictive test. Forensic Sci Int Genet. 2014 Nov;13:34-44.

[55] van Oven M, Vermeulen M, Kayser M. Multiplex genotyping system for efficient inference of matrilineal genetic ancestry with continental resolution. Investig Genet. 2011 Mar 23;2:6.

[56] Royal CD, Novembre J, Fullerton SM, Goldstein DB, Long JC, Bamshad MJ, Clark AG. Inferring genetic ancestry: opportunities, challenges, and implications. Am J Hum Genet. 2010 May 14;86(5):661-73.

[57] Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H, Zegura SL.Hierarchical patterns of global human Y-chromosome diversity. Mol Biol Evol. 2001 Jul;18(7):1189-203.

[58] Jobling MA, Gill P. Encoded evidence: DNA in forensic analysis. Nat Rev Genet. 2004 Oct;5(10):739-51. Review. Erratum in: Nat Rev Genet. 2005 Mar;6(3):246.

[59] Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, Holmlund G, Kouvatsi A, Macek M, Mollet I, Parson W, Palo J, Ploski R, Sajantila A, Tagliabracci A, Gether U, Werge T, Rivadeneira F, Hofman A, Uitterlinden AG, Gieger C, Wichmann HE, Rüther A, Schreiber S, Becker C, Nürnberg P, Nelson MR, Krawczak M, Kayser M. Correlation between genetic and geographic structure in Europe. Curr Biol. 2008 Aug 26;18(16):1241-8.

[60] Shriver MD, Parra EJ, Dios S, Bonilla C, Norton H, Jovel C, Pfaff C, Jones C, Massac A, Cameron N, Baron A, Jackson T, Argyropoulos G, Jin L, Hoggart CJ, McKeigue PM, Kittles RA. Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet. 2003 Apr;112(4):387-99.

[61] Kidd JR, Friedlaender FR, Speed WC, Pakstis AJ, De La Vega FM, Kidd KK. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples.Investig Genet. 2011 Jan 5;2(1):1.

[62] Pereira R, Phillips C, Pinto N, Santos C, dos Santos SE, Amorim A, Carracedo Á, Gusmão L. Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. PLoS One. 2012;7(1):e29684.

[63] Ginther C, Issel-Tarver L, King MC. Identifying individuals by sequencing mitochondrial DNA from teeth. Nat Genet. 1992 Oct;2(2):135-8.

[64] Allen, M., Engstrom, A. S., Meyers, S., et al. Mitochondrial DNA sequencing of shed hairs and saliva on robbery caps: sensitivity and matching probabilities. J Forensic Sci 1998; 43(3):453-64.

[65] Lutz, S., Weisser, H. J., Heizmann, J., et al. mtDNA as a tool for identification of human remains. Identification using mtDNA. Int J Legal Med 1996; 109(4):205-9.

[66] Szibor, R., Michael, M., Plate, I., et al. Efficiency of forensic mtDNA analysis. Case examples demonstrating the identification of traces. Forensic Sci Int 2000; 113(1-3):71-8.

[67] Wilson, M. R., Polanskey, D., Butler, J., et al. Extraction, PCR amplification and sequencing of mitochondrial DNA from human hair shafts. Biotechniques 1995; 18(4):662-9.

[68] Parson W, Brandstätter A, Alonso A, Brandt N, Brinkmann B, Carracedo A, Corach D, Froment O, Furac I, Grzybowski T, Hedberg K, Keyser-Tracqui C, Kupiec T, Lutz-Bonengel S, Mevag B, Ploski R, Schmitter H, Schneider P, Syndercombe-Court D, Sørensen E, Thew H, Tully G, Scheithauer R. The EDNAP mitochondrial DNA population database (EMPOP) collaborative exercises: organisation, results and perspectives.Forensic Sci Int. 2004 Jan 28;139(2-3):215-26.

[69] Parson W, Bandelt HJ. Extended guidelines for mtDNA typing of population data in forensic science. Forensic Sci Int Genet. 2007 Mar;1(1):13-9.

[70] Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG. Sequence and organization of the human mitochondrial genome. Nature. 1981 Apr 9;290(5806):457-65.

[71] Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet. 1999 Oct;23(2):147.

[72] Brandstatter, A., Parsons, T. J. and Parson, W. Rapid screening of mtDNA coding region SNPs for the identification of west European Caucasian haplogroups. Int J Legal Med 2003; 117(5):291-8.

[73] Schlebusch CM, Naidoo T, Soodyall H. SNaPshot minisequencing to resolve mitochondrial macro-haplogroups found in Africa. Electrophoresis. 2009 Nov;30(21): 3657-64.

[74] Connor A, Stoneking M. Assessing ethnicity from human mitochondrial DNA types determined by hybridization with sequence-specific oligonucleotides. J Forensic Sci. 1994 Nov;39(6):1360-71.

[75] Wallace DC, Brown MD, Lott MT. Mitochondrial DNA variation in human evolution and disease. Gene. 1999 Sep 30;238(1):211-30. Review.

[76] Tanaka M, Cabrera VM, González AM, Larruga JM, Takeyasu T, Fuku N, Guo LJ, Hirose R, Fujita Y, Kurata M, Shinoda K, Umetsu K, Yamada Y, Oshida Y, Sato Y, Hattori N, Mizuno Y, Arai Y, Hirose N, Ohta S, Ogawa O, Tanaka Y, Kawamori R, Shamoto-Nagai M, Maruyama W, Shimokata H, Suzuki R, Shimodaira H. Mitochondrial genome variation in eastern Asia and the peopling of Japan. Genome Res. 2004 Oct;14(10A):1832-50.

[77] Roostalu U, Kutuev I, Loogväli EL, Metspalu E, Tambets K, Reidla M, Khusnutdinova EK, Usanga E, Kivisild T, Villems R. Origin and expansion of haplogroup H, the dominant human mitochondrial DNA lineage in West Eurasia: the Near Eastern and Caucasian perspective. Mol Biol Evol. 2007 Feb;24(2):436-48.

[78] Parson, W., Parsons, T. J., Scheithauer, R., et al. Population data for 101 Austrian Caucasian mitochondrial DNA d-loop sequences: application of mtDNA sequence analysis to a forensic case. Int J Legal Med 1998; 111(3):124-32.

[79] Coble, M. D., Just, R. S., O'Callaghan, J. E., et al. Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. Int J Legal Med 2004; 118(3):137-46.

[80] Herrnstadt, C., Elson, J. L., Fahy, E., et al. Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. Am J Hum Genet 2002; 70(5):1152-71.

[81] Ingman, M., Kaessmann, H., Paabo, S., et al. Mitochondrial genome variation and the origin of modern humans. Nature 2000; 408(6813):70813.

[82] Ottoni C, Primativo G, Hooshiar Kashani B, Achilli A, Martínez-Labarga C, Biondi G, Torroni A, Rickards O. Mitochondrial haplogroup H1 in north Africa: an early holocene arrival from Iberia. PLoS One. 2010 Oct 21;5(10):e13378.

[83] Underhill PA, Kivisild T. Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. Annu Rev Genet. 2007;41:539-64. Review.

[84] Gill P, Ivanov PL, Kimpton C, Piercy R, Benson N, Tully G, Evett I, Hagelberg E, Sullivan K.Identification of the remains of the Romanov family by DNA analysis. Nat Genet. 1994 Feb;6(2):130-5.

[85] Lutz S, Weisser HJ, Heizmann J, Pollak S. Location and frequency of polymorphic positions in the mtDNA control region of individuals from Germany. Int J Legal Med. 1998;111(2):67-77. Erratum in: Int J Legal Med 1999;112(2):145-50. Int J Legal Med 1998;111(5):286.

[86] Brandstätter A, Niederstätter H, Pavlic M, Grubwieser P, Parson W. Generating population data for the EMPOP database - an overview of the mtDNA sequencing and data evaluation processes considering 273 Austrian control region sequences as example. Forensic Sci Int. 2007 Mar 2;166(2-3):164-75.

[87] Holland MM, Parsons TJ. Mitochondrial DNA Sequence Analysis - Validation and Use for Forensic Casework. Forensic Sci Rev. 1999 Jun;11(1):21-50. Review.

[88] Palo JU, Hedman M, Söderholm N, Sajantila A. Repatriation and identification of the Finnish World War II soldiers. Croat Med J. 2007 Aug;48(4):528-35.

[89] Coble MD, Vallone PM, Just RS, Diegoli TM, Smith BC, Parsons TJ. Effective strategies for forensic analysis in the mitochondrial DNA coding region. Int J Legal Med. 2006 Jan;120(1):27-32.

[90] Nilsson M, Andréasson-Jansson H, Ingman M, Allen M. Evaluation of mitochondrial DNA coding region assays for increased discrimination in forensic analysis. Forensic Sci Int Genet. 2008 Jan;2(1):1-8.

[91] Daca P, Mielnik M, Rogalla U, Skonieczna K, Linkowska K, Grzybowski T. [The application of minisequencing reactions for haplogroup assignment of mitochondrial DNA]. Arch Med Sadowej Kryminol. 2008 Oct-Dec;58(4):212-7. Review.

[92] Quintáns B, Alvarez-Iglesias V, Salas A, Phillips C, Lareu MV, Carracedo A. Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing.Forensic Sci Int. 2004 Mar 10;140(2-3):251-7.

[93] Ballantyne KN, van Oven M, Ralf A, Stoneking M, Mitchell RJ, van Oorschot RA, Kayser M. MtDNA SNP multiplexes for efficient inference of matrilineal genetic ancestry within Oceania. Forensic Sci Int Genet. 2012 Jul;6(4):425-36.

[94] Phillips C, Fernandez-Formoso L, Gelabert-Besada M, Garcia-Magariños M, Santos C, Fondevila M, Carracedo A, Lareu MV. Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing. Electrophoresis. 2013 Apr;34(8):1151-62.

[95] Fondevila M, Phillips C, Santos C, Freire Aradas A, Vallone PM, Butler JM, Lareu MV, Carracedo A.Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies. Forensic Sci Int Genet. 2013 Jan;7(1):63-74.

[96] Hammer MF and Zegura SL. The Human Y Chromosome Haplogroup Tree: Nomenclature and Phylogeography of Its Major Divisions. Annual Review of Anthropology, Vol. 31: 303 -321

[97] Jobling MA, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet. 2003 Aug;4(8):598-612. Review.

[98] Shriver MD, Kittles RA. Genetic ancestry and the search for personalized genetic histories. Nat Rev Genet. 2004 Aug;5(8):611-8.

[99] Carracedo A, Bär W, Lincoln P, Mayr W, Morling N, Olaisen B, Schneider P, Budowle B, Brinkmann B, Gill P, Holland M, Tully G, Wilson M. DNA commission of the international society for forensic genetics: guidelines for mitochondrial DNA typing. Forensic Sci Int. 2000 May 15;110(2):79-85.

[100] Tully G, Lareu M; European DNA Profiling (EDNAP) Group. Controversies over heteroplasmy. Electrophoresis. 2001 Jan;22(1):180-2.

[101] Bandelt HJ, Parson W. Consistent treatment of length variants in the human mtDNA control region: a reappraisal. Int J Legal Med. 2008 Jan;122(1):11-21.

[102] Behar DM, Rosset S, Blue-Smith J, Balanovsky O, Tzur S, Comas D, Mitchell RJ, Quintana-Murci L, Tyler-Smith C, Wells RS; Genographic Consortium. The Genographic Project public participation mitochondrial DNA database.PLoS Genet. 2007 Jun;3(6):e104. Erratum in: PLoS Genet. 2007 Sep 14;3(9):1785.

[103] Parson W, Dür A. EMPOP--a forensic mtDNA database.Forensic Sci Int Genet. 2007 Jun;1(2):88-92.

[104] Bandelt HJ, Quintana-Murci L, Salas A, Macaulay V. The fingerprint of phantom mutations in mitochondrial DNA data. Am J Hum Genet. 2002 Nov;71(5):1150-60.

[105] Bandelt HJ, Salas A, Lutz-Bonengel S. Artificial recombination in forensic mtDNA population databases.Int J Legal Med. 2004 Oct;118(5):267-73.

[106] Yao YG, Bravi CM, Bandelt HJ. A call for mtDNA data quality control in forensic science. Forensic Sci Int. 2004 Apr 20;141(1):1-6.

[107] Salas A, Bandelt HJ, Macaulay V, Richards MB. Phylogeographic investigations: the role of trees in forensic genetics. Forensic Sci Int. 2007 May 3;168(1):1-13.Review.

[108] Zimmermann B, Röck A, Huber G, Krämer T, Schneider PM, Parson W. Application of a west Eurasian-specific filter for quasi-median network analysis: Sharpening the blade for mtDNA error detection. Forensic Sci Int Genet. 2011 Mar;5(2):133-7.

[109] Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. J Genet Genomics. 2011 Mar 20;38(3):95-109. Review.

[110] Metzker ML. Emerging technologies in DNA sequencing.Genome Res. 2005 Dec;15(12):1767-76. Review.

[111] Warshauer DH, Lin D, Hari K, Jain R, Davis C, Larue B, King JL, Budowle B. STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. Forensic Sci Int Genet. 2013 Jul;7(4):409-17.

[112] Xue Y, Wang Q, Long Q, Ng BL, Swerdlow H, Burton J, Skuce C, Taylor R, Abdellah Z, Zhao Y; Asan, MacArthur DG, Quail MA, Carter NP, Yang H, Tyler-Smith C. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr Biol. 2009 Sep 15;19(17):1453-7.

[113] Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature. 2000 Sep 28;407(6803):513-6.

[114] Miller W, Drautz DI, Janecka JE, Lesk AM, Ratan A, Tomsho LP, Packard M, Zhang Y, McClellan LR, Qi J, Zhao F, Gilbert MT, Dalén L, Arsuaga JL, Ericson PG, Huson DH, Helgen KM, Murphy WJ, Götherström A, Schuster SC. The mitochondrial genome sequence of the Tasmanian tiger (Thylacinus cynocephalus). Genome Res. 2009 Feb;19(2):213-20.

[115] Irwin JA, Saunier JL, Strouss KM, Diegoli TM, Sturk KA, O'Callaghan JE, Paintner CD, Hohoff C, Brinkmann B, Parsons TJ. Mitochondrial control region sequences from a Vietnamese population sample. Int J Legal Med. 2008 May;122(3):257-9.

[116] Stiller M, Knapp M, Stenzel U, Hofreiter M, Meyer M. Direct multiplex sequencing (DMPS)--a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA. Genome Res. 2009 Oct;19(10):1843-8.

[117] Willerslev E, Gilbert MT, Binladen J, Ho SY, Campos PF, Ratan A, Tomsho LP, da Fonseca RR, Sher A, Kuznetsova TV, Nowak-Kemp M, Roth TL, Miller W, Schuster SC. Analysis of complete mitochondrial genomes from extinct and extant rhinoceroses reveals lack of phylogenetic resolution. BMC Evol Biol. 2009 May 11;9:95.

[118] Lindqvist C, Schuster SC, Sun Y, Talbot SL, Qi J, Ratan A, Tomsho LP, Kasson L, Zeyl E, Aars J, Miller W, Ingólfsson O, Bachmann L, Wiig O. Complete mitochondrial genome of a Pleistocene jawbone unveils the origin of polar bear. Proc Natl Acad Sci U S A. 2010 Mar 16;107(11):5053-7.

[119] Ho SY, Gilbert MT. Ancient mitogenomics. Mitochondrion. 2010 Jan;10(1):1-11. Review.

[120] Green RE, Malaspinas AS, Krause J, Briggs AW, Johnson PL, Uhler C, Meyer M, Good JM, Maricic T, Stenzel U, Prüfer K, Siebauer M, Burbano HA, Ronan M, Rothberg JM, Egholm M, Rudan P, Brajković D, Kućan Z, Gusić I, Wikström M, Laakkonen L, Kelso J, Slatkin M, Pääbo S. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell. 2008 Aug 8;134(3):416-26.

[121] Davis C, Peters D, Warshauer D, King J, Budowle B. Sequencing the hypervariable regions of human mitochondrial DNA using massively parallel sequencing: Enhanced data acquisition for DNA samples encountered in forensic testing. Leg Med (Tokyo). 2015 Mar;17(2):123-7.

[122] Parson W, Huber G, Moreno L, Madel MB, Brandhagen MD, Nagl S, Xavier C, Eduardoff M, Callaghan TC, Irwin JA. Massively parallel sequencing of complete mitochondrial genomes from hair shaft samples. Forensic Sci Int Genet. 2015 Mar;15:8-15.

[123] Ralf A, van Oven M, Zhong K, Kayser M. Simultaneous analysis of hundreds of Y-chromosomal SNPs for high-resolution paternal lineage classification using targeted semiconductor sequencing.Hum Mutat. 2015 Jan;36(1):151-9.

[124] Vallone PM, Just RS, Coble MD, Butler JM, Parsons TJ. A multiplex allele-specific primer extension assay for forensically informative SNPs distributed throughout the mitochondrial genome. Int J Legal Med. 2004 Jun;118(3):147-57.

[125] Sanchez JJ, Phillips C, Børsting C, Balogh K, Bogus M, Fondevila M, Harrison CD, Musgrave-Brown E, Salas A, Syndercombe-Court D, Schneider PM, Carracedo A, Morling N. A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis. 2006 May;27(9):1713-24.

[126] Phillips C, Salas A, Sánchez JJ, Fondevila M, Gómez-Tato A, Alvarez-Dios J, Calaza M, de Cal MC, Ballard D, Lareu MV, Carracedo A; SNPforID Consortium. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genet. 2007 Dec;1(3-4):273-80.

[127] Alessandrini F, Turchi C, Onofri V, Buscemi L, Pesaresi M, Tagliabracci A. Multiplex PCR development of Y-chromosomal biallelic polymorphisms for forensic application. J Forensic Sci. 2005 May;50(3):519-25.

[128] Onofri V, Alessandrini F, Turchi C, Pesaresi M, Buscemi L, Tagliabracci A. Development of multiplex PCRs for evolutionary and forensic applications of 37 human Y chromosome SNPs. Forensic Sci Int. 2006 Feb 10;157(1):23-35.

[129] Bouakaze C, Keyser C, Amory S, Crubézy E, Ludes B. First successful assay of Y-SNP typing by SNaPshot minisequencing on ancient DNA. Int J Legal Med. 2007 Nov;121(6): 493-9.

[130] Krjutskov K, Viltrop T, Palta P, Metspalu E, Tamm E, Suvi S, Sak K, Merilo A, Sork H, Teek R, Nikopensius T, Kivisild T, Metspalu A. Evaluation of the 124-plex SNP typing microarray for forensic testing. Forensic Sci Int Genet. 2009 Dec;4(1):43-8.

[131] Paneto GG, Köhnemann S, Martins JA, Cicarelli RM, Pfeiffer H. A single multiplex PCR and SNaPshot minisequencing reaction of 42 SNPs to classify admixture populations into mitochondrial DNA haplogroups. Mitochondrion. 2011 Mar;11(2):296-302.

[132] Tully G, Sullivan KM, Nixon P, Stones RE, Gill P. Rapid detection of mitochondrial sequence polymorphisms using multiplex solid-phase fluorescent minisequencing. Genomics. 1996 May 15;34(1):107-13.

[133] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. 1977. Biotechnology. 1992;24:104-8.

[134] Chial, H. DNA sequencing technologies key to the Human Genome Project. Nature Education. 2008 1(1):219

[135] Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods. 2010 Jun;7(6):461-5.

Chapter 2

Collaborative EDNAP Exercise on the IrisPlex system for DNA based prediction of human eye color

Forensic Science International: Genetics 11 (2014) 241 – 251

Abstract

The IrisPlex system is a DNA-based test system for the prediction of human eye colour from biological samples and consists of a single forensically validated multiplex genotyping assay together with a statistical prediction model that is based on genotypes and phenotypes from thousands of individuals. IrisPlex predicts blue and brown human eye colour with, on average, >94% precision accuracy using six of the currently most eye colour informative single nucleotide polymorphisms (*HERC2* rs12913832, *OCA2* rs1800407, *SLC24A4* rs12896399, *SLC45A2 (MATP)* rs16891982, *TYR* rs1393350, and *IRF4* rs12203592) according to a previous study, while the accuracy in predicting nonblue and non-brown eye colours is considerably lower. In an effort to vigorously assess the IrisPlex system at the international level, testing was performed by 21 laboratories in the context of a collaborative exercise divided into three tasks and organized by the European DNA Profiling (EDNAP) Group of the International Society of Forensic Genetics (ISFG). Task 1 involved the assessment of 10 blood and saliva samples provided on FTA cards by the organising laboratory together with eye colour phenotypes; 99.4% of the genotypes were correctly reported and 99% of the eye colour phenotypes were correctly predicted. Task 2 involved the assessment of 5 DNA samples extracted by the host laboratory from simulated casework samples, artificially degraded, and provided to the participants in varying DNA concentrations. For this task, 98.7% of the genotypes were correctly determined and 96.2% of eye colour phenotypes were correctly inferred. For Tasks 1 and 2 together, 99.2% (1875) of the 1890 genotypes were correctly generated and of the 15 (0.8%) incorrect genotype calls, only 2 (0.1%) resulted in incorrect eye colour phenotypes. The voluntary Task 3 involved participants choosing their own test subjects for IrisPlex genotyping and eye colour phenotype inference, while eye photographs were provided to the organising laboratory and judged; 96% of the eye colour phenotypes were inferred correctly across 100 samples and 19 laboratories. The high success rates in genotyping and eye colour phenotyping clearly demonstrate the reproducibility and the robustness of the IrisPlex assay as well as the accuracy of the IrisPlex model to predict blue and brown eye colour from DNA. Additionally, this study demonstrates the ease with which the IrisPlex system is implementable and applicable across forensic laboratories around the world with varying pre-existing experiences.

Introduction

The field of forensic genetics is making great strides with the rapid scientific and technological evolution in obtaining new knowledge and creating innovative tools for solving crimes more and more effectively. Forensic DNA Phenotyping (FDP), a nascent advancement in this field, is one example of recent innovative developments in forensic genetics and involves the prediction of an individual's externally visible characteristics (EVCs) using biological samples obtained at a crime scene or from an anonymous body (parts) that may belong to a missing person [1–4]. Conventional DNA identification involves the comparison of DNA profiles derived from short tandem repeat (STR) marker genotypes obtained from evidence and reference samples, which is useful in cases when the sample donor is known from their DNA profile. In certain circumstances, sample donors may not be identified, i.e. a match (or familial match) of the DNA profiles with known suspects such as those in criminal offender DNA (profile) databases or with ante-mortem samples in cases of missing persons is not successful, or when DNA profile comparisons with putative relatives of missing persons does not reveal the degree of similarities indicating biological relationship. In these situations, FDP can be used to help investigative authorities focus their search for unknown suspects or missing persons towards individuals with particular DNA-predicted externally visible phenotypes. The DNA-based prediction of EVCs can thus aid investigations by police and other authorities by reducing the number of possible suspects or other individuals if conventional STR typing of the evidence fails to produce identification [1,2]. Furthermore, reconstructing appearance information from biological samples such as bones or teeth or other remains of deceased individuals is relevant in anthropological research disciplines including those relying on ancient DNA analysis [1,2,5].

Several model-based approaches, amongst others [4,6], have been developed for predicting a particular phenotype from DNA most notably human eye (iris) colour [7]; the IrisPlex system is one such tool [8]. IrisPlex can accurately predict blue and brown eye colour with a precision of >94%, according to a previous study [9], using six of the most informative eye colour markers: rs12913832 (*HERC2*), rs1800407 (*OCA2*), rs12896399 (*SLC24A4*), rs16891982 (*SLC45A2 (MATP)*), rs1393350 (*TYR*) and rs12203592 (*IRF4*) in a single genotyping assay and a prediction model based on thousands of individuals for which IrisPlex genotype and eye colour phenotype data are available [7,8,10]. The 94%

accuracy is based on using a threshold of p >0.7, however it is possible to use IrisPlex prediction with a lower p >0.5 threshold. The assessment of precision accuracy is based on a broad European dataset of >3800 individuals using IrisPlex can be found in Table 3 of that publication [9]. The IrisPlex assay represents the first FDP system that successfully underwent developmental validation using the Scientific Working Group on DNA Analysis Methods (SWGDAM) guidelines for use in forensic casework [10]. The IrisPlex prediction model, first established on thousands of Dutch Europeans, has been evaluated in several populations within and outside Europe and was shown to perform reliably, independent of the bio-geographic origin of the individual tested [9,11].

In an international effort to test the reliability and consistency of the IrisPlex system for eye colour prediction through an inter laboratory exercise, the European DNA Profiling (EDNAP) Group, a working group of the International Society for Forensic Genetics (ISFG), carried out a collaborative study led by the Department of Forensic Molecular Biology of the Erasmus University Medical Center Rotterdam (Netherlands), who initially developed and validated the IrisPlex system [8–10] and for this reason were chosen to conduct this further assessment on the IrisPlex tool alone. Of the 21 participating laboratories, 18 were from Europe, 2 were from Australia and 1 was from the U.S.A. The prime aim of this exercise was to implement the method and assess the performance of the system across different forensic laboratories with varying levels of experience, from complete novices with no SNP typing experience to participants with SNaPshot experience and those with specific IrisPlex experience. Notably, some authors previously raised issues about marker content and model outcomes of the IrisPlex system [12–15]. The present collaborate EDNAP exercise, however, represents a rather technical exercise to test the performance of the IrisPlex system across laboratories with varying levels of pre-existing experience. Therefore, issues about marker and model choice for predicting eye colour from DNA may be addressed in more dedicated future studies. Here, we present the results of this collaborative EDNAP exercise, placing emphasis on the reliability and consistency in using the IrisPlex system for blue and brown eye colour prediction from DNA.

Materials and methods

Samples and materials provided to the participating laboratories

The organising laboratory (Department of Forensic Molecular Biology, Erasmus MC University Medical Center Rotterdam, Netherlands) divided the entire exercise into three different tasks. All participants were provided with a detailed written laboratory protocol [8] as well as the eye colour prediction model that is an interactive excel spread-sheet as published earlier [10]. Furthermore, for assay interpretation guidelines, participants were given a protocol stating a 50 relative fluorescent units (rfu) peak height threshold should be used for allele calls using the IrisPlex specific GeneMapper software (Applied Biosystems) Bin and Panel set provided. For a broader understanding of average peak heights and balance ratios, participants were asked to refer to the previous developmental validation publication of the IrisPlex system [10]. In addition to the samples and the primers provided for each task, all reagents, which include: 1X PCR buffer, 2.7 mM $MgCl_2$, 200 mM of each dNTP, 0.5 U AmpliTaq Gold Polymerase, SNaPshot Multiplex chemistry for the single base extension (SBE) reactions (Applied Biosystems, Foster City, CA), Exonuclease Shrimp Alkaline Phosphatase (ExoSAP-IT), and Shrimp Alkaline phosphatase (SAP) (USB Corporation, Cleveland, OH), required for running the IrisPlex system were shipped on dry ice to each of the 21 participating laboratories. The laboratories were asked to use their own internal sizing standard (LIZ 120) and formamide for the capillary electrophoresis run. Due to an ExoSAP-IT degradation issue noted during the early phase of the exercise, which subsequently was acknowledged by the producing company as a bad batch of enzyme, aliquots of a newly delivered and tested ExoSAP-IT were shipped again to the requesting laboratories, while the others opted to use their in-house standard cleaning protocols. As this was a clean-up procedure, it did not impede on the testing of the IrisPlex assay overall. Purified products were run by the laboratories using their in-house Genetic Analysers (for type, see Table 1) and analysed with the previously published eye colour prediction model [10] provided by the organising laboratory for predicting human eye colour from IrisPlex genotypes.

Lab ID	Extraction Protocol	Quantification Protocol	Polymer	Genetic Analyzer
1			POP 4	3130xl
2	EZ1 DNA Investigator Kit on an EZ1 Advanced XL (Qiagen)	Quantifiler Duo DNA Quantification Kit (Life Technologies)	POP 6	3100
3			POP 4	3130
4	Task 1: EZ1 advanced (Qiagen); Task 3:DNA Blood Mini Kit (Qiagen)	ABI prism® 7900 (Life Technologies-LT) using Quantifiler® Human DNA Quantification Kit (LT).	POP 4	3130xl
5	Chelex + Qiaquick	Task 1: Qubit and Task 2: RT-PCR Quantifiler	POP 4	3130xl
6	QIAamp® DNA Mini (Qiagen, Hilden, Germany)	qPCR using the 7900HT Fast Real-time PCR System (Applied Biosystems, Darmstadt, Germany) and Alu Primers	POP 4	3500xl
7	Chelex	Quantifiler Duo	POP 7	3130xl
8	5% solution of BT Chelex 100 resin (Bio-Rad)	Quantifiler® Human DNA Quantification Kit (Life Technologies) on AB 7900 RT-PRC	POP 4	3130
9	QIAmp DNA blood Mini kit from Qiagen	Quantifiler Human kit on a ABI7500	POP 7	3500
10	QIAmp DNA blood Mini kit from Qiagen	Nanodrop	POP 7	3500
11		Quantifiler Human	POP 4	3130xl
12	QIAamp mini columns (Qiagen)	real-time quantitative PCR assay using ALU repeats from Nicklas et el.	POP 7	3130xl
13	EZ1 robot (Qiagen)	Quantifiler	POP 7	3130xl
14	Qiagen EZ1 Advanced XL extraction robot with the EZ1 DNA Investigator Kit	Quantifiler Human DNA Quantification Kit (Life Technologies)	POP 4	3130xl
15			POP 4	
16	QiAmp DNA Mini kit (Qiagen)		POP 7	3500
17	Qiagen EZ1 Investigator Handbook	Thermo Scientific NanoDrop 2000/2000c spectrophotometer	POP 4	3130
18	Applied Byosistems Prepfiler Forensic DNA Extraction kit	Task 1 and Task 3 a 1% agarose gel; Task 2 - a RT PCR	POP 4	3130
19	Chelex	Quantifiler DUO in an 7500 Real-Time PCR System	POP 7	3130
20	Phenol–Chloroform	Quantifiler Human DNA Quantification Kit and AB 7300.	POP 4	3130
21	Qiagen M48 robotic station and MagAttract DNA Mini M48 Kit	Quantifiler kit and ABI 7500	POP 7	3130xl

Table 1: DNA extraction and quantification protocols used by the 21 laboratories for both Task 1 and 2. Grey boxes indicate no data received from the participating laboratory.

As a disclaimer for the choice of samples used in this assessment, please note that it is well established and documented [7,9,16,17], that the IrisPlex system through its use of six eye colour associated SNPs performs very well in predicting blue and brown eye colour with Area Under the receiving operator Curve (AUC) values >0.9; however its use for predicting intermediate eye colour (current AUC of ~0.7) is not at an optimum level yet. This is due to the current lack of knowledge on DNA predictors for these non-blue, non-brown eye colours i.e. green eye colour, individuals with heterochromia etc. which is not only a limitation of the IrisPlex but of all currently available DNA test systems for eye colour [4,6]. DNA variants with similarly high prediction effects on non-blue/brown eye colours as the IrisPlex SNPs have on blue and brown eye colour have yet to be identified. Therefore, the IrisPlex system was previously promoted for the prediction of blue and brown eye colours and thus the organising laboratory opted to test variations in blue and brown eye colour alone for Tasks 1 and 2 to evaluate the current IrisPlex system assay and prediction performance on these categories. Task 3 however incorporates all three categories as this task was based on samples provided by the participating laboratories who were not asked to focus on blue and brown eye colour alone when selecting their Task 3 volunteers.

Tasks 1 and 2 contain samples from individuals of European (80% per task) and non-European bio-geographic origin (20% per task) including one admixed individual in Task 1. The individuals used in Task 3, including information about their bio-geographic background, were at the discretion of the participating laboratories and were unknown to the organising lab.

Task 1- IrisPlex eye colour prediction from biological samples with eye colour knowledge

The organising laboratory provided all participating laboratories with five blood samples (labelled Ind1–Ind5) and five saliva samples (labelled Ind6–Ind10) on FTA cards of 10 individuals with blue (N=5) or brown (N=5) eye colour. To produce these samples, fresh venous blood and saliva samples were collected from ten different individuals and 100 mL of each of the samples were pipetted on to the FTA cards. A digital eye image from each of these 10 individuals who donated blood or saliva was also provided to the participants. All the laboratories were instructed to use their own in-house DNA extraction and quantification protocols. All participating laboratories were asked to generate the IrisPlex genotype profile from each of the 10 samples and using the provided IrisPlex prediction model, to conclude

the eye colour prediction of the 10 donor individuals by noting the probability and precision accuracy per each sample and individual using the guidelines implemented in a previous publication [9] as provided to the participants. An example report was also provided to each laboratory to ease the fill-out for return. As the participants were provided with eye pictures, they already knew the eye colour of the study individuals of Task 1 before analyses. However, the conclusion of the eye colour phenotypes had to be based on the genotypes determined by each individual laboratory considering the probability and precision accuracy guidelines provided by the organising laboratory. The term 'precision accuracy' relates to the previous publication [9], which undertook a study on the final prediction called by the IrisPlex model in terms of probability values on over 3800 European individuals. It assesses the highest probability value (which is defined as the eye colour of the individual) and how correct the eye colour prediction was at thresholds that increase in increments of 0.05 p; from no threshold to p >0.95 [9].

Task 2 –IrisPlex eye colour prediction from DNA of simulated casework samples without eye colour knowledge

The laboratories were provided with five DNA samples extracted from simulated casework samples (labelled CW1-CW5) from 5 individuals with blue (N=2) or brown (N=3) eye colour. DNA samples were extracted by the organising laboratory with the QIAamp DNA mini kit (Qiagen, Hagen, Germany). The following biological materials were used for DNA extractions: 2 buccal swabs (CW1 and CW2) – both samples subjected to UV radiation using the Bio-Link (Vilber Lourmat) for 1 min at a strength of 50 J/cm2 before DNA extraction; saliva on glass slide (CW3) stored at room temperature for 1 week before DNA extraction, whole blood on glass slide (CW4) stored at room temperature for 1 week before DNA extractions, and semen DNA extracted from a frozen donated sample (CW5). DNA concentra-tions were measured using the nanodrop system and 3 µL of the following solutions were provided: CW1: 0.5 ng/µL, CW2: 0.1 ng/ µL, CW3: 0.25 ng/µL, CW4: 2 ng/µL, and CW5: 50 ng/µL (see Table 2 for overview). The samples CW1–CW4 were freshly extracted, quantified and run using the IrisPlex system prior to DNA degradation, thereby serving as a control to their degraded counterparts. In contrast to Task 1, in Task 2 the laboratories were not provided with any eye colour phenotype information of the sample donors. Participants were also not

provided with any other sample information such as DNA concentration or treatment prior to DNA extractions of the DNA extracts they received for Task 2. They were asked to generate the IrisPlex genotype profile for each individual and report back the obtained eye colour probabilities and accuracies using the model and materials provided, and to finally conclude the most likely eye colour category per individual. An example report was also provided.

Sample #	Sample Type	Treatment	Concentration (ng/µl)
CW1	Buccal Swab	UV for 1 min	0.5
CW2	Buccal Swab	UV for 1 min	0.1
CW3	Saliva on Slide	RT for 1 week	0.25
CW4	Blood on Slide	RT for 1 week	2
CW5	Semen	-	50

Table 2: Details of the simulated casework samples provided to the participating laboratories for the Task 2.

Task 3 – Participant-driven IrisPlex eye colour prediction

This part of the study was optional. Each participating laboratory was instructed to collect and genotype samples from five different individuals of any eye colour. Selection of volunteers and biological materials was at the discretion of the participants. An important caveat in this task is that IrisPlex cannot guarantee a high prediction accuracy of the non-blue and non-brown eye colours; however, in contrast to Tasks 1 and 2 no eye colour phenotype restrictions were imposed on the participants in their choice of volunteers for Task 3. The laboratories were asked to report the DNA concentration, IrisPlex genotypes, eye colour probability outcomes and accuracy percentages, and a digital high resolution eye image of the genotyped individuals. The organising laboratory instructed that the iris photo should be taken in natural light conditions (no fluorescent bulb light) with and without flash lens using a digital camera focusing on eyes only (no full portrait).

Results and discussion

Sample extractions and quantifications by each participating laboratory

As the DNA extraction and quantification method can influence genotyping outcomes due to the quality and quantity of DNA extracted and consequently input into a downstream reaction, it can thus influence phenotype inference from the genotypes produced in FDP

systems. Therefore we included extraction and quantification monitoring in this exercise. As part of Task 1, the host laboratory provided the participating laboratories with biological samples (blood and saliva samples on FTA cards) from which the participants extracted and quantified DNA using their methods of choice. To note, the affiliated laboratory number in the author list does not represent the laboratory number described throughout the paper. Protocols used for DNA extraction and quantification were different and are listed in Table 1. As evident, the different extraction and quantification methods used by the participating laboratories provided varying results, as summarised in a box plot diagram (Figure 1), even though the same volume of biological sample was provided to each of the participants on FTA cards. Lab #2, 4, 13, 14 and 17 used the Qiagen EZ1 investigator kit for extraction and reported on average higher quantification values as compared to Lab #6, 9, 10, 12 and 15 that used the Qiagen QIAamp DNA mini kit extraction protocol. Lab #20 applied a Phenol-Chloroform extraction approach, which yielded on average higher quantification values compared to all the other methods used. Lab #18 used the Prepfiler Forensic DNA extraction kit and also displays on average higher final DNA amounts than all other methods, except Phenol-Chloroform. Lab #5, 7, 8, and 19, which used the Chelex extraction protocol, reported comparatively lower quantification values than all other methods used in this exercise. Worthy to note, this figure assumes that all laboratory extraction volumes were similar (i.e. the recommended 1 mL). The precise extraction volumes used by all labs were not available to the organising laboratory. This figure merely represents the differing extraction methods yielding varying final DNA concentrations, however, it is expected that the participants followed all recommendations provided by the organising lab which specifically states a 1 mL volume with at least a concentration of 32 pg, DNA input for IrisPlex profiling.. The DNA samples provided for Task 2 were previously extracted by the host laboratory using the QIAamp DNA mini kit (Qiagen). The participating laboratories were requested to measure DNA concentrations using their method of choice and to report back the values. Because different quantifica-tion methods were used, the obtained concentration estimates differed (Fig. 1 and Supplementary Table 1), similar to Task 1, even though equal aliquots of the very same DNA solutions per each sample were provided to each of the participants. As evident, sample CW2 was recorded as the most variable (0.01–2.61 ng/mL), which contradicts recorded measurements by the organising laboratory of 100 pg (Figure. 1).

Overall, the recorded DNA quantification data indicate that all samples shipped,

both the biological samples on FTA cards of Task 1 and the extracted DNA samples of Task 2, remained rather stable during transportation and short-term storage at the participating laboratories. For the impact of the varying amounts of DNA obtained by the participants in Task 1 and the varying DNA concentration measures obtained in Task 2 on genotype and phenotype accuracy, see the specific chapters on Tasks 1 and 2 below. From the DNA quantification data reported by the participating laboratories for the samples used in Task 3 (Supplementary Table 1) it is evident that all the samples genotyped for this portion of the exercise were of reasonable quantity. When conducting genotyping analyses and calling the peaks, the 50 rfu fluorescence threshold was set for calling alleles for a locus in all tasks and samples.

Task 1 - IrisPlex eye colour prediction from biological samples with eye colour knowledge

All participating laboratories reported the predicted eye colour and their probabilities in the format as requested by the organising laboratory. Figure 2 depicts the accurate genotype and eye colour phenotype calls for all the ten samples as obtained by the 21 participating laboratories. Supplementary Table 2 lists the genotypes of the ten individuals with their respective eye colour probability and accuracy. Figure 3 shows the eye colour images of the 10 individuals used in this task.

Twenty of the 21 laboratories (95%) predicted the eye colour of all 10 individuals included in Task 1 correctly from IrisPlex (Figure 2; green bars). Overall, 208 (99%) of the 210 samples analysed in this task by all the 21 laboratories were reported with the correct eye colour phenotype prediction. An overview of the samples with incorrect genotypes that were discordant with the organising laboratory is provided in Table 3. Only one laboratory (lab #3) faced difficulties in concluding the correct eye colour phenotype for two samples (Individual 1 and 10). The phenotype for both individuals was reported as inconclusive, although the correct IrisPlex genotypes were obtained and reported. These two individuals had eye colour probabilities for blue, intermediate and brown of 0.306, 0.142, 0.552 and 0.299, 0.253, 0.448 respectively (Figures 3(a) and 3(k)), and did not cause a problem for the other 20 laboratories to conclude the correct brown eye colour for both samples.

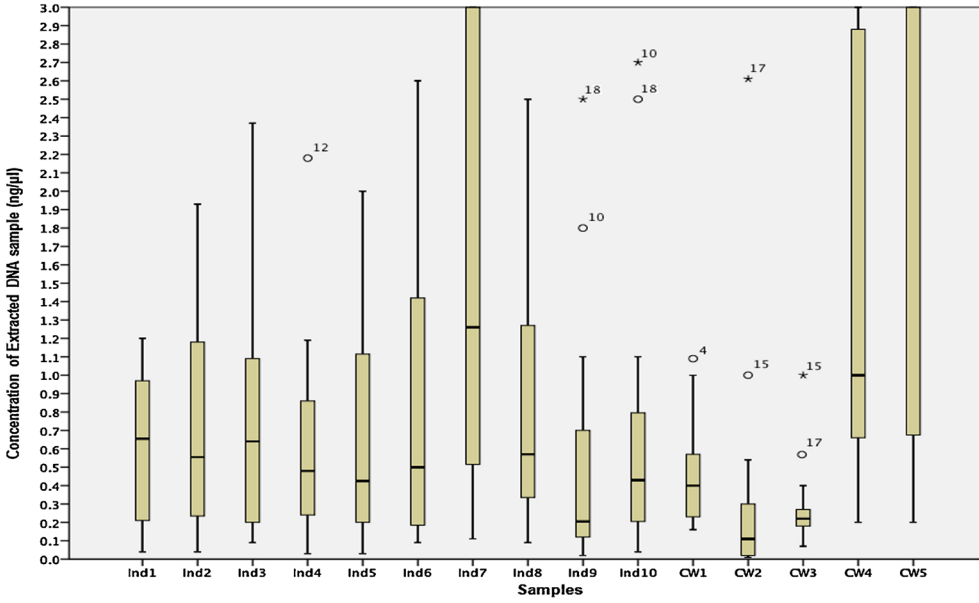
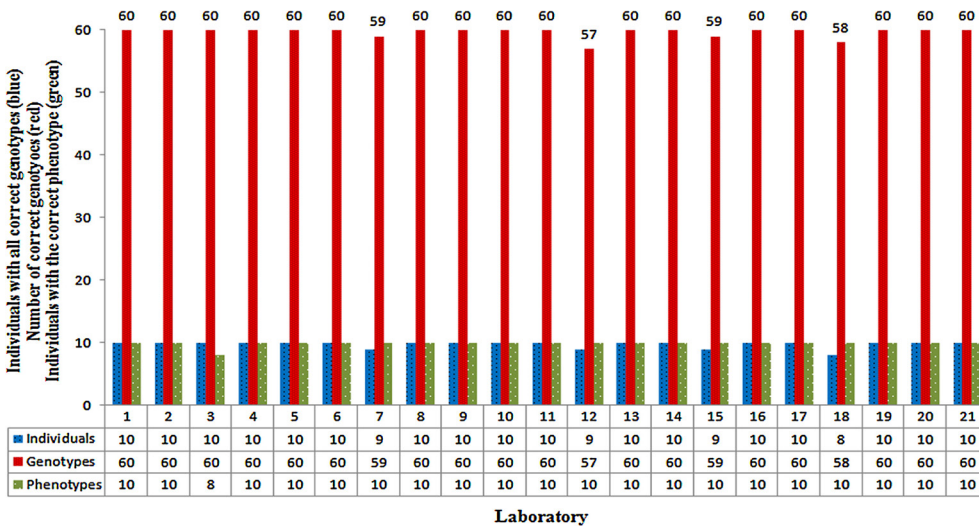


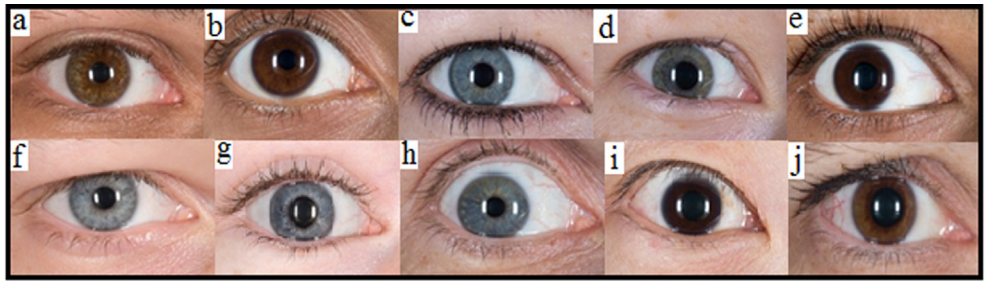
Figure 1: Box-Plot (scaled to 3ng/µL) showing the variation in obtained DNA concentrations using different methods for extraction and quantification between the 21 laboratories for the 15 samples used in Task 1 and 2. Note that for Task 1 (samples Ind1-Ind10), biological samples were provided to the participants so that DNA extraction and DNA quantifications were carried out by the participants on the same volumes of biological materials provided on FTA cards, while for Task 2 (samples CW1-CW5) participants were provided with already extracted DNA samples that varied in treatment and DNA concentrations. Medians are represented by the horizontal lines and the boxes depict the 25%-75% quartiles. The whiskers represent the minimum and the maximum values. Outliers are marked with the laboratory number where they were reported.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Individuals	10	10	10	10	10	10	9	10	10	10	10	9	10	10	9	10	10	8	10	10	10
Genotypes	60	60	60	60	60	60	59	60	60	60	60	57	60	60	59	60	60	58	60	60	60
Phenotypes	10	10	8	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10

Figure 2: Accuracy of the IrisPlex genotype calls (6 SNPs) and the IrisPlex-based eye colour phenotype prediction of the 10 samples provided in Task 1 as reported by each of the 21 participating laboratories. Blue indicates the number of individuals that were correctly genotyped at all 6 IrisPlex SNPs (i.e. for which a correct IrisPlex profile was reported). Red indicates the total number of genotypes across all 6 SNPs and all 10 samples that were correctly reported. Green indicates the number of individual samples for which the correct eye colour phenotype was reported.

Task 1



Task 2

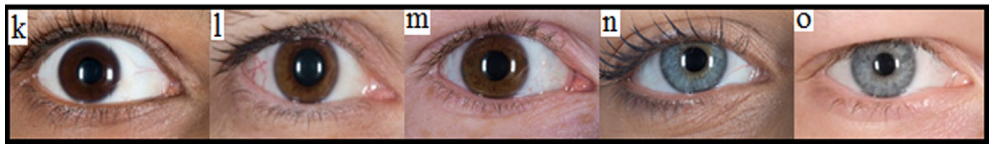


Figure 3: Eye colour images of the 10 individuals whose samples were used in Task 1 and the 5 individuals whose samples were used in Task 2.

	Sample	Lab#	Locus	Comments
Task 1	Individual 8	7	rs12203592	drop-out of C
	Individual 9	12	rs12913832	drop-in of C
	Individual 9	12	rs16891982	drop-in of G
	Individual 9	12	rs1393350	drop-in of T
	Individual 2	15	rs16891982	drop-out of C
	Individual 3	18	rs12203592	drop-in of T
	Individual 8	18	rs12203592	drop-out of T
Task 2	CW2	6	rs12913832	drop-out of T
	CW2	7	rs12913832	drop-out of C
	CW2	15	rs1393350	drop-in of T
	CW2, CW3	17	rs12913832	drop-out of T, C respectively
	CW2	17	rs1800407	drop-out of A
	CW3	18	rs1393350	drop-out of T
	CW2	21	rs12896399	drop-out of T

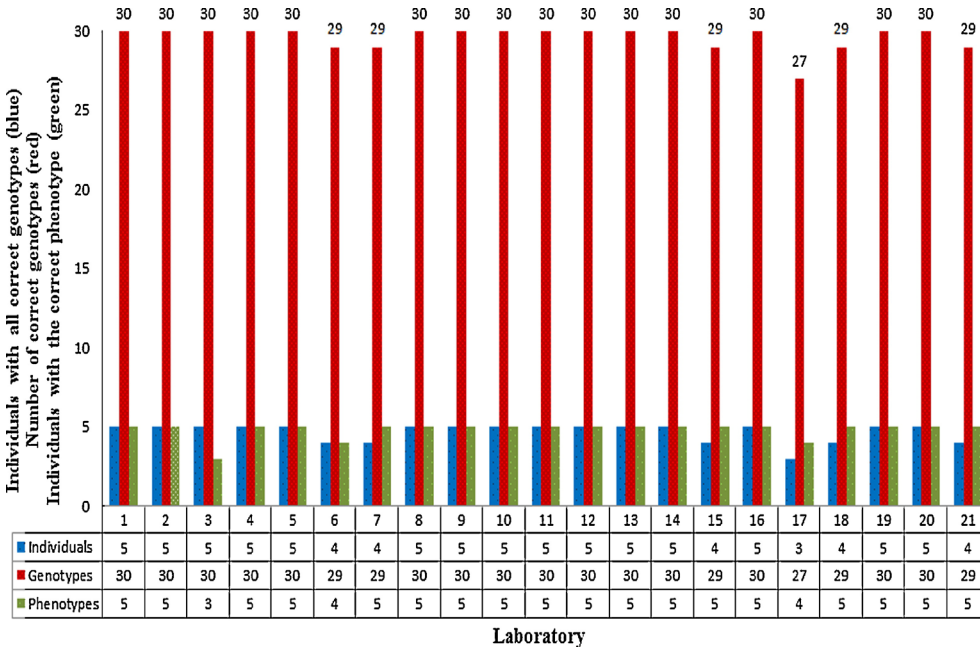
Table 3: Tasks 1 and 2 genotyping results that were discordant between the host and the participating laboratories.

Overall, 1253 (99.4%) of the 1260 genotypes generated for the 6 IrisPlex SNPs in the 210 samples analysed by all the 21 laboratories were reported correctly. The 7 (0.6%) incorrect genotypes were reported by 4 laboratories, while 17 participants reported the correct 6-SNP IrisPlex profiles for all 10 samples. Importantly, none of these incorrect genotypes led to erroneous eye colour phenotype predictions being reported. Lab #7 reported an incorrect homozygous genotype T instead of the true heterozygous CT for rs12203592 for Individual 8 due to a dropout of the C allele, since the respective peak was below the 50 rfu threshold. Lab #12 reported discordant heterozygous instead of correct homozygous genotypes for Individual 9 across three markers (rs12913832, rs16891982 and rs1393350). Drop-in of the alleles C, G and T was observed for each of the markers rs12913832, rs16891982 and rs1393350 respectively. However, the DNA concentration for this particular sample was reported by Lab #12 to be very low (0.02 ng/µL), much lower than the concentration obtained by the other laboratories for this sample (Supplementary Table 1). In accordance with the provided protocol, 1 µL of DNA solution of this sample was used for the IrisPlex PCR. Therefore, for this sample the amount of DNA input was below the previously established sensitivity threshold of the IrisPlex assay [10], which explains the high failure rate for genotypes of this sample. Lab #15 reported incorrect homozygous genotype of the G allele (instead of the correct heterozygous GC genotype) for individual 2 at rs16891982 due to a dropout of the C allele. Lab #18 experienced at marker rs12203592 a drop-in of allele T for individual 3 and dropout of allele T in individual 8. One explanation could be primer degradation due to incorrect storage of the primer or an incorrect volume addition of this primer to the assay, but unlikely due to a sample issue as the DNA concentrations reported for the individuals 2, 3 and 8 by the Lab #15 and 18 were more than 1.1 ng/µL (provided that the correct input of 1 µL was used). For cases such as these, an erroneous result can be avoided by re-running several analyses of the sample, and is usually recommended when using all genotyping platforms, this includes the IrisPlex system.

Task 2 - IrisPlex eye colour prediction from DNA of simulated casework samples without eye colour knowledge

In Task 2, the provided five DNA extraction aliquots from simulated casework samples (Table 2) were evaluated by each participating laboratory. Notably, the samples used for DNA extractions not only came from different biological sources (saliva, blood,

and semen) but also experienced different environmental conditions (UV radiation, storage at room temperature) and were provided to the participants in varying DNA concentrations (0.1 to 50 ng/µL), all unknown to the participants. Therefore, and due to the fact that no eye colour phenotypes were provided of the sampled individuals, this task was more challenging than Task 1. Figure 4 depicts the accurate genotype and eye colour phenotype calls for all the 5 samples as reported by the 21 participating laboratories. Supplementary Table 2 shows the reported genotypes of the five samples with their respective eye colour prediction probabilities and precision accuracy. Figure 3 shows the eye colour images of the 5 individuals used in this task. An overview of the samples with incorrect genotypes that were discordant with the organising and the other participating laboratories is provided in Table 3.



Laboratory	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Individuals	5	5	5	5	5	4	4	5	5	5	5	5	5	5	4	5	3	4	5	5	4
Genotypes	30	30	30	30	30	29	29	30	30	30	30	30	30	30	29	30	27	29	30	30	29
Phenotypes	5	5	3	5	5	4	5	5	5	5	5	5	5	5	5	5	4	5	5	5	5

Figure 4: Accuracy of the IrisPlex genotype calls (6 SNPs) and the IrisPlex-based eye colour phenotype prediction of the 5 samples provided in Task 2 as reported by each of the 21 participating laboratories. Blue indicates the number of individuals that were correctly genotyped at all 6 IrisPlex SNPs (i.e. for which a correct IrisPlex profile was reported). Red indicates the total number of genotypes across all 6 SNPs and all 5 samples that were correctly reported. Green indicates the number of individual samples for which the correct eye colour phenotype was reported.

Eighteen (86%) of the 21 laboratories predicted the eye colour of all 5 individuals correctly from IrisPlex (Figure. 4; green bars). Overall, 101 (96.2%) of the 105 samples analysed by all the 21 laboratories together were reported with the correct eye colour phenotype. The 4 samples (3.8%) for which the eye colour phenotypes were incorrect had been reported by 3 laboratories. Lab #3 predicted the eye colour of 2 of the 5 individuals (CW2 and CW3) as inconclusive, although the genotypes for these samples were reported correctly. Both samples clearly had to be designated as brown from the obtained probabilities (p=0.448 and p=0.552, respectively), and the phenotypes indeed were brown (Figure. 3(l) and (m) respectively), as was correctly interpreted by 18 other laboratories. The other 2 incorrectly phenotyped samples were reported by Labs #6 and 17 due to the drop-out of the T allele at rs12913832 in sample CW2. The 2 laboratories reported an incorrect homozygous C allele instead of a heterozygous CT allele, thereby, reporting an incorrect blue eye colour instead of the correct brown eye colour phenotype (Figure. 3(l)).

Overall, 622 (98.7%) of the 630 genotypes generated for the 6 IrisPlex SNPs in the 105 samples analysed by all laboratories together were correctly reported in Task 2. The 8 (1.3%) incorrect genotypes were produced in 2 samples (CW2 and CW3) by 6 laboratories, while 15 of the 21 laboratories (71.4%) reported the correct 6-SNP IrisPlex profile for all 5 samples. In contrast to the 2 (25%) genotype errors in sample CW2 by Lab #6 and 17 that caused phenotype errors as mentioned in the previous paragraph, the remaining 6 incorrect genotypes (75%) did not have any impact on the eye colour phenotype accuracy. At rs12913832, Lab #7 reported an incorrect genotype due to a drop-out of the C allele for sample CW2. A drop-in of the T allele for CW2 and drop-out of the T allele for CW3 at rs1393350 resulted in incorrect genotyping by Lab #15 and 18 respectively. Furthermore, incorrect genotypes were reported by Lab #21 at rs12896399 for sample CW2 due to the drop-out of the T allele. Lab #17 experienced problems in the first typing of samples CW1, 2 and 3 and subsequently retyped these samples in different DNA dilutions. At rs12913832, drop-out of the T allele for CW2 (as mentioned above); drop-out of the C allele for CW3, and drop-out of the A allele at rs1800407 for CW2 were reported which resulted in erroneous results for this laboratory. A dilution step performed by the participating laboratory, due to a misleading quantification result, of the already low quantity degraded samples provides a likely explanation for the drop-out of the alleles in this set of samples.

Several laboratories (n=3; Lab #3, 6 and 17) experienced difficulties with correct phenotyping of the simulated and treated casework samples in Task 2 for which no eye colour phenotypes were provided as opposed to the untreated biological samples provided together with eye colour phenotypes in Task 1 (n=1; Lab #3). Similarly, more laboratories (n=6; Lab# 6, 7, 15, 17, 18 and 21) had difficulties in correct genotyping of Task 2 samples in relation to Task 1 samples (n=4; Lab #7, 12, 15 and 18). Within Task 2, the most genotyping and phenotyping difficulties i.e. allelic drop-outs and drop-ins were reported for 2 particular samples (CW2 and CW3). Sample CW2 was reported with different incorrect genotypes by 5 of the laboratories (Lab #6, 7, 15, 17 and 21) and sample CW3 was reported incorrectly by 2 laboratories (Lab #17 and 18) (see Table 3 for overview). Sample CW2 must therefore be noted as being a difficult sample to genotype. From Figure 1, it is evident that, of the laboratories that reported quantification data for Task 2, sample CW2 was recorded as the most variable (0.01 ng/µL – 2.61 ng/µL), which strongly deviates from the recorded measurements by the organising laboratory of 100 pg. Given its unusual quantification range, severe degradation and heterozygosity at 3 (rs12913832, rs1800407 and rs12896399) of the 6 SNPs, increased incidence of allelic drop-out may be expected in sample CW2 as compared to the homozygous sample CW1 (which also experienced UV degradation) that caused no problems for genotyping. This demonstrates, as expected and as also known for any other genotyping assay, that the combination of low quality and low quantity template DNA provides challenges for correct genotyping including for the IrisPlex assay. However, it should be emphasized that 244 of the 252 (96.8%) genotypes of the most challenging samples CW2 and CW3 were generated correctly by 15 of the 21 (71.4 %) participating laboratories, which demonstrates the reliability of the IrisPlex assay for difficult DNA samples. This also represents the necessity of employing duplicate analysis when genotyping samples of low DNA quantity in final case work applications.

Task 3 –Participant-driven IrisPlex testing

The optional Task 3 of the exercise, where participants were asked to recruit their own volunteers for IrisPlex genotyping and eye colour prediction, was performed by 20 of the 21 laboratories. Lab #7 could not perform this task due to reported ethical issues. Supplementary Table 3 summarises the data for this task. Based on the digital eye images provided by the participants (Figure 5), the organising laboratory judged the correct

phenotypes by two independent experienced observers. Lab #12 performed this exercise and reported the genotype and phenotype, but provided no eye images to the host laboratory for inspection. As it was not possible to judge the accuracy of the results provided by this participant, they were excluded from the analyses. As can be seen in Figure 6; 16 of the 19 laboratories (84.2%) predicted the eye colour of all analysed individual samples correctly, while 3 laboratories faced difficulties in concluding the correct eye colour from the estimated probability combinations for some samples. Overall, 96 (96%) of the 100 samples analysed by the 19 laboratories were reported with the correct eye colour prediction, as judged by the organising laboratory based on the digital eye images sent by the participants. The 4 samples reported with incorrect eye colour were from 3 different laboratories (1x#2, 2x#3, and 1x#5). Lab #2 reported blue eye colour (p=0.678) for their sample 1 but the eye image showed brown colour and the estimated brown eye probability was only 0.191 (Figure 5(a)). Lab #3 obtained the following probabilities for their sample 1: Blue-0.207, Intermediate-0.161 and Brown-0.632 and reported an inconclusive result, while the probability for brown was by far higher than for the other two categories so that brown should have been concluded instead and indeed the respective eye image showed brown (Figure 5(b)). Sample 4 of Lab #3 appeared blue from the images (Figure 5(c)) but a high brown eye probability (p=0.892) was obtained while the blue eye probability was low (p=0.024). Lab #5 obtained probabilities of Blue – 0.375, Intermediate – 0.264 and Brown – 0.361 for their sample 1, and reported blue eye colour but the image indicates brown eye colour (Figure 5(d)). In this case, however, the eye colour could have been reported inconclusive since the brown and the blue eye colour probabilities were very similar. It is therefore important to use and report the level of precision accuracy based on each probability threshold with the final prediction, i.e. $p < 0.5$ highest probability value, $p > 0.5$ highest probability value. This can be found in Table 2 of our previous publication [9]. It is noteworthy to emphasise that the IrisPlex genotypes in Task 3 were not verified independently in contrast to those in Task 1 and 2. Therefore we cannot know for sure if any of the 4 incorrect phenotype predictions in Task 3 may have been caused by incorrect genotypes, although the high genotyping accuracy rates achieved in Task 1-3 suggest this might be somewhat unlikely. Due to violation of anonymity, the geographic origin of these individuals cannot be determined.

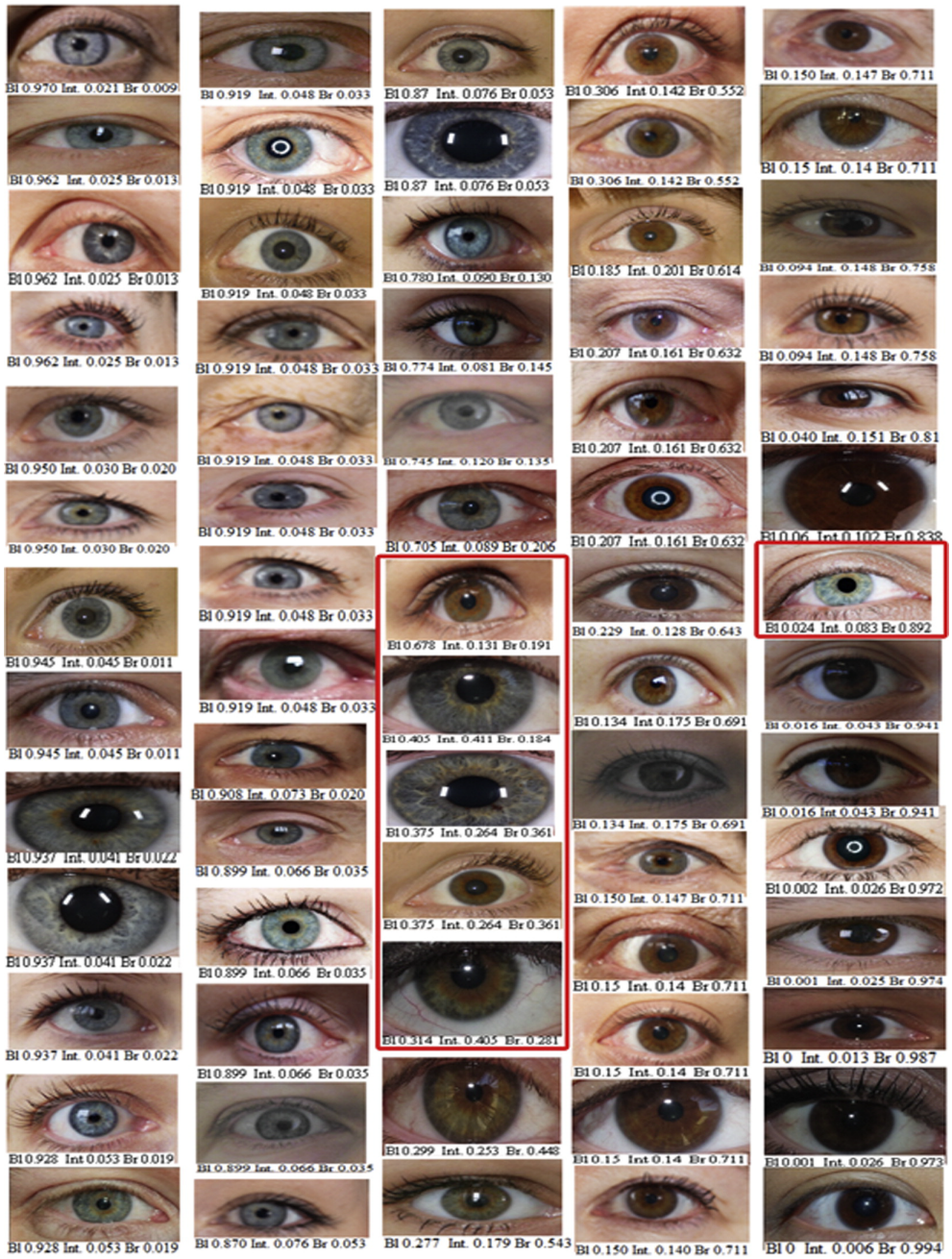


Figure 5: Eye images sent by the participating laboratories used for the voluntary aspect of the study, Task 3. Eye images include probability values for blue, intermediate and brown eye colour provided by the participants as determined from the IrisPlex genotypes. The area surrounded by the red lines indicates the incorrect eye colour prediction as assessed by the host laboratory from inspection of the eye images provided, and compared with the eye colour phenotype reported by the participants based on IrisPlex analysis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

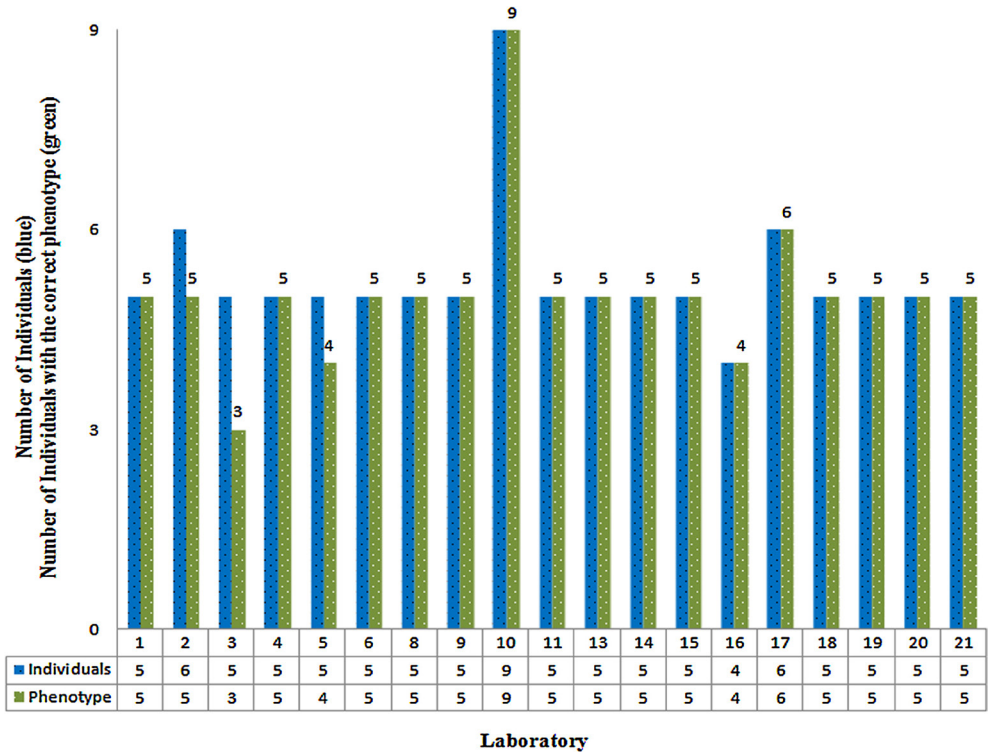


Figure 6: Accuracy of the IrisPlex-based eye colour phenotype prediction of the samples selected by all 19 laboratories participating in Task 3 (Lab #7 did not participate in this task). Lab #12 though performed the task, did not include images and hence was not considered here. The correct eye colour phenotype was assessed by the host laboratory from inspection of the eye images provided, and compared with the eye colour phenotype reported by the participants based on IrisPlex analysis. Blue indicates the number of individuals that were genotyped and green indicates the number of individuals for which the correct phenotype was reported. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

Further to note, participants for Task 3 were not asked to restrict their choice of volunteers to blue and brown eye colour only. This was different from Tasks 1 and 2 where only volunteers with blue and brown eye colour were used due to known limitations of the IrisPlex system to accurately predict non-blue and non-brown eye colours [8–10]. However only Lab #1 reported 2 individuals as intermediate ($p = 0.411$ and $p = 0.405$) and from the eye images (Figures. 5(e) and (f) respectively), we can confirm that the individuals were correctly predicted as intermediate as they contain substantial pupillary rings of a different colour (i.e. majority of iris blue colour with obvious brown pupillary ring). It is also worth noting that although no restrictions were imposed on the choice of samples for Task 3, all the laboratories (except Lab #1) chose individuals with either blue or brown eyes and hence

it is most likely that all the participants were guided by knowledge of (or clearly considered the) limitations of IrisPlex for accurately predicting non-blue and non-brown eye colour phenotypes. Furthermore, according to general knowledge, the frequency of blue and/or brown eyes is comparatively higher than intermediate in the population, which could explain the rarity of volunteers with intermediate eye colour phenotype used in Task 3. The overall performances of the participating laboratories in all 3 tasks, is shown in Supplementary Table 4.

Conclusions

Overall, the high level of consistency achieved throughout this collaborative effort in all 3 tasks illustrates the reliability of the IrisPlex assay in producing highly accurate 6-SNP genotypes and of the IrisPlex prediction model in producing accurate blue and brown eye colour phenotypes from IrisPlex genotypes. As shown here and previously [8], the IrisPlex assay provides reproducible results despite differing levels of experience of the laboratory personnel involved and differing DNA extraction and quantifica-tion methods used. The results obtained in this collaborative exercise demonstrate the robustness and reproducibility of DNA-based eye colour prediction when using the IrisPlex system in different forensic laboratories world-wide. As emphasised before [8–10], future focus shall be placed on improving DNA based prediction of non-blue and non-brown eye colours, for which the IrisPlex system is less suitable than for blue and brown eye colour prediction from DNA.

Acknowledgements

We would like to thank Gabriela Huber, Nelli Roininen and Mintu Hedman for outstanding technical assistance and Vlastimil Stenzl for support. This work was funded in part by the Netherlands Forensic Institute (NFI) and a grant from the Netherlands Genomic Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands (FGCN). Funding for Adrian Linacre and Renee Ottens was provided by the Attorney General's Office, South Australia, for Theresa Gross by the EU 7th Framework Programme (FP7/2007–2013) under Grant Agreement N8 285487 (EUROFORGEN-NoE), and for Antti Sajantila by the Finnish Foundations' Pool Professorship (Paulo Foundation).

References

[1] Kayser M, Schneider PM, DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical consid-erations, Forensic Sci. Int. Genet. 3 (June (3)) (2009) 154–161.

[2] Kayser M, de Knijff P, Improving human forensics through advances in genetics, genomics and molecular biology, Nat. Rev. Genet. 12 (March (3)) (2011) 179–192.

[3] Han J, Kraft P, Nan H, Guo Q, Chen C, Qureshi A, Hankinson SE, Hu FB, Duffy DL, Zhao ZZ, Martin HG, Montgomery GW, Hayward NK, Thomas G, Hoover RN, Chanock S, Hunter DJ, A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation, PLoS Genet. 4 (May (5)) (2008) e1000074.

[4] Ruiz Y, Phillips C, Gomez-Tato A, Alvarez-Dios J, Casares de Cal M, Cruz R, Maronas O, Sochtig J, Fondevila M, Rodriguez-Cid MJ, Carracedo A, Lareu MV, Further development of forensic eye color predictive tests, Forensic Sci. Int. Genet. 7 (January (1)) (2013) 28–40.

[5] Fortes GG, Speller CF, Hofreiter M, King TE, Phenotypes from ancient DNA: approaches, insights and prospects, Bioessays 35 (August (8)) (2013) 690–695.

[6] Hart KL, Kimura SL, Mushailov V, Budimlija ZM, Prinz M, Wurmbach E, Improved eye- and skin-color prediction based on 8 SNPs, Croat. Med. J. 54 (June (3)) (2013) 248–256.

[7] Liu F, van Duijn K, Vingerling JR, Hofman A, Uitterlinden AG, Janssens AC, Kayser M, Eye color and the prediction of complex phenotypes from genotypes, Curr. Biol. 19 (March (5)) (2009) R192–R193.

[8] Walsh S, Liu F, Ballantyne KN, 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. 5 (June (3)) (2011) 170–180.

[9] Walsh S, Wollstein A, Liu F, Chakravarthy U, Rahu M, Seland JH, Soubrane G, Tomazzoli L, Topouzis F, Vingerling JR, Vioque J, Fletcher AE, Ballantyne KN, Kayser M, DNA-based eye colour prediction across Europe with the IrisPlex system, Forensic Sci. Int. Genet. 6 (May (3)) (2012) 330–340.

[10] Walsh S, Lindenbergh A, Zuniga SB, Sijen T, de Knijff P, Kayser M, Ballantyne KN, Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence, Forensic Sci. Int. Genet. 5 (November (5)) (2011) 464–471.

[11] Purps J, Geppert M, Nagy M, Roewer L, Evaluation of the IrisPlex eye color prediction tool in a German population sample, Forensic Sci. Int. Genet. 3 (December (1)) (2011) 202–203.

[12] Martinez-Cadenas C, Pen˜a-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. 7 (July (4)) (2013) 453–460.

[13] Prestes PR, Mitchell RJ, Daniel R, Ballantyne KN, van Oorschot RAH, Evalua-tion of the IrisPlex system in admixed individuals, Forensic Sci. Int. Genet. 3 (December (1)) (2011) e283–e284.

[14] Dembinski GM, Picard CJ, Evaluation of the IrisPlex DNA-based eye color prediction assay in a United States population, Forensic Sci. Int. Genet. 9 (March) (2014) 111–117.

[15] Yun L, Gu Y, Rajeevan H, Kidd KK, Application of six IrisPlex SNPs and comparison of two eye color prediction systems in diverse Eurasia populations, Int. J. Legal Med. (January) (2014),

[16] Kastelic V, Pospiech E, Draus-Barini J, Branicki W, Drobnic K, Prediction of eye color in the Slovenian population using the IrisPlex SNPs, Croat. Med. J. 54 (August (4)) (2013) 381–386.

[17] 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. 7 (January (1)) (2013) 98–115.

Supplementary Material

Quantification Values (ng/µl)																										
	Task 1										Task 2					Task 3										
Lab ID	Ind 1	Ind 2	Ind 3	Ind 4	Ind 5	Ind 6	Ind 7	Ind 8	Ind 9	Ind 10	CW 1	CW 2	CW 3	CW 4	CW 5	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9		
1	0,22	0,28	0,13	0,19	0,03	0,50	0,50	0,50	0,50	0,50						0,50	0,50	0,50	0,50	0,50						
2	0,86	0,93	0,76	0,64	0,58	0,35	0,53	0,70	0,74	0,87	0,16	0,01	0,18	0,66	0,65	0,79	0,79	0,50	1,16	1,20						
3	0,04	0,04	0,09	0,03	0,05	0,12	1,29	0,09	0,04	0,08	0,23	0,01	0,07	0,75	17,00	4,53	6,65	2,30	4,00	2,73						
4	0,58	0,55	0,54	0,31	0,36	1,54	3,10	2,19	0,60	0,46	1,09	0,14	0,23	2,17	49,44	1,43	8,23	4,53	2,62	3,25						
5	0,44	0,26	0,33	0,30	0,20	0,38	0,16	0,36	0,12	0,33	0,57	0,54	0,22	4,39	111,90	0,11	0,37	0,16	0,13	<0.100						
6	0,13	0,21	0,20	0,29	0,33	0,54	0,54	0,53	0,66	0,54						0,54	0,61	0,57	0,52	0,48						
7	0,30	0,33	0,32	0,28	0,40	0,19	0,73	0,22	0,09	0,40																
8	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20	0,20						
9	0,05	0,12	0,17	0,12	0,19	0,35	0,68	0,28	0,09	0,09						0,30	0,16	0,26	0,21	0,12						
10	1,20	0,60	0,80	0,80	2,00	2,60	2,80	2,50	1,80	2,70						3,9	3,3	2,9	3,5	3	2,4	2,5	3,3	3,2		
11	0,19	0,12	0,16	0,07	0,07	0,18	0,36	0,31	0,07	0,07						1	0,42	1,1	0,5	0,37						
12	3,80	1,93	2,37	2,18	1,03	0,11	11,17	1,32	0,02	0,04	0,28	0,01	0,15	2,88	62,36	0,5	0,2	0,01	0,1	0,05	1,000	30,95	6,060			
13	0,83	1,36	1,08	1,19	1,94	0,50	3,48	0,42	0,20	0,24	0,60	0,11	0,27	5,12	Undetermined	15,62	9,57	11,49	30,72	5,07						
14	1,04	1,26	1,18	0,42	0,94	1,61	3,39	1,22	0,19	0,26						12,62	8,34	8,2	4,99	9,76						
15	0,90	1,10	1,10	0,90	1,20	1,30	1,40	1,10	1,10	1,10	1,00	1,00	1,00	1,00	1,00	1	1,2	1,2	1,1	1						
16																										
17	0,72	0,56	0,56	0,60	0,45	0,50	0,58	0,61	0,61	0,59	0,46	2,61	0,58	0,59	0,51	0,640	0,708	0,747	0,678	0,784	0,696					
18	10,00	10,00	10,00	5,00	10,00	5,00	5,00	2,50	2,50	2,50	0,50	0,30	0,40	0,50	0,70											
19	0,78	0,78	0,72	0,82	0,88	0,09	1,23	0,41	0,21	0,21	0,28	0,02	0,26	1,00	0,98	0,36	0,44	0,70	0,74	0,61						
20	4,50	8,78	10,02	6,37	5,94	12,95	89,63	16,23	1,08	27,79	0,16	0,02	0,17	2,90	81,08	26,23	11,53	20,11	1,86	9,53						
21	0,59	0,53	1,00	0,54	0,35	0,52	1,60	0,76	0,12	0,72	0,40	0,03	0,20	2,00	50,00	0,53	0,53	0,46	0,46	0,46						
Average Quantification Values (ng/µl)	1,37	1,50	1,59	1,06	1,36	1,48	6,42	1,62	0,55	1,98	0,46	0,39	0,30	1,86	31,32	3,93	2,99	3,11	3,00	2,30	1,37	16,73	4,68	3,20		

Supplementary Table 1: Quantification values in (ng/µL) for Task 1, Task 2 and Task 3 as reported by the laboratories. Grey boxes indicate no data received from the participating laboratories. For Task 3, some of the laboratories reported the data from more than the five requested samples.

TASK 1

	Genotype						Probabilities			Final colour prediction and accuracy based on a European dataset.
	rs12913832	rs1800407	rs12896399	rs16891982	rs1393350	rs12203592	Blue	Int.	Brown	
Ind. 1	C/T	G/G	T/T	G/G	C/T	C/C	0.306	0.142	0.552	The most probable eye colour of this individual is brown with 91% accuracy
Ind. 2	C/T	G/G	G/G	G/C	C/C	C/C	0.024	0.083	0.892	The most probable eye colour of this individual is brown with 95.58% accuracy
Ind. 3	C/C	G/G	T/T	G/G	C/T	C/C	0.962	0.025	0.013	The most probable eye colour of this individual is blue with 99% accuracy
Ind. 4	C/C	G/G	G/T	G/G	C/C	C/C	0.919	0.048	0.033	The most probable eye colour of this individual is blue with 97.4% accuracy
Ind. 5	T/T	G/G	G/G	C/C	C/C	C/C	0	0.006	0.994	The most probable eye colour of this individual is brown with 99% accuracy
Ind. 6	C/C	G/G	G/G	G/G	C/C	C/C	0.87	0.076	0.053	The most probable eye colour of this individual is blue with 95.58% accuracy
Ind. 7	C/C	G/G	G/T	G/G	C/T	T/T	0.944	0.049	0.007	The most probable eye colour of this individual is blue with 97.4% accuracy
Ind. 8	C/C	G/G	T/T	G/G	C/T	C/T	0.965	0.027	0.007	The most probable eye colour of this individual is blue with 99% accuracy
Ind. 9	T/T	G/G	G/T	C/C	C/C	C/C	0	0.006	0.994	The most probable eye colour of this individual is brown with 99% accuracy
Ind. 10	C/T	G/A	G/T	G/G	C/C	C/C	0.299	0.253	0.448	The most probable eye colour of this individual is brown with 87.5% accuracy

TASK 2

	Genotype						Probabilities			Final colour prediction and accuracy based on a European dataset.
	rs12913832	rs1800407	rs12896399	rs16891982	rs1393350	rs12203592	Blue	Int.	Brown	
CW 1	T/T	G/G	G/G	C/C	C/C	C/C	0	0.006	0.994	The most probable eye colour of this individual is brown with 99% accuracy
CW 2	C/T	G/A	G/T	G/G	C/C	C/C	0.299	0.253	0.448	The most probable eye colour of this individual is brown with 87.5% accuracy
CW 3	C/T	G/G	T/T	G/G	C/T	C/C	0.306	0.142	0.552	The most probable eye colour of this individual is brown with 91% accuracy
CW 4	C/C	G/G	G/T	G/G	C/C	C/C	0.919	0.048	0.033	The most probable eye colour of this individual is blue with 97.4% accuracy
CW 5	C/C	G/G	G/G	G/G	C/C	C/C	0.87	0.076	0.053	The most probable eye colour of this individual is blue with 95.58% accuracy

Supplementary Table 2: The results from the Tasks 1 and 2 as reported by the organising laboratory. The correct genotypes and the predicted eye colour with their probabilities and the respective accuracies are detailed in the table.

Lab ID	Polymer	Genetic Analyzer	Samples with correct genotype assignments of all 6 SNPs (Task 1 and 2)	Samples with correct phenotype (Task 1 and 2)
1	POP 4	3130xl	15	15
2	POP 6	3100	15	15
3	POP 4	3130	15	11
4	POP 4	3130xl	15	15
5	POP 4	3130xl	15	15
6	POP 4	3500xl	14	14
7	POP 7	3130xl	13	15
8	POP 4	3130	15	15
9	POP 7	3500	15	15
10	POP 7	3500	15	15
11	POP 4	3130xl	15	15
12	POP 7	3130xl	14	15
13	POP 7	3130xl	15	15
14	POP 4	3130xl	15	15
15	POP 4		6	15
16	POP 7	3500	15	15
17	POP 4	3130	13	15
18	POP 4	3130	12	14
19	POP 7	3130	15	15
20	POP-4	3130	15	15
21	POP 7	3130xl	14	15

Supplementary Table 4: The overall performances of the 21 participating laboratories and the equipment used for the exercise. Grey boxes indicate no data received from the participating laboratory.

Lab#	Number of individuals	Individuals with correct phenotype	Comments
1	5	5	
2	6	5	For sample 1, image appears brown but probability indicates blue
3	5	3	Reported inconclusive for sample 1 which clearly apeared brown from the image and the probability value. Sample 4 appeared blue but the probability indicated brown.
4	5	5	
5	5	4	For sample 1, image appears brown but reported blue based on probability
6	5	5	
7	-	-	Did not participate in this task
8	5	5	
9	5	5	
10	9	9	
11	5	5	
12	8	-	No images included
13	5	5	
14	5	5	
15	5	5	
16	4	4	
17	6	6	
18	5	5	
19	5	5	
20	5	5	
21	5	5	

Supplementary Table 3: The results of Task 3 are summarised in this table. The correct phenotype was judged by the organising laboratory based on the images sent by the participating laboratories.

Chapter 3

The HIrisPlex system: DNA-based hair colour prediction tool

Chapter 3.1

Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage

Forensic Science International: Genetics 9 (2014) 150 – 161

Abstract

Forensic DNA Phenotyping or 'DNA intelligence' tools are expected to aid police investigations and find unknown individuals by providing information on externally visible characteristics of unknown suspects, perpetrators and missing persons from biological samples. This is especially useful in cases where conventional DNA profiling or other means remain non-informative. Recently, we introduced the HIrisPlex system, capable of predicting both eye and hair colour from DNA. In the present developmental validation study, we demonstrate that the HIrisPlex assay performs in full agreement with the Scientific Working Group on DNA Analysis Methods (SWGDAM) guidelines providing an essential prerequisite for future HIrisPlex applications to forensic casework. The HIrisPlex assay produces complete profiles down to only 63 pg of DNA. Species testing revealed human specificity for a complete HIrisPlex profile, while only non-human primates showed the closest full profile at 20 out of the 24 DNA markers, in all animals tested. Rigorous testing of simulated forensic casework samples such as blood, semen, saliva stains, hairs with roots as well as extremely low quantity touch (trace) DNA samples, produced complete profiles in 88% of cases. Concordance testing performed between five independent forensic laboratories displayed consistent reproducible results on varying types of DNA samples. Due to its design, the assay caters for degraded samples, underlined here by results from artificially degraded DNA and from simulated casework samples of degraded DNA. This aspect was also demonstrated previously on DNA samples from human remains up to several hundreds of years old. With this paper, we also introduce enhanced eye and hair colour prediction models based on enlarged underlying databases of HIrisPlex genotypes and eye/ hair colour phenotypes (eye colour: N=9188 and hair colour: N=1601). Furthermore, we present an online web-based system for individual eye and hair colour prediction from full and partial HIrisPlex DNA profiles. By demonstrating that the HIrisPlex assay is fully compatible with the SWGDAM guidelines, we provide the first forensically validated DNA test system for parallel eye and hair colour prediction now available to forensic laboratories for immediate casework application, including missing person cases. Given the robustness and sensitivity described here and in previous work, the HIrisPlex system is also suitable for analysing old and ancient DNA in anthropological and evolutionary studies.

Introduction

When investigating a case that involves DNA, current forensic practice calls for short tandem repeat (STR) DNA analysis to identify the donor of a biological sample found at a crime scene. However, in certain circumstances an STR profile obtained from evidence material does not match that of a known suspect including any from a criminal DNA database, which often leads to progress in the investigation being halted (so-called cold cases). Due to this, and by taking advantage of scientific progress in the genetic understanding of human appearance, advancement has been made in the DNA prediction of externally visible characteristics (EVCs) [1–18]. This concept of DNA intelligence, also termed Forensic DNA Phenotyping (FDP), adds a new dimension to forensic DNA analysis, and provides a promising alternative to help with future investiga-tions in cases where conventional STR profiling fails to identify a crime scene sample donor [19–21]. At this moment, only eye and hair colour DNA prediction is feasible with accuracies high enough to base police investigations on, while other group-specific EVCs such as skin colour, male baldness, hair morphology are under investigation to identify the underlying DNA variants and to estimate their phenotype predictive value (for recent reviews see [20,21]).

DNA prediction of eye colour is now achievable when it comes to broad categories such as blue or brown, with systems such as IrisPlex that uses model-based probability prediction [22–24] and that have been extensively tested [24] and forensically validated [25]. Liklihood association ratio methods [9], a bayesian classifier approach based on likelihood ratios [13] or lastly a prediction guide process [11], amongst others also currently exist for eye colour DNA prediction. Furthermore, DNA prediction of hair colour is already feasible via the HIrisPlex system [26] for parallel prediction of hair and eye colour from DNA, generating leading intelligence for two of the most obvious externally visible traits. Combining our previous eye colour prediction system IrisPlex with knowledge from early systems used for hair colour prediction [1,10,27], and considering knowledge from our previous study where we established the hair colour predictive value of a larger number of hair colour associated DNA variants [28], the HIrisPlex system was finally developed [26]. The HIrisPlex system consists of a single multiplex genotyping assay targeting 24 DNA variants identified in European population studies to be highly informative for eye and hair colour prediction [22,28], as well as two statistical prediction models, one for eye colour and one for hair colour

and shade. These prediction models were previously developed from thousands (eye colour) [22] and hundreds (hair colour) [26] of individual genotype and phenotype datasets from European populations. Combined in this highly effective multiplex genotyping assay are the six most eye colour informative SNPs previously used in the IrisPlex system [23–25] and the twenty-two most hair colour informative DNA variants previously identified as carrying the most hair colour predictive information from a larger number of hair colour associated DNA variants [22,23,26,28]. According to previously established knowledge [22–24], the use of the IrisPlex eye colour prediction model within the IrisPlex or the HIrisPlex system can correctly predict human blue and brown eye colour with >90% precision, and the use of the HIrisPlex hair colour and shade prediction model and guide can predict on average hair colour accuracies of 79% [26].

In the present study, we performed the developmental validation of the HIrisPlex genotyping assay following the Scientific Working Group on DNA Analysis Methods (SWGDAM) guidelines [29]. These guidelines allow an assessment of the assay for use in certified forensic laboratories. It examines the assay's performance quality and limitations under differing conditions such as sensitivity, reproducibility and concordance as well as non-human species amplification, mixtures, degraded DNA and simulated casework samples. In completing these evaluations, and by demonstrating that the HIrisplex assay fully meets all SWGDAM requirements, we provide a necessary step towards the implementation of the HIrisPlex system in forensic (and other) laboratories to be used for enhancing police investigations in suitable cases. We also introduce here an enhanced prediction model for hair colour, which includes additional individuals to the underlying prediction databases relative to our previously intro-duced eye and hair colour prediction models. Furthermore, we developed an online web-based tool freely available under www.erasmusmc.nl/fmb/resources for eye and hair colour DNA prediction from complete and partial IrisPlex/HIrisPlex genotype profiles.

Materials and Methods

Human samples

A selection of human body fluid and tissue samples were collected from donors with informed consent. The hair and eye colour phenotype of these sample donors was also recorded. Test samples included single and multiple source samples, and simulated casework samples (blood, saliva, semen, hair with roots, touched items). DNA was extracted from the samples using the QIAamp DNA Mini kit (Qiagen, Hagen, Germany) according to the manufacturer's guidelines or an in-house extraction protocol (unpublished) and quantified using the Quantifiler Human DNA Quantification Kit (Applied Biosystems Inc., Foster City, USA) following manufacturers guidelines.

Multiplex Protocol

As outlined previously [26], the HIrisPlex assay consists of 24 SNPs, 6 of these make up the eye colour prediction portion of the tool, also termed IrisPlex [22–25], these are rs12913832 (*HERC2*), rs1800407 (*OCA2*), rs12896399 (*SLC24A4)*, rs16891982 (*SLC45A2 (MATP)*), rs1393350 (*TYR*) and rs12203592 (*IRF4*), while the other 18 (including IrisPlex SNPs 1, 2, 4 and 6), are used for hair colour and shade prediction. These 18 DNA variants are 1 insertion/deletion (INDEL) polymorphism N29insA and 10 SNPs from the MC1R gene, rs11547464, rs885479, rs1805008, rs1805005, rs1805006, rs1805007, rs1805009, Y152OCH, rs2228479, and rs1110400, rs28777 (*SLC45A2*), rs12821256 (*KITLG*), rs4959270 (*EXOC2*), rs1042602 (*TYR*), rs2402130 (*SLC24A4*), rs2378249 (*ASIP/PIGU*), and rs683 (*TYRP1*). All marker details and primer sequences, including a redesign of the extension primer for N29insA and slight alterations to the previously published protocols concentrations [26], can be found in Table 1. The protocol consists of a single multiplex two step PCR using 1 µl genomic DNA extract (varying concentrations) and primers in a 10 µl reaction which includes 1X PCR buffer, 2.5 mM MgCl2, 220 µM of each dNTP and 1.75 U Amplitaq Gold DNA polymerase following published thermocycling conditions [26]. This is followed by product purification and a further multiplex single base extension (SBE) reaction using 2 µl cleaned product with 1 µl ABI Prism SNaPshot chemistry (Applied Biosystems) using primer concentrations found in Table 1 and our previously published thermocycling conditions [26]. Lastly, all cleaned products were analysed on the ABI 3130xl Genetic Analyser

(Applied Biosystems) with POP-7 on a 36 cm capillary length array. Run parameters were optimised to increase sensitivity, with an injection voltage of 2.5 kV for 10 s, and run time of 500 s at 60°C.

Assay Position	SNP	Chr Pos			Gene	Major allele	Minor allele	PCR Primers		Concentration		Product Size	SBE Primers	Concentration
1	N29insA	16	89985753	Exonic	MC1R	C	insA	MC1Rset1F	Set 1	0.5µm[a]	GCAGGGATCCCAGAGAAGAC	117bp	CCAGCTGGGGCTGGCTGCCAA[a]	1.3µm
2	rs11547464	16	89986091	Exonic	MC1R	G	A	MC1Rset1R		0.5µm[a]	TCAGAGATGGACACCTCCAG		tttttttttttGCCATCGCCGTGGACC	0.1µm
3	rs885479	16	89986154	Exonic	MC1R	C	T	MC1Rset2F	Set 2	0.5µm	CTGGTGAGCTTGGTGGAGA	158bp	ttttttttttttttttGATGGCCGCAACGGCT	1.25µm
4	rs1805008	16	89986144	Exonic	MC1R	C	T	MC1Rset2R		0.5µm	TCCAGCAGGAGGATGACG		ttttttttttttACAGCATCGTGACCCTGCCG	0.375µm
5	rs1805005	16	89985844	Exonic	MC1R	G	T	MC1Rset3F	Set3	0.5µm	GTCCAGCCTCTGCTTCCTG	147bp	ttttttttttttttTGGTGGAGAACGCGCTGGTG	1µm[a]
6	rs1805006	16	89985918	Exonic	MC1R	C	A	MC1Rset3R		0.5µm	AGCGTGCTGAAGACGACAC		ttttttttttttttttttCTGCCTGGCCTTGTCGGA	0.8µm[a]
7	rs1805007	16	89986117	Exonic	MC1R	C	T	MC1Rset4F	Set 4	0.4µm	CAAGAACTTCAACCTCTTTCTCG	106bp	tttttttttttttttttttttttCTCCATCTTCTACGCACTG	1.1µm[a]
8	rs1805009	16	89986546	Exonic	MC1R	G	C	MC1Rset4R		0.4µm	CACCTCCTTGAGCGTCCTG		tttttttttttttttttttttttttttATCTGCAATGCCATCATC	0.12µm[a]
9	Y152OCH	16	89986122	Exonic	MC1R	C	A						tttttttttttttttttttttttttttCATCTTCTACGCACTGCGCTA	0.5µm[a]
10	rs2228479	16	89985940	Exonic	MC1R	G	A						tttttttttttttttttttttttttttttttttCTGGTGAGCGGGAGCAAC	0.375µm
11	rs1110400	16	89986130	Exonic	MC1R	T	C						tttttttttttttttttttttttttttCTTCTACGCACTGCGCTACCACAGCA	0.1µm[a]
12	rs28777	5	33994716	Intronic	SLC45A2	A	C	rs28777_F	Set5	0.4µm	TACTCGTGTGGGAGTTCCAT	150bp	tttttttttttttttttttttttttttttttttttCATGTGATCCTCACAGCAG	1.2µm
								rs28777_R		0.4µm	TCTTTGATGTCCCCTTCGAT			
13	rs16891982	5	33987450	Exonic	SLC45A2	G	C	rs16891982_F	Set 6	0.4µm	TCCAAGTTGTGCTAGACCAGA	128bp	ttttttttttttttttttttttttttttttttttttttttAAACACGGAGTTGATGCA	0.9µm[a]
								rs16891982_R		0.4µm	CGAAAGAGGAGTCGAGGTTG			
14	rs12821256	12	87852466	Intergenic	KITLG	A	G	rs12821256_F	Set 7	0.4µm	ATGCCCAAAGGATAAGGAAT	118bp	tttttttttttttttttttttttttttttttttttGGAGCCAAGGGCATGTTACTACGGCAC	0.12µm[a]
								rs12821256_R		0.4µm	GGAGCCAAGGGCATGTTACT			
15	rs4959270	6	402748	Intergenic	EXOC2	C	A	rs4959270_F	Set 8	0.4µm	TGAGAAATCTACCCCCACGA	140bp	tttttttttttttttttttttttttttttttttttttGGAACACATCCAAACTATGACACTATG	0.3µm[a]
								rs4959270_R		0.4µm	GTGTTCTTACCCCCTGTGGA			
16	rs12203592	6	341321	Intronic	IRF4	C	T	rs12203592_F	Set 9	0.4µm	AGGGCAGCTGATCTCTTCAG	126bp	tttttttttttttttttttttttttttttttttttttttttTCCACTTTGGTGGGTAAAAGAAGG	0.2µm[a]
								rs12203592_R		0.4µm	GCTTCGTCATATGGCTAAACCT			
17	rs1042602	11	88551344	Exonic	TYR	G	T	rs1042602_F	Set 10	0.4µm	CAACACCCATGTTTAACGACA	124bp	tttttttttttttttttttttttttttttttttttttttttttttttTCAATGTCTCTCCAGATTTCA	1.25µm
								rs1042602_R		0.4µm	GCTTCATGGGCAAAATCAAT			
18	rs1800407	15	25903913	Exonic	OCA2	G	A	rs1800407_F	Set 11	0.4µm	AAGGCTGCCTCTGTTCTACG	124bp	tttttttttttttttttttttttttttttttttttttttttttttttttttttttGCATACCGGCTCTCCC	0.1µm
								rs1800407_R		0.4µm	CGATGAGACAGAGCATGATGA			
19	rs2402130	14	91870956	Intronic	SLC24A4	A	G	rs2402130_F	Set 12	0.4µm	ACCTGTCTCACAGTGCTGCT	150bp	ttttttttttttttttttttttttttttttttttttttttttttttttttttttTGAACCATACGGAGCCCGTG	0.75µm
								rs2402130_R		0.4µm	TTCACCTCGATGACGATGAT			
20	rs12913832	15	26039213	Intronic	HERC2	C	T	rs12913832_F	Set 13	0.4µm	TCAACATCAGGGTAAAAATCATGT	150bp	ttttttttttttttttttttttttttttttttttttttttttttttttttttttttttTAGCGTGCAGAACTTGACA	1µm[a]
								rs12913832_R		0.4µm	GGCCCCTGATGATGATAGC			
21	rs2378249	20	32681751	Intronic	ASIP/PIGU	T	C	rs2378249_F	Set 14	0.4µm	CGCATAACCCATCCCTCTAA	136bp	ttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttCCACACCTCTCCTCAGCCCA	0.1µm[a]
								rs2378249_R		0.4µm	CATTGCTTTTCAGCCCACAC			
22	rs12896399	14	91843416	Intergenic	SLC24A4	T	G	Rs12896399_F	Set 15	0.4µm	CTGGCGATCCAATTCTTTGT	125bp	ttttttttttttttttttttttttttttttttttttttttttttttttttttttttttTCTTTAGGTCAGTATATTTTGGG	1µm[a]
								Rs12896399_R		0.4µm	GACCCTGTGTGAGACCCAGT			
23	rs1393350	11	88650694	Intronic	TYR	C	T	Rs1393350_F	Set 16	0.4µm	TTTCTTTATCCCCCTGATGC	124bp	ttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttCATTTGTAAAAGACCACACAGATTT	1.1µm[a]
								Rs1393350_R		0.4µm	GGGAAGGTGAATGATAACACG			
24	rs683	9	12699305	Exonic	TYRP1	T	G	rs683_F	Set 17	0.4µm	CACAAAACCACCTGGTTGAA	138bp	tttttttttttttttttttttttttttttttttttttttttttttttttttttttttttGCTTTGAAAAGTATGCCTAGAACTTTAAT	0.3µm[a]
								rs683_R		0.4µm	TGAAAGGGTCTTCCCAGCTT			

Table 1: Information about the 24 DNA variants of the HIrisPlex assay, including PCR and single base extension (SBE) primer sequences and concentrations. An asterisk indicates where changes were made to the previously published initial HIrisPlex protocol [26].

Sensitivity, peak height balance and consistency

The sensitivity limit of the assay was previously published at 63 pg for a full 24-SNP profile [26]. Here we repeat the sensitivity testing to assess the impact of the minor changes to the assay we introduced. DNA samples from 3 individuals were prepared by serial dilution from 500 pg and re-measured with the Quantifiler Human DNA Quantification kit. The three individuals were genotyped at 500, 250, 125, 63, and 32 pg DNA input with the HIrisPlex assay to assess its sensitivity. Allele peaks were only called if they were above 50 RFU in their respective size range within a custom designed HIrisPlex bin set, using the GeneMapper 3.7 software (Applied Biosystems). The individuals were chosen to display as many heterozygote SNP combinations as possible. Phenotypes include blue eye colour and red hair (Individual 1), brown eye colour and light brown hair (Individual 2) and brown eye colour and red hair (Individual 3). To ensure consistency of genotyping calls and relative fluorescence units (RFUs) at a particular DNA input, DNA samples from several individuals with differing genotype combinations were amplified at the optimum concentration of 250 pg and run several times (minimum 2 and maximum 10 measurements for each SNP in both heterozygous and homozygous forms if available). The average peak height at each allele, and heterozygote peak height ratios for each SNP locus were also ascertained at DNA inputs of 1000, 500 and 125 pg. The heterozygote peak height ratio was calculated by dividing the peak height of the lower molecular weight allele (first peak at the SNP locus) by the peak height of the higher molecular weight allele (second peak at the SNP locus) for all SNPs.

Mixture studies

DNA mixtures from two individuals were made in ratios of 1:1, 1:5 and 1:10, with a total concentration of around 500 pg. An assessment of samples using heterozygote balance and peak height knowledge for single-source at approximately 500 pg DNA input was performed based on this.

Species specificity

A variety of animal DNA samples were genotyped with the HIrisPlex assay to ensure human amplification specificity. These samples included cat, dog, mouse, rat, bovine, pig, and chicken DNA at 2.5 ng input and a chimpanzee sample of 1 ng input. DNA was obtained

commercially from Novagen, Inc. (Madison, WI) for all but the chimpanzee sample, which was described elsewhere [30].

Reproducibilty

Concordance testing was performed on 30 samples of varying DNA concentrations and profiles by five independent laboratories, three of which had no previous experience with the HIrisPlex assay (see Supplementary Table 1 for more details).

Case-type samples and stability studies

Simulated casework samples were genotyped using the HIrisPlex assay under a blind proficiency test, as shown in Supplementary Table 2. Samples include blood, semen, saliva, hair, inhibited samples (intentional haem inhibition) and touched items. DNA from one individual measured at 300 pg DNA was exposed to UV light for time intervals of 0, 0.5, 1, 5, 10, 20 and 30 min, using the Bio-Link (Vilber Lourmat) at a strength of 50 J/cm2 and genotyped to test the stability of the assay under the degraded conditions often found in casework samples.

HIrisPlex eye and hair colour prediction tool and guide

The human eye and hair colour prediction model and guide previously published [26] are core components of the HIrisPlex system. Statistical prediction probabilities are assigned to each of the SNP alleles obtained from the HIrisPlex genotyping assay, for eye colour it was based on 3804 Dutch individuals in the model-building set of a previous study [22–24]. When the number of alleles at a particular SNP is input and combined in the model (for eye colour it is rs12913832 (*HERC2*), rs1800407 (*OCA2*), rs12896399 (*SLC24A4*), rs16891982 (*SLC45A2 (MATP)*), rs1393350 (*TYR*) and rs12203592 (*IRF4*)), it gives three probability estimates for brown, blue and intermediate eye colour of the unknown donor, with the highest probability being the predicted phenotype. For hair colour it is based upon statistical probabilities of 1243 Polish, Irish and Greek individuals in the model-building set of our previous study [26]. When the number of alleles at a particular genotype are input and combined in the model, (for hair colour, it is 11 DNA variants from the MC1R gene, N29insA, rs11547464, rs885479, rs1805008, rs1805005, rs1805006, rs1805007, rs1805009, Y152OCH, rs2228479, rs1110400, and rs28777 (*SLC45A2*), rs12821256 (*KITLG*), rs4959270 (*EXOC2*), rs1042602 (*TYR*), rs2402130 (*SLC24A4*), rs2378249 (*ASIP/PIGU*), rs683

(*TYRP1*), rs12913832 (*HERC2*), rs1800407 (*OCA2*), rs16891982 (*SLC45A2 (MATP)*), and rs12203592 (*IRF4*), it gives four probability estimates for hair colour categories blond, brown, red and black, including hair shade probabilities of light and dark. Using these probability values and following the prediction guide as previously published [26], a final hair colour phenotype can be predicted. Therefore using this HIrisPlex prediction tool, one can predict both eye and hair colour (including shade) from a DNA source. To enhance the accuracy of the models for eye and hair colour prediction, additional individuals were added to both databases, which include full genotypic and phenotypic information. The eye colour database now exceeds 9180 individuals (original model contained 3804 individuals [22, 23, 25]) and the hair colour database now exceeds 1601 (initial number was 1551 [26]). It includes 50 Japanese Individuals for which bioethical approval was obtained (Ethics Committee of Osaka City University Graduate School of Medicine). An updated Excel macro termed "HIrisPlex hair & eye colour prediction tool – enhanced version 1.0" was created to allow simplified calculations of these probability estimates, and can be found in this paper in Supplementary Table 3. As an addition to provide easy access for users, this enhanced HIrisPlex tool can also be found online as a prediction tool on our departmental website www.erasmusmc.nl/fmb/resources. Here you can access the tools prediction probabilities using full and partial profiles produced from the HIrisPlex assay. Also a slightly modified hair colour prediction guide for correct interpretation of the hair colour prediction portion of the tool using this enhanced model can be found in Supplementary Figure. 2.

Population Studies

The HIrisPlex assay was previously genotyped on a subset of HGDP-CEPH samples which includes 945 individuals from 51 worldwide populations as published elsewhere [26].

Results and discussion

Multiplex design and protocol

The multiplex design of the HIrisPlex assay presented and used here was slightly altered from its previously published version [26] with the aim to improve the peak balance of the several SNPs at the lower concentration levels of input DNA. This involved some adjustment of primer concentrations and a redesign of the N29insA (DNA variant 1) SBE primer (see Table 1). Overall, these modifications enhanced the HIrisPlex assay's performance at the

lower DNA input levels, as seen in Figure 1.

Sensitivity, peak height balance and consistency

HIrisPlex genotyping from three individuals with differing genotype combinations was repeated in duplicate with the modified protocol using serial DNA dilutions, which confirmed its high sensitivity as described previously [26]. As evident from Figure 1, we obtained full HIrisPlex profiles in samples tested down to 63 pg of DNA input, in agreement with results from our initial protocol [26]. Partial profiles were observed at 32 pg of DNA in several amplifications. In Fig. 1 we illustrate this using red circles for individual 1, where dropout occurred at N29insA (DNA variant 1) and rs683 (DNA variant 24). For N29insA (DNA variant 1), the C allele was not observed above the recommended 50 RFU threshold level in this heterozygote example at 32 pg of input DNA. Notably, N29insA (DNA variant 1) is an important indicator for red hair colour prediction but is one of the most difficult HIrisPlex DNA variants to amplify when low quality/quantity template DNA is present, as reported previously [28].

Consequently, in all profiles at less than 1 ng DNA input levels, this DNA variants RFU peak height tends to be the lowest in comparison to the other DNA variants in the same HIrisPlex profile (see Figure 2 and Table 2 for more peak height information). It is however worthy to note that it is the A allele that is required for a red hair prediction. In our heterozygote tests, the A allele did not drop out so that a red hair prediction was not missed for this individual. In the second drop out case for this individual at 32 pg DNA input, heterozygote G allele of rs683 (DNA variant 24) was not eminent in the HIrisPlex profile. This DNA marker does not seem to have a special tendency for allele drop out as far as we can say based on our current data. Not observing rs683-G in this sample may therefore reflect the stochastic nature of allelic drop-out at low template DNA with the SNaPshot chemistry used, and given the relatively large number of 24 DNA variants within the same multiplex assay. For individual 2 and 3, drop out was not strictly observed in the duplicate samples that were run at all DNA input levels. However, as seen in Figure. 1 at 32 pg DNA input, there are additional DNA variants (DNA variants 1: N29insA and 4: rs1805008) marked in black circles that appear very close to the 50 RFU threshold. It therefore is possible that drop out (i.e., peak heights less than 50 RFU) could occur at these two DNA variants depending on the heterozygosity of the sample.

A general trend observed with the HIrisPlex assay with decreasing DNA input is that the *MC1R* DNA markers (DNA variants 1–11) display lower peak heights and therefore more potential for drop-out relative to non-*MC1R* DNA markers (DNA variants 12–24). Therefore, depending on the type of sample (i.e. ancient DNA (aDNA) samples) and the sample volume and DNA amount available, it may be advisable to split and run the HIrisPlex assay in two parts on extremely difficult samples. This would involve splitting the HIrisPlex assay into two multiplexes, the *MC1R* red hair colour DNA marker set (DNA variants 1–11) and the eye and remaining hair colour DNA marker set (DNA variants 12–24). However, this modification is only suggested for problematic samples e.g. when running of the full HIrisPlex assay leads to incomplete results particularly for the *MC1R* DNA markers. In general, it is highly recommended to use newly ordered primers and reagents when running the HIrisPlex assay to achieve optimal performance and results; we noted that the freeze–thaw action may interfere with the quality of the result. Our experience with SNaPshot assays other than HIrisPlex tells us (data not shown) that this appears to be a more general SNaPshot related effect rather than a particular effect only noticed with the HIrisPlex assay.

Peak height ratio	1000pg	500pg	125pg
SNP1	0.72±0.18	0.77±0.15	1.43±0.52
SNP2	1.32±0.12	2.03±0.81	1.91±0.33
SNP3	1.07±0.09	1.17±0.05	1.54±0.09
SNP4	0.53±0.04	0.43±0.03	0.51±0.11
SNP5	2.43±0.16	2.66±0.45	2.25±0.64
SNP6	1.13±0.14	1.31±0.02	0.97±0.35
SNP7	0.27±0.01	0.26±0.04	0.25±0.06
SNP8	4.6±1.56	4.52±0.76	2.55±0.09
SNP10	1.39±0.12	1.37±0.13	0.94±0.14
SNP11	0.84±0.16	0.75±0.02	0.92±0.11
SNP12	0.96±0.01	0.98±0.02	0.88±0.01
SNP13	2.61±0.06	1.77±0.02	3.25±0.14
SNP14	1.22±0.02	1.33±0.01	1.31±0.04
SNP15	0.69±0.02	0.67±0.01	0.71±0.07
SNP16	0.49±0.03	0.51±0.01	0.5±0.06
SNP17	3.54±0.4	3.46±0.19	3.76±0.99
SNP18	1.13±0.17	0.99±0.14	1.24±0.39
SNP19	3.05±0.14	3.17±0.19	3.71±1.3
SNP20	1.09±0.15	0.92±0.03	0.86±0.15
SNP21	2.14±0.89	0.86±0.12	0.78±0.1
SNP22	3.35±0.3	3.06±0.12	2.76±0.24
SNP23	1.16±0.34	0.98±0.12	0.9±0.18
SNP24	2.59±0.15	2.63±0.09	2.46±0.15

Table 2: Heterozygote peak height ratios of the HIrisPlex DNA markers at DNA input levels of 1000 pg, 500 pg and 125 pg showing the increase in variability in the peak height ratios with decreasing template DNA amount at some of the loci. Y152OCH (DNA variant 9) is not present due to lack of a heterozygote sample within the validation set used. 95% SE measurements of the average peak height ratios are included.

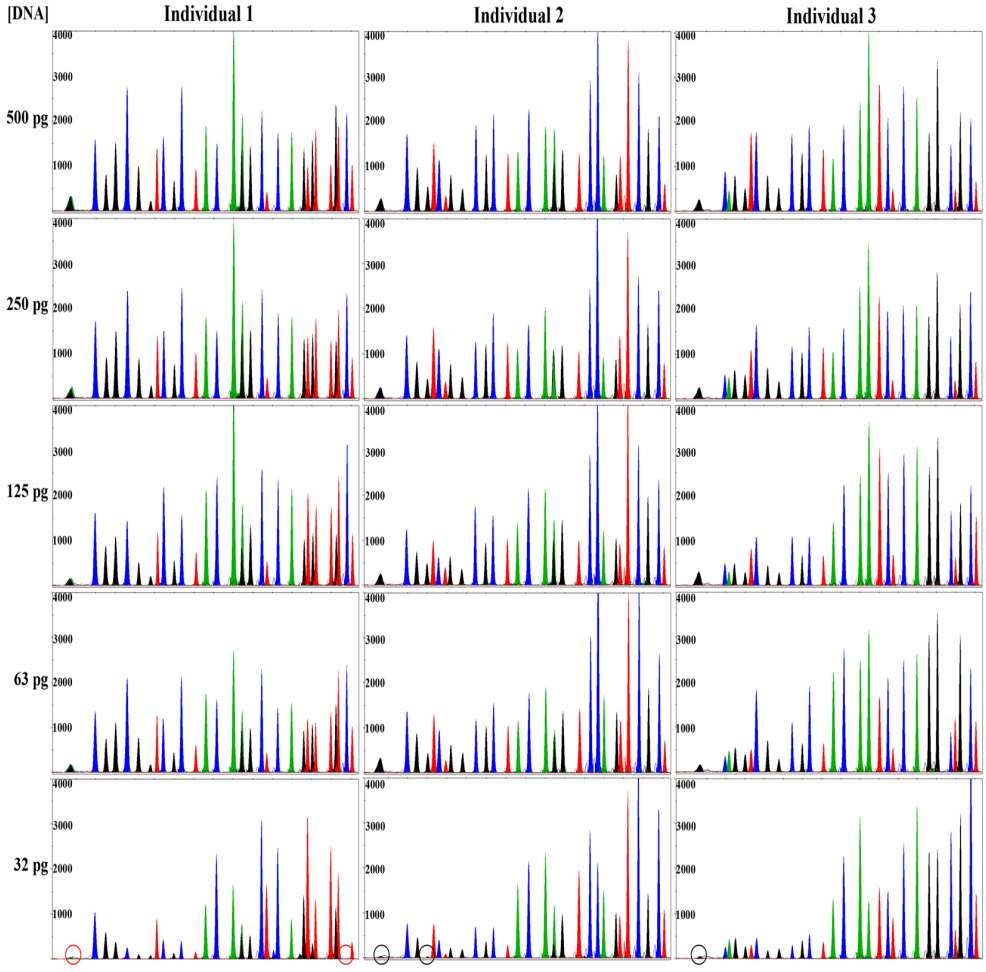


Figure 1: Sensitivity testing of the HIrisPlex assay using three individuals with different eye and hair colour phenotypes and different HIrisPlex genotypes using DNA samples ranging from 500 to 32 pg DNA input, with profiles presented at 500 pg, 250 pg, 125 pg, 63 pg and 32 pg DNA input. Individual 1 displayed phenotypic blue eye colour and red hair colour, individual 2 displayed phenotypic brown eye colour and light brown hair colour, and individual 3 displayed phenotypic brown eye colour and red hair colour. Alleles that have dropped out depending on DNA input amounts are indicated by red circles (<50 RFU) and alleles that appeared near the 50 RFU threshold are indicated in black circles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

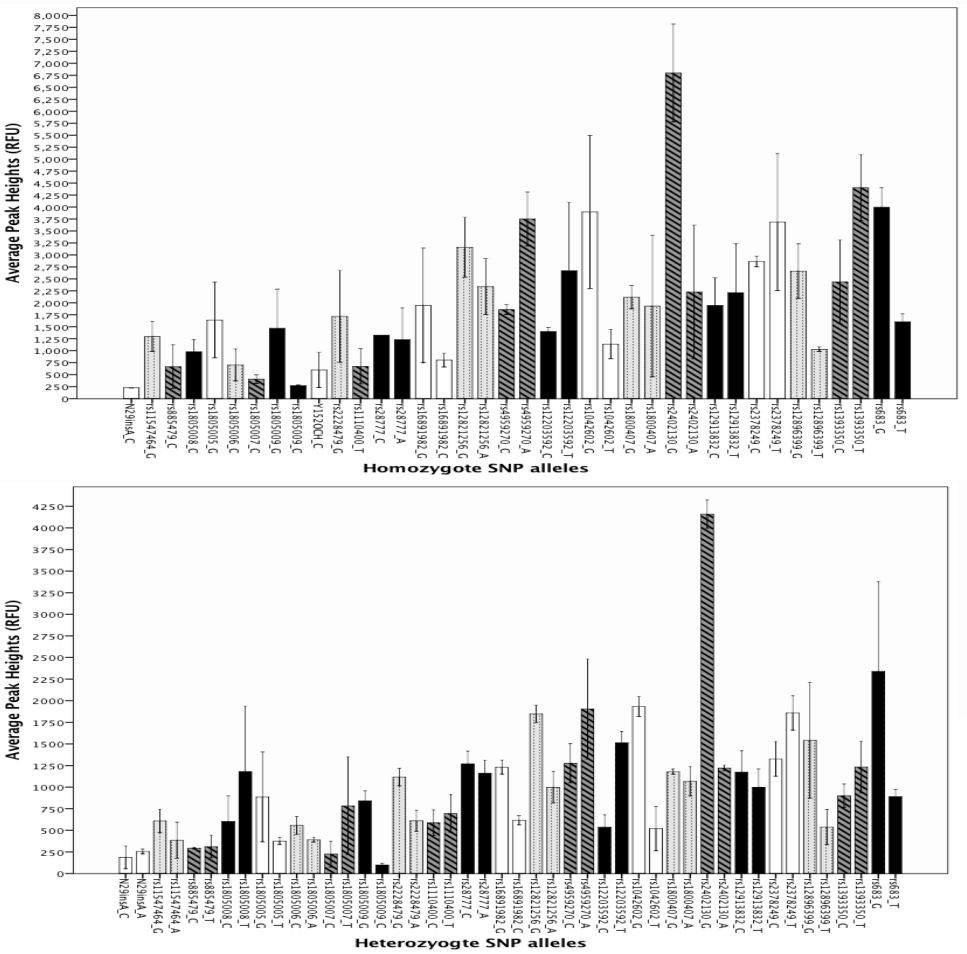


Figure 2: Homozygote and heterozygote average peak heights from several replicates (ranging 2–10 measured RFU values per DNA variant if available) from 8 individuals with varying degrees of heterozygosity at 250 pg of DNA using the HIrisPlex assay. Samples were run with an ABI 3130xl Genetic Analyser, using POP-7 on a 36 cm capillary length array. The A/A homozygote allele of N29insA (DNA variant 1), rs11547464 (DNA variant 2), rs1805006 (DNA variant 6), Y152OCH (DNA variant 9), and rs2228479 (DNA variant 10), the T/T homozygote allele of rs885479 (DNA variant 3), rs1805008 (DNA variant 4), rs1805005 (DNA variant 5), and rs1805007 (DNA variant 7), and the C/C homozygote allele of rs1110400 (DNA variant 11) all from the *MC1R* gene, are absent from 2(A) and the heterozygote SNP profile of Y152OCH (DNA variant 9) is not represented in part 2(B). Error bars represent 95% standard error of the averages.

Overall, the HIrisplex sensitivity achieved with the modified protocol is not only comparable to the initial protocol published [26], but also to several autosomal SNP assays currently being used in forensic laboratories, such as the 18-plex designed by Freire-Aradas et al. [31] requiring a minimum of 78 pg for a full profile. Furthermore, the minimum 63 pg DNA input to generate a full HIrisPlex profile is comparable with the reported sensitivity

limits of some commercial STR kits such as the Powerplex ESX systems from Promega [32] or lower than the Minifiler kit from Applied Biosystems [33].

One important aspect of sensitivity testing with SNP assays is assessing the balance and difference in peak heights (RFUs) of homozygote versus heterozygote alleles. As peak heights tend to be lower in heterozygote cases there is an increased risk of allelic drop out of heterozygote alleles at lower level DNA inputs producing false homozygote effects. Therefore, it is relevant to understand the average peak heights of both genotype combinations as illustrated for HIrisPlex in Figure 2. All available homozygous and heterozygous SNP peak heights from the 24-SNP assay can be seen in Figure 2 measured in duplicate at an optimum DNA input range of 250 pg from 8 different individuals with diverse HIrisPlex profiles. It is important to note that due to the nature of the *MC1R* DNA markers (DNA variants 1–11), their compound heterozygosity [34] and their influence on red hair prediction, it is rarely the case to find all red hair alleles in their homozygous and sometimes heterozygous state. Therefore we did not contain both ancestral and derived alleles for all DNA variants in our test set. In Figure 2(A), balance is achieved as much as possible at the DNA input level of 250 pg. However, there are cases where a large peak height range is observed for a particular DNA marker. This does not reduce the genotyping success of the assay; it simply means that large peak height variation is possible between DNA variants in different individuals depending on the overall profile and the degree of heterozygosity within the sample. In our current validation sample set, average peak heights could not be recorded for the A/A homozygote allele of N29insA (DNA variant 1), rs11547464 (DNA variant 2), rs1805006 (DNA variant 6), Y152OCH (DNA variant 9), and rs2228479 (DNA variant 10), the T/T homozygote allele of rs885479 (DNA variant 3), rs1805008 (DNA variant 4), rs1805005 (DNA variant 5), and rs1805007 (DNA variant 7), and the C/C homozygote allele of rs1110400 (DNA variant 11) all from the *MC1R* gene. Figure 2(B) displays the range of heterozygote DNA variants and their peak heights at 250 pg input DNA. Many DNA markers show higher peaks of one allele in the heterozygote stage, and this is noted by the difference in peak heights for the different alleles within the same DNA marker. Rather than preferential amplification, this observation reflects a common feature of SNP genotyping using fluorescently labelled dyes and can be explained by the differences in intensity levels between the four dyes used to label the four bases in the primer extension reaction (SNaPshot). Consequently, fluorescence and thus peak height also depends on the genotype combination at that particular DNA variant and locus.

From Figure 2(B) we can also see a diverse range of peak heights within a DNA variant i.e. the G allele of rs683 (DNA variant 24) displays a large RFU range in the profiles observed at a 250 pg DNA input. As previously mentioned, due to the large numbers of DNA variants within the multiplex, there is competition for reagents and template. These RFU values are based on 8 unique individuals with very different profiles and therefore such a range is to be expected as it is dependent on all 24 genotypes, their allelic combinations and the level of heterozygosity observed per individual. The heterozygote SNP profile of the *MC1R* red hair DNA marker Y152OCH (DNA variant 9) is not represented in this figure due to its rarity and thus sufficient data could not be collected from our validation sample set. Supplementary Figure 3(I) from our previous paper [26] illustrates its worldwide allele distribution in the CEPH-HGDP set where there were no heterozygote C/A or homozygote A/A genotypes found among the 945 worldwide individuals tested. Alike several other *MC1R* variants in this assay, Y152OCH variation is known to be rare and limited to Europe.

Further examination of peak height ratios between heterozygous alleles at different DNA input levels are provided in Table 2. Equal peak heights at homologous alleles give a balance ratio of 1, while 0.5 represents the smaller molecular weight (first) allele having a peak height half that of the larger molecular weight (second) allele, and the opposite for a ratio of 2. The ratio of peak heights observed between the alleles at all three DNA input levels tested is comparatively consistent. There is slightly more variability seen in both peak heights and peak balance at some HIrisPlex DNA markers with lower template amounts such as N29insA (DNA variant 1) and rs1805009 (DNA variant 8) but this can be expected due to inherent stochastic variation (as explained ab ove). Overall, it is possible to assess the quality of the amplification and completeness of a HIrisPlex profile using information provided with this Table 2 and peak height ratios as a guide, which has been suggested for other assays [35]. However, one should be cautious to solely base conclusions on this information when using very low levels of template DNA and should also consider additional information on the sample itself such as quantity, quality and previous amplification success.

Mixture studies

It has already been stated elsewhere that the presence of a mixture is more easily determined by STRs than by SNP or other biallelic marker analysis and mixture recognition can be difficult to impossible with biallelic markers such as SNPs [36]. As standard crime

scene sampling typically ensures STR typing prior to any downstream DNA analysis for investigative purposes such as the use of the HIrisPlex, STR typing is the preferred method of mixture determination. However, for added informative purposes we illustrate here some interesting factors that may suggest the presence of a 2-person mixture using the HIrisPlex assay. While it appears impossible to detect a mixture at a 1:1 mixture ratio with the HIrisPlex assay due to the peak balance of the SNPs in the assay, there are general indications of a sample mixture seen in 1:5 and 1:10 ratio examples (Figure. 3). For ratios 1:5 and 1:10; N29insA (DNA variant 1), rs1805007 (DNA variant 7), rs4959270 (DNA variant 15), rs1042602 (DNA variant 17), rs12913832 (DNA variant 20), rs2378249 (DNA variant 21), and rs1393350 (DNA variant 23) display substantially lower peak heights for one of the alleles in their genotype combinations that differ from knowledge of known standard single source heterozygotes at the 500 pg DNA input level. Therefore, it is possible to recognise that there may be complications with the profile and that a mixture may be present. Overall, although from this it is not conceivable to be absolutely sure that a mixture is present or even differentiate SNP allele profiles from different contributors with this assay, the well characterised average peak heights and peak height ratios presented in this study allow further investigation of the single or multi-source status of a sample. However, this statement comes with re-emphasising that standard STR typing is the preferred method of mixture determination rather than the HIrisPlex (or any other bi-allelic marker) assay alone.

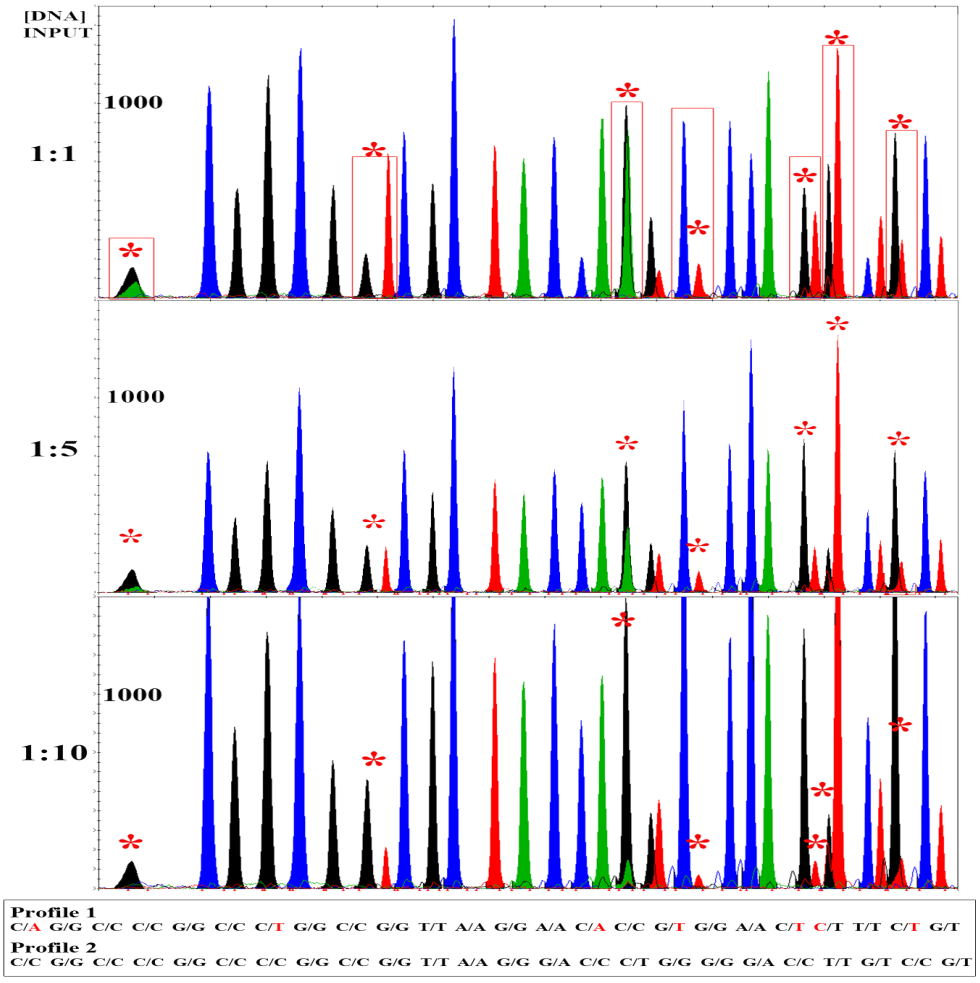


Figure 3: HIrisPlex mixture analysis of 1:1, 1:5 and 1:10 from DNA samples of two individuals with different eye and hair colour phenotypes and different HIrisPlex genotypes. Individual 1 had phenotypic brown eyes and red hair, individual 2 had phenotypic blue eyes and brown hair. The total amount of DNA used was approximately 500 pg. The asterisks represent the alleles present in the 1:1 mixture that may display unusual peak balance in the 1:5 and 1:10 mixtures indicating the presence of a mixture.

Species Specificity

Species specificity was assessed with the HIrisPlex assay by testing samples from several animal species. Supplementary Figure 1 displays the peaks observed and called above the 50 RFU threshold in each of the species at 2.5 ng except for the chimpanzee at 1 ng. From all species tested, only the chimpanzee gave an almost complete HIrisPlex profile (20/24). For example, the SNP rs12913832, a strong predictor of both eye and hair colour is strongly conserved across non-human primate, monkey and lemur species [37]. Other species usually used in validation studies, such as cat, dog, mouse, rat, bovine, pig, and chicken, did not show amplification of alleles in more than two DNA variants. As such, the species specificity of the HIrisPlex assay is sufficient for routine forensic casework situations. However, more investigation into the degree of phylogenetic conservation for all 24 HIrisPlex markers across non-human primates and monkeys may be of interest, in particular the *MC1R* DNA variants and shall be tested in further studies.

Reproducibility

Thirty DNA samples of varying concentrations and from various types of forensically relevant scenarios were assessed for reproducibility and concordance testing of the HIrisPlex assay by five independent laboratories. The four external laboratories involved were the Forensic Laboratory for DNA Research (FLDO) at Leiden University Medical Centre, the Netherlands Forensic Institute (NFI), both involved in routine forensic DNA analysis, the Laboratory of Forensic Medicine & Toxicology (FMT) in Greece and the Institute of Forensic Research (IFR) in Poland. One external test site (IFR) had previous experience with the HIrisPlex assay from successfully applying it to DNA samples obtained from modern and several hundreds of years old human remains [38], while two other external institutes (FLDO and NFI) had general knowledge of SNP genotyping and SNaPshot technology, albeit not on HIrisPlex. The fourth external institution (FMT) had no previous SNP typing and SNaPshot experience. All participating groups were provided with DNA samples, a written protocol as previously published [26] using the altered primer concentrations as described in Table 1, the Excel macro for statistical prediction of eye colour and hair colour including hair colour shade (Supplementary Table 3), and the hair colour prediction guide [26]. Therefore, a full test of the HIrisPlex assays reliability and ease of use was undertaken when performed by laboratories and analysts with and without experience of SNaPshot in general and/or

HIrisPlex in particular. Supplementary Table 1 provides full descriptions of the samples tested, and the level of concordance achieved by the five laboratories. As evident, the vast majority of the samples (91%) were genotyped correctly in all external laboratories. One laboratory had difficulty with the lower level DNA input samples, and stated that is was possibly due to evaporation of these samples during PCR. In general, even with the different experience levels in using the SNaPshot® technology and the HIrisPlex SNP assay in particular, all four laboratories commented on the consistency and ease of use of the HIrisPlex assay.

Case-type samples and stability studies

Twenty-five samples were produced to simulate a forensic casework type scenario to test the HIrisPlex assay's range and ability to produce a full profile and subsequently to also test for reliable phenotype prediction from varying sample types with differing DNA qualities and quantities. Genotype designation and profile assessment, including eye and hair colour phenotype prediction from the assay results were also blind-tested to check the proficiency level of the validation. Supplementary Table 2 gives a full list of the types of samples tested, their concentration and the analyst's phenotype assessment and prediction results. The vast majority of samples (84%) from blood, semen, saliva, hair and touched objects, were correctly predicted (21 out of 25). On the correctly genotyped single-source samples (15 out of 18), the phenotype prediction matched those of the volunteers reported eye and hair colour, apart from one instance. In this individual HIrisPlex revealed probability values of 0.5 p blue, 0.3 p intermediate and 0.2 p brown indicating blue as the most likely eye colour, while the reported phenotype was categorised as an intermediate eye colour. Of the three incorrect single-source samples, two concern very low input DNA or intentional contamination samples, where there was no profile produced (Supplementary Table 2). The final incorrect sample was a low input DNA single source sample that was interpreted incorrectly as a possible multi-source sample due to imbalanced allele indicators (STR profiling would rule out these difficult source predictions). As seen, overall, the HIrisPlex system is able to accept a range of sample types, and showed high levels of accuracy in genotyping and phenotyping prediction across a number of casework type samples.

Aiming to test how the HIrisPlex assay copes with degraded DNA, single source DNA at an input of 300 pg was treated with UV rays at varying time intervals to simulate DNA degradation, and samples were genotyped with the HIrisPlex assay. Notably, the assay

was designed to facilitate the amplification of degraded DNA with target templates less than 160 bp in size. Complete HIrisPlex profiles were obtained from DNA samples treated with UV light for 0.5, 1, 2 and 5 min, while partial profiles at 10 min (21 of 24 markers) and 20 min (17 of 24 markers), and a complete lack of amplification at all 24 HIrisPlex DNA markers was seen after 30 min. Also important to note is that at the 20 min UV time, there were 2 allele drop-outs (variant 5 and 17) and 1 allele drop-in (variant 13). When dealing with low template degraded samples, it is advantageous to amplify a sample several times to ensure consistency of genotyping. Furthermore, the HIrisPlex system was used recently on DNA samples from modern up to hundreds of years old human remains with proven success and consistency [38]; when dealing with such difficult samples, it is useful to follow the guidelines as outlined in this previous publication.

Enhanced hair colour prediction model

The HIrisPlex system is composed of an eye and hair colour prediction tool. As previously mentioned, the eye colour prediction part of the tool was based on a multinomial logistic regression model trained on genotype and eye colour phenotype data from 3804 Dutch European individuals [22] and further tested with data on over 3840 individuals from seven sites throughout Europe, which yielded >94% prediction accuracy for blue and brown eye colour prediction [24]. The hair colour prediction part of the HIrisPlex tool was based on a multinomial logistic regression model trained genotype and hair colour phenotype data on 1243 European individuals from Poland, Ireland and Greece and validated on data from 308 individuals from these countries, which yielded on average prediction accuracies of 69.5% for blond, 78.5% for brown, 80% for red and 87.5% for black using the prediction guide approach as previously published [26]. To further enhance eye and hair colour prediction, we now include additional individuals with genotype/phenotype data into both the eye and hair colour prediction models (Supplementary Table 3). For eye colour prediction, the model underlying database now includes samples united from both publications by Liu et al. (2009) [22] and Walsh et al. (2012) [24], with a total number of 9188 individuals from eight parts of Europe (Netherlands, Norway, Estonia, UK, France, Italy, Spain, and Greece). For hair colour prediction, the model-underlying database now includes 1601 individuals, 1551 from the publication Walsh et al. (2013) [26] from Ireland, Greece and Poland and 50 individuals from a new Japanese set for which a black hair phenotype is recorded.

Notably, the inclusion of these Japanese samples into the model underlying dataset represents a non-European black hair subset. Although being monomorphic in hair colour by nature, adding these Japanese individuals to the model increases the phenotypic variation range through the introduction of an extreme black hair phenotype, with the majority of causal alleles displaying the ancestral genotype. Both phenotype and genotype are not typically observed in Europeans, and thus they help enhance black hair, blond hair and hair shade prediction accuracies.

With such model-underlying reference database enlargements, we now obtain overall prediction accuracies, as measured by the Area Under the receiver operating characteristic Curve (AUC) values, for eye colour at 0.94 for blue, 0.74 for intermediate, and 0.95 for brown, and for hair colour at 0.81 for blond, 0.75 for brown, 0.92 for red, and 0.85 for black hair with a hair colour shade at 0.90. Overall, these AUC values represent only a slight improvement on prediction accuracies compared with what we obtained from our previous eye and hair colour prediction models using smaller underlying genotype and phenotype database. However, more importantly, these improvements are expected to provide additional strength with more difficult samples such as rare genotype–phenotype combinations as with increased database size come increased genotype combinations. Using an independent test set of 119 additional Polish individuals who were not included in model building, an overall look at accuracy was undertaken using the enhanced eye and hair models (Supplementary Table 4). Eye colour could be predicted with 84% accuracy on average (100 out of 119 predictions were correct). Worthy to note is that this value increases to 93% accuracy (100 out of 108) when only brown and blue prediction is assessed and the intermediate eye colour category was not included. As emphasised before [22,24], intermediate eye colours are more difficult to predict than blue and brown either due to lack of information on the SNPs/genes responsible for these colours (in case they indeed have different underlying genes), or because of the absence of expressed eumelanin at different quantities within irises that is not easily captured by the IrisPlex SNPs alone. Using the same test set, hair colour was predicted with 73% accuracy on average. We have reported in our previous publication [26], that there are individuals whose hair colour is more difficult to predict with HIrisPlex (and expectedly with any other hair colour test with currently known DNA markers). This is due to age-related hair colour changes overtime from a child to an adult, where the current adult hair colour is predicted incorrectly. However, this chance of error (DNA predicted blond but phenotype is brown,

in rare circumstances brown/black) has been included in our accuracy analysis and notably many individuals who were blond as a child still remain in the blond category as an adult, and in most cases a slightly darker blond. Although these age-dependent hair colour changes have been linked to hormonal changes, to date there have been no studies that specifically explore this phenomenon and hence a predictive test has yet to be developed that considers this. Overall, the enhanced HIrisPlex tool performs well with 3 out of 4 cases of eye and hair colour prediction being accurate within this 119 Polish test-set. This enhanced HIrisPlex prediction tool version 1 is available via an excel spreadsheet published as Supplementary Table 3 and employs the enhanced prediction guide provided in Supplementary Figure 2.

Online eye and hair colour DNA prediction tool for complete and partial HIrisPlex profiles

Additionally, we developed an online web-based tool for HIrisPlex eye and hair colour prediction based on the enhanced prediction tool introduced here, and with this announce its public availability for practical use in forensic case work and other applications such as anthropological studies. It can be found at www.erasmusmc.nl/fmb/resources. With this online tool, individual eye and hair colour probabilities are obtainable not only from complete IrisPlex/HIrisPlex profiles as was possible before with our published spreadsheets [23–26] but can now be generated from partial IrisPlex/HIrisPlex profiles. As outlined above and as expected with low quality/quantity DNA samples, partial or incomplete IrisPlex/HIrisPlex profiles may be observed in some forensic casework applications and any other applications where DNA quality and/or quantity is highly limited such as anthropological studies including ancient DNA studies. With the use of this online prediction tool for partial IrisPlex/HIrisPlex profiles, the underlying prediction models (one for eye colour and another for hair colour, including shade) are re-estimated every time a partial profile is loaded so that the models are adjusted towards the particular DNA markers available from the input partial profile. This approach provides a more realistic probability estimate from partial profiles because only the DNA variants present in the profile are used for model building compared to the use of a model that is based on a full profile for probability estimation. However, it is important to note that, as expected, partial profile predictions will not perform to the full levels of accuracy relative to full profile predictions. Furthermore, it shall be considered that different DNA variants used in the eye and hair colour prediction models have different impacts on the eye and hair colour prediction accuracies as deterined by their differences in

predictive strength each DNA marker has on eye and/or hair colour. It therefore does not simply depend on having a partial or a complete profile, but the accuracy obtainable with a partial profile is strongly determined by which DNA markers are present in the partial profiles (or missing from a complete profile). For instance, the strongest predictor for eye colour is *HERC2s* rs12913832, which also impacts considerably on several hair colours; consequently, missing this SNP in a partial profile severely limits the prediction accuracy of all eye colour categories and of some hair colour categories such as blond and black. To illustrate this effect, AUC for blue eyes drops from 0.94 AUC with all 6 SNPs used for eye colour prediction in IrisPlex/HIrisPlex to 0.64 AUC when missing rs12913832 in a partial IrisPlex/HIrisplex profile, and for brown eyes from 0.95 AUC to 0.65 AUC as estimated when using our revised eye colour prediction model introduced here. To note, as an AUC of 0.5 means random prediction, no predictive DNA marker is needed to achieve 0.5 AUC. For red hair, the by far strongest predictors are DNA variants from the *MC1R* gene (i.e. DNA variants 1–11 in HIrisPlex assay); consequently, the red hair phenotype is likely to be missed if several of these DNA markers are missing in a partial HIrisPlex profile. To illustrate: AUC for red hair drops from 0.92 AUC with all 11 *MC1R* markers in the HIrisPlex profile to 0.64 AUC when all 11 HirisPlex DNA variants are missing from a partial profile, when using our revised hair colour prediction model introduced here. For non-red hair colours, three SNPs *HERC2*s rs12913832, *SLC45A2*s rs16891982, and *IRF4*s rs12203592 have the strongest predictive effect. Missing this combination from a HIrisPlex profile severely affects hair colour and shade prediction accuracy with an AUC loss for black from 0.85 AUC with complete 24 marker HIrisPlex profile to 0.78 AUC, for brown from 0.75 to 0.66, and for blond from 0.81 to 0.70 with the 21 marker profile missing these three DNA variants, as estimated with our revised hair colour prediction model. As a partial profile has many possibilities and the missing genotypes would reflect different accuracy losses in particular colour categories, we allow the model to still produce probabilities based on the genotypes input, but at the same time strongly advise that the AUC losses due to missing DNA markers are reported together with the prediction outcomes. This allows a comparison for the analyst of the limitations a particular partial profile relative to the full profile. The analyst can then make the report stating the result as acceptable or inconclusive. The ‘loss in AUC’ information is provided as an additional output of the online tool as estimated from the reference database using a complete profile versus the partial profile. We made several exceptions due to the severity

of the eye or hair colour effects as described above: *HERC*2 rs12913832, the *MC1R* DNA markers and a combination of *HERC2-SLC45A2-IRF4.* If *HERC2* rs12913832 is missing from an input partial IrisPlex/HIrisPlex profile, no eye colour probabilities are provided in the output. If all 11 *MC1R* DNA markers are missing, no hair colour probability is provided. If a combination of rs12913832-*HERC2*, rs16891982-*SLC45A2*, and rs12203592-*IRF4* are missing no eye or hair colour probability is provided.

We would like to re-emphasise that the typical forensic application of FDP tools such as HIrisPlex concern cases where a full STR profile is obtained, but it did not match a known person's STR profile. Given the sensitivity of the HIrisPlex assay as described here and before [26], which is similar if not more sensitive that commercial STR kits used for human identification, we expect that in a typical forensic case for which a complete STR profile is obtained, a complete HIrisPlex profile is also obtainable assuring maximal power for eye and hair colour prediction. However, there will be cases where DNA quality and/or quantity is insufficient for generating complete STR profile and in such cases there is a risk to also generate a partial HIrisPlex profiles. The newly provided online HIrisPlex prediction tool can now be used in such cases, but care has to be taken in the interpretation of the obtained eye and hair colour probabilities depending on the particular DNA marker(s) missing in a partial IrisPlex profile. Lastly, for both full and partial profiles, DNA phenotyping is probability-based prediction; hence the result is based on knowledge from a combination of genotypes from thousands of individuals in a model-training set. In some circumstances i.e. green eye colour, or age-related changes in blond/black hair colour, there are additional factors that are not included in current prediction models such as the underlying genetic or other biological factors that have not been found yet. Hence HIrisPlex prediction of eye and hair colour should be taken with the understanding that it is not and cannot be, given current scientific knowledge, 100% correct. Instead, a final prediction result from the HIrisPlex system provides the most likely eye and hair colour category outcome given the markers used in the assay and the reference datasets used in the prediction modelling, but it does not rule out other colour categories completely. However, as of today, HIrisPlex is the most accurate DNA tool available for human phenotype prediction of eye and hair colour in parallel.

Finally, and to re-emphasise, categorical eye and hair colour prediction as achieved with the HIrisPlex system or any other DNA test previously developed for eye or hair colour prediction [13,39] can potentially come with limitations in its final use in terms of

the prediction outcomes in police investigation. Individuals may grade eye and hair colour differently into the typically used categories. Although it is unlikely that these observer differences lead to blue eyes being graded as brown, or blond hair graded as black, this may happen with the more similar categories e.g. blue eyes may be graded as intermediate, or dark blond hair may be graded as brown. Obviously, unless the categorical prediction approach can be overcome and eye and hair colours can be predicted on a continuous scale, this limitation cannot be solved. However, research on the genetic basis and predictability of continuous eye colour has already started [40].

Population Studies

The HIrisPlex system was previously tested on the HGDP-CEPH H952 worldwide samples as a population study [26]. In summary, the HIrisPlex assay gave the genotypes of 945 individuals from 51 worldwide populations. This population study (Figure 5 [26]) showed the restriction of the prediction of different hair colour categories to Europeans and individuals from immediate surrounding areas (Middle East and West Asia), whereas individuals from the rest of the world were all predicted with black hair colour. In this previous study we applied the highest-probability-indicates-predictedcategory approach and used the initial hair colour model, while Supplementary Figure 3 of the present study provides a worldwide look on hair colour prediction using the enhanced hair colour model introduced here and the newly developed hair colour prediction guide (Supplementary Figure 2) on the same HGDP-CEPH H952 HIrisPlex 797 genotype data. One minor difference between the two figures is that there is an over-representation of blond hair prediction using the highest category approach over the prediction guide approach, which represents a more accurate view within Europe. Similarly, we predicted eye colour using the IrisPlex system in these worldwide samples and showed the restriction of different predicted eye colour categories to Europeans and individuals from immediate surrounding areas whereas ndividuals from the rest of the world were only predicted brown [23]. Athough no eye and hair colour phenotypes are available for the HGDP-CEPH samples, in general the predicted worldwide hair and eye colour distribution fit with the expected phenotype distribution given existing general knowledge and available data around the world [41]. This displays the accuracy and reliability of the HIrisPlex system for hair colour prediction as well as eye colour prediction of these pigmentation phenotypes worldwide and illustrates that the accuracy

of categorical hair and eye colour prediction with the HIrisPlex system is independent of biogeographic origin or genetic ancestry of the individual analysed. Also worthy to note, and an added bonus to the prediction probability outcomes generated by the HIrisPlex system, is the ability of the HIrisPlex system to differentiate between a brown eyed, black haired individual of European bio-geographic origin (and from areas immediately surrounding Europe) from a brown eyed, black haired non-European (outside surrounding areas). Overall accuracies of up to 86.5% are predicted when using a threshold of ≥0.7 p for black hair and ≥ 0.99 p for brown eyes as wedescribed previously ([26], Supplementary Figure 2)

Conclusions

The developmental validation of the HIrisPlex assay matches all SWGDAM guidelines in terms of sensitivity, stability, reproducibility, precision and accuracy, species specificity, casework samples and population studies. The observed problems with mixture detection are not typical to the HIrisPlex assay but concern all bi-allelic DNA marker systems. HIrisPlex surpasses many genotyping assays currently used in a forensic environment in terms of sensitivity, with only 63 pg minimum DNA input required to achieve complete profiles. Accurate and reliable genotyping results were produced from simulated casework samples that included blood, semen, saliva, hair, and low quantity trace DNA samples. The generation of consistent HIrisPlex results both internally and externally in five laboratories with varying background experiences demonstrates its ease of use and its reliability. The enlargements of the genotype and phenotype reference databases underlying the now enhanced eye and hair colour prediction models contribute towards increased prediction accuracies especially for more rare genotype-phenotype combinations and for black hair phenotypes in people of European descent. The publically available web-based online prediction tool, which adjusts the eye and hair colour prediction models according to the DNA markers available in an input HIrisPlex profile, now allows the estimation of individual eye and hair colour probabilities from complete and partial profiles with the exception of missing *HERC2*s rs12913832 and all 11 *MC1R* DNA markers. The HIrisPlex assay together with both eye and hair colour prediction models and the easy to use online tool is ready for immediate implementation and use in any accredited forensic laboratory for aiding police investigations in the prediction of eye and hair colour pigmentation traits of unknown suspects, perpetrators and missing persons from DNA. Furthemore, the presented validation results demonstrate the robustness

and reliability of the HIrisplex system and may also encourage scientists outside the forensic arena such as those in anthropological or evolutionary studies including aDNA researchers to use this tool in the DNA prediction of eye and hair colour.

.**Acknowledgements**

We are grateful to all volunteers who provided DNA samples for this study. We are additionally grateful to S. van Baal and A.Verkerk for help in establishing the HIrisPlex online prediction tool. This work was funded in part by the Netherlands Forensic Institute (NFI), Erasmus MC University Medical Center Rotterdam, and by a grant from the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands (FGCN). WB has received funding support from the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement no 285487 (EUROFORGEN-NoE).

References

[1] Sulem P, Gudbjartsson D, Stacey S, et al., Genetic determinants of hair, eye and skin pigmentation in Europeans, Nat. Genet. 39 (2007) 1443–1452.

[2] Duffy DL, Montgomery GW, Chen W, Zhao ZZ, Le L, James MR, Hayward NK, Martin NG, Sturm RA, A three single-nucleotide polymorphism haplotype in intron 1 of *OCA2* explains most human eye-color variation, Am. J. Hum. Genet. 80 (2007) 241–252.

[3] Han J, Kraft P, Nan H, et al., A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation, PLoS Genet. 4 (2008) e1000074.

[4]Branicki W, Brudnik U, Kupiec T, Wolanska-Nowak P, Szczerbinska A, Wojas-Pelc A, Association of polymorphic sites in the OCA2 gene with eye colour using the tree scanning method, Ann. Hum. Genet. 72 (2008) 184–192.

[5] Eiberg H, Troelsen J, Nielsen M, Mikkelsen A, Mengel-From J, Kjaer K, 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. 123 (2008) 177–187.

[6] Sulem P, Gudbjartsson D, Stacey S, et al., Two newly identified genetic determinants of pigmentation in Europeans, Nat. Genet. 40 (2008) 835–837.

[7] Kayser M, Liu F, Janssens A, et al., Three genome-wide association studies and a linkage analysis identify *HERC2* as a human iris color gene, Am. J. Hum. Genet. 82 (2008) 411–423.

[8] Sturm RA, Larsson M, Genetics of human iris colour and patterns, Pigment Cell Melanoma Res. 22 (2009) 544–562.

[9] Mengel-From J, Wong T, Morling N, Rees J, Jackson I, Genetic determinants of hair and eye colours in the Scottish and Danish populations, BMC Genet. 10 (2009) 88.

[10] Valenzuela RK, Henderson MS, Walsh MH, et al., Predicting phenotype from genotype: normal pigmentation, J. Forensic Sci. 55 (2010) 315–322.

[11] Spichenok O, Budimlija Z, Mitchell 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. 5 (2011) 472–478.

[12]Pospiech E, Draus-Barini J, Kupiec T, Wojas-Pelc A, Branicki W, Gene–gene interactions contribute to eye colour variation in humans, J. Hum. Genet. 56 (2011) 447–455.

[13] Ruiz Y, Phillips C, Gomez-Tato A, et al., Further development of forensic eye color predictive tests, Forensic Sci. Int. Genet. 7 (2013) 28–40.

[14] Valverde P, Healy E, Jackson I, Rees J, Thody A, Variants of the melanocytestimulating hormone receptor gene are associated with red hair and fair skin in humans, Nat. Genet. 11 (1995) 328–330.

[15] Kanetsky PA, Swoyer J, Panossian S, Holmes R, Guerry D, Rebbeck TR, A polymorphism in the agouti signaling protein gene is associated with human pigmentation, Am. J. Hum. Genet. 70 (2002) 770–775.

[16] Graf J, Hodgson R, van Daal A, Single nucleotide polymorphisms in the *MATP* gene are associated with normal human pigmentation variation, Hum. Mutat. 25 (2005) 278–284.

[17] Branicki W, Brudnik U, Draus-Barini J, Kupiec T, A. Wojas-Pelc, Association of the SLC45A2 gene with physiological human hair colour variation, J. Hum. Genet. 53 (2008) 966–971.

[18] Shekar SN, Duffy DL, Frudakis T, Montgomery GW, James MR, Sturm RA, Martin NG, Spectrophotometric methods for quantifying pigmentation in human hair—influence of MC1R genotype and environment, Photochem. Photobiol. 84 (2008) 719–726.

[19] Kayser M, Schneider P, DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations, Forensic Sci. Int. Genet. 3 (2009) 154–161.

[20] Kayser M, de Knijff P, Improving human forensics through advances in genetics, genomics and molecular biology, Nat. Rev. Genet. 12 (2011) 179–192.

[21] Kayser M, Forensic DNA Phenotyping: DNA testing for externally visible characteristics, in: J.A. Siegel, P.J. Saukko (Eds.), Encyclopedia of Forensic Sciences, Academic Press, Waltham, 2013, pp. 369–374.

[22] Liu F, van Duijn K, Vingerling J, Hofman A, Uitterlinden A, Janssens A, Kayser M, Eye color and the prediction of complex phenotypes from genotypes, Curr. Biol. 19 (2009) 192–193.

[23] Walsh S, Liu F, Ballantyne K, 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. 5 (2011) 170–180.

[24] Walsh S, Wollstein A, Liu F, et al., DNA-based eye colour prediction across Europe with the IrisPlex system, Forensic Sci. Int. Genet. 6 (2012) 330–340.

[25] Walsh S, Lindenbergh A, Zuniga S, Sijen T, de Knijff P, Kayser M, Ballantyne KN, Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence, Forensic Sci. Int. Genet. 5 (2011) 464–471.

[26] Walsh S, Liu F, Wollstein A, Kovatsi L, Ralf A, Kosiniak-Kamysz A, Branicki W,. Kayser, The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA, Forensic Sci. Int. Genet. 7 (2013) 98–115.

[27] Grimes EA, Noake PJ, Dixon L, Urquhart A, Sequence polymorphism in the human melanocortin 1 receptor gene as an indicator of the red hair phenotype Forensic Sci. Int. 122 (2001) 124–129.

[28] Branicki W, Liu F, van Duijn K, Draus-Barini J, Pospiech E, Walsh S, Kupiec T, Wojas-Pelc A, Kayser M, Model-based prediction of human hair color using DNA variants, Hum. Genet. 129 (2011) 443–454.

[29] Scientific Working Group on DNA Analysis Methods (SWGDAM), Revised validation guidelines, Forensic Sci. Commun. 6 (2004).

[30] Erler A, Stoneking M, Kayser M, Development of Y-chromosomal microsatellite markers for nonhuman primates, Mol. Ecol. 13 (2004) 2921–2930.

[31] Freire-Aradas A, Fondevila M, Kriegel AK, Phillips C, Gill P, Prieto L, Schneider PM, Carracedo A, Lareu MV, A new SNP assay for identification of highly degraded human DNA, Forensic Sci. Int. Genet. 6 (2012) 341–349.

[32] Tucker VC, Hopwood AJ, Sprecher CJ, McLaren RS, Rabbach DR, Ensenberger MG, Thompson JM, Storts DR, Developmental validation of the PowerPlex® ESX 16 and PowerPlex® ESX 17 Systems, Forensic Sci. Int. Genet. 6 (2012) 124–131.

[33] Mulero JJ, Chang CW, Lagace´ RE , Wang DY, Bas JL, McMahon TP, Hennessy LK, Development and validation of the AmpF‘STR1 MiniFilerTM PCR amplification kit: a MiniSTR multiplex for the analysis of degraded and/or PCR inhibited DNA, J. Forensic Sci. 53 (2008) 838–852.

[34] Liu F, Struchalin MV, van Duijn K, Hofman A, Uitterlinden AG, van Duijn C, Aulchenko YS, Kayser M, Detecting low frequent loss-of-function alleles in genome wide association studies with red hair color as example, PLoS ONE 6 (2011) e28145.

[35] Børsting C, Rockenbauer E, Morling N, Validation of a single nucleotide polymorphism (SNP) typing assay with 49 SNPs for forensic genetic testing in a laboratory accredited according to the ISO 17025 standard, Forensic Sci. Int. Genet. 4 (2009) 34–42.

[36] Sobrino B, Brion M, Carracedo A, SNPs in forensic genetics: a review on SNP typing methodologies, Forensic Sci. Int. 154 (2005) 181–194.

[37] Bradley BJ, Pedersen A, Mundy NI, Brief communication: blue eyes in lemurs and humans: same phenotype, different genetic mechanism, Am. J. Phys. Anthropol. 139 (2009) 269–273.

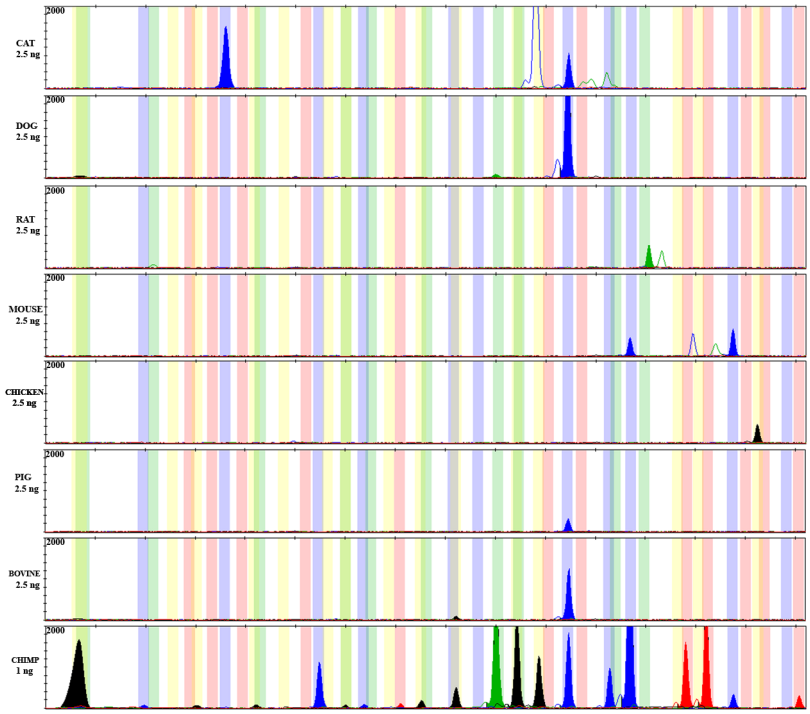
[38] Draus-Barini J, Walsh S, Pospiech E, Kupiec T, Glab H, Branicki W, Kayser M, Bonafide colour: DNA prediction of human eye and hair colour from ancient and contemporary skeletal remains, Invest. Genet. 4 (2013) 3.

[39] Hart KL, Kimura SL, Mushailov V, Budimlija ZM, Prinz M, Wurmbach E, Improved eye- and skin-color prediction based on 8 SNPs, Croat. Med. J. 54 (2013) 248–256.

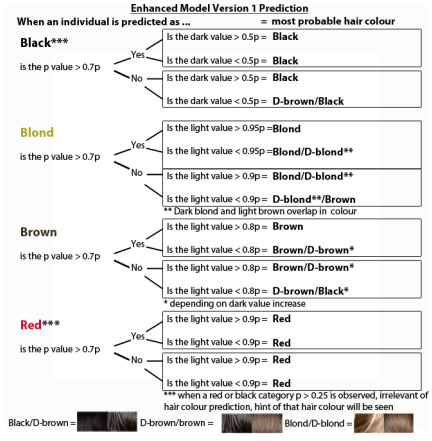
[40] Liu F, Wollstein A, Hysi PG, et al., Digital quantification of human eye color highlights genetic association of three new loci, PLoS Genet. 6 (2010) e1000934.

[41] R.L. Beals, H. Hoijer, An Introduction to Anthropology, MacMillan, New York, 1965.

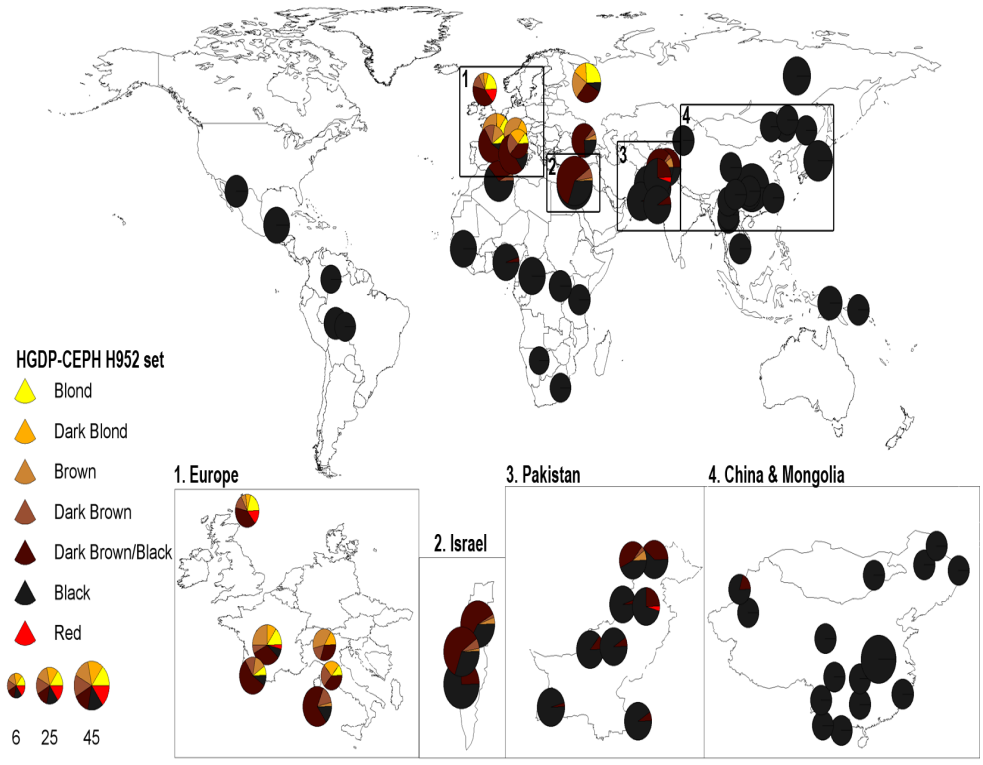
Supplementary Materials



Supplementary Figure 1*:* Human specificity testing of the HIrisPlex assay on several animal species, cat, dog, rat, mouse, chicken, pig, bovine and chimp. The presence or absence of a HIrisPlex DNA variant after genotyping is highlighted (threshold 50 RFU, scale set to 2000 RFU).



Supplementary Figure 2: Enhanced prediction guide for interpretation of hair colour prediction.



Supplementary Figure 3: A worldwide look on hair colour prediction using the enhanced hair colour model introduced here and the newly developed hair colour prediction guide.

Concordance Sample	Type	Concentration	Site 1	Site 2 & comments	Site 3 & comments	Site 4 & comments	Site 5 & comments
1	Saliva	1000 pg	FP	FP	FP	FP	FP
2	Buccal	800 pg	FP	FP	X Partial profile 20/24 (D/O of variant 1,5,6,7)	FP	FP
3	Saliva	1100 pg	FP	FP	FP	FP	X Partial Profile 23/24 (D/O of T allele at variant 5
4	Buccal	450 pg	FP	FP	FP	FP	FP
5	Saliva	1050 pg	FP	FP	FP	FP	FP
6	Buccal	200 pg	FP	FP	FP	FP	FP
7	Buccal	50 pg	FP	FP	X Partial profile 23/24 (D/O of C allele at variant 1)	X Partial profile 19/24 D/O allele variant 1, 15 (A allele),17(G allele),24(G allele), Drop in G allele at variant 22	X Partial Profile (D/O of C allele at variant 1)
8	Saliva	137 pg	FP	FP	X Partial profile 6/24 (D/O in all but variant 13,18,20,21,22,23)	FP	FP
9	Buccal	870 pg	FP	FP	FP	FP	FP
10	Saliva	7000 pg	FP	FP	FP	FP	FP
11	Buccal	75 pg	FP	FP	FP	FP	FP
12	Buccal	40 pg	FP	X Partial Profile 23/24 (D/O of T allele at SNP 22)	FP	FP	X Partial Profile 23/24 (D/O of C allele at variant 8)
13	Buccal	9000 pg	FP	FP	FP	FP	FP
14	Buccal	16000 pg	FP	FP	FP	FP	FP
15	Buccal	148 pg	FP	FP	FP	FP	FP
16	Frozen Blood	500 pg	FP	FP	X Partial Profile 23/24 (Drop In of A allele at variant 1)	FP	FP
17	Buccal	250 pg	FP	FP	FP	FP	FP
18	Frozen Blood	7000 pg	FP	FP	FP	FP	FP
19	Saliva	4000 pg	FP	FP	FP	FP	FP
20	Saliva	350 pg	FP	FP	FP	FP	FP
21	Saliva	1900 pg	FP	FP	FP	FP	FP
22	Control DNA	70 pg	FP	FP	FP	FP	FP
23	Mucus	175 pg	FP	FP	FP	FP	FP
24	Dried Saliva	80 pg	FP	FP	FP	FP	FP
25	Trace DNA	79 pg	FP	X Partial Profile 21/24 (Drop In of T allele at variant 3 & 4 and 24)	FP	FP	Loss of entire sample due to error
26	Control DNA	94 pg	FP	FP	FP	FP	FP
27	Control blank	-	NP	-	-	-	-
28	Control blank	-	NP	-	-	-	-
29	Saliva	10000 pg	FP	FP	FP	FP	FP
30	Cell line	320 pg	FP	FP	FP	FP	FP

Supplementary Table 1: Results of the HIrisPlex system concordance testing from five laboratories with varying background experiences including profile interpretation and comments. X indicates an incorrect genotype and drop out (d/o) was called if an allele was missing from the concordance comparison. The threshold was set at 50 RFU.

Sample	Type	Simulated casework scenario	DNA amount (pg)	Profile quality	Brown	Red	Black	Blond	Light	Dark	Blue	Int.		Analyst interpretation	Correct/Incorrect on source and prediction?
1	Frozen blood	Single source	4200	good profile	0.248	0.004	0.050	0.698	0.931	0.069	0.953	0.032	0.015	single source, dark blond hair, blue eyed individual	Correct on both
2	Frozen blood	Single source	3000	good profile	0.047	0.000	0.952	0.001	0.003	0.997	0.000	0.003	0.997	single source, black hair, brown eyed individual	Correct on both
3	Dried blood	Dried on glass: single source	130	good profile	0.119	0.137	0.005	0.739	0.996	0.004	0.891	0.068	0.041	single source, light blond hair, blue eyed individual	Correct on both
4	Semen	Single source	500	good profile	0.248	0.004	0.050	0.698	0.931	0.069	0.953	0.032	0.015	single source, dark blond hair, blue eyed individual	Correct on both
5	Saliva	Single source	240	good profile	0.275	0.010	0.007	0.708	0.988	0.012	0.953	0.032	0.015	single source, light to dark blond hair, blue eyed individual	Correct on both
6	Vaginal secretions	Single source	3000	good profile	0.864	0.087	0.028	0.021	0.432	0.568	0.951	0.043	0.006	single source, brown hair, blue eyed individual	Correct on both
7	Saliva	Single source	7600	good profile	0.000	1.000	0.000	0.000	0.951	0.049	0.252	0.165	0.583	single source, red hair, brown to intermediate eyed individual	Correct on both
8	Semen & blood	Mixture: semen & blood	1960	possible mixture										possible mixture present, no brown eye individual in mixture, unsure of hair colour due to mixture	Correct on source
9	Saliva	Single source	11900	good profile	0.103	0.862	0.005	0.030	0.886	0.114	0.492	0.294	0.214	single source,red brown auburn hair, blue to intermediate eyed individual	Correct on source, incorrect on eye colour
10	Mucus	Single source	5390	good profile	0.265	0.005	0.029	0.701	0.952	0.048	0.858	0.081	0.061	single source, dark blond hair, blue eyed individual	Correct on both
11	Saliva	Dried on glass, single source	1520	some low unbalanced peaks in profile possible mixture										unsure of both hair and eye as possible mixture	Incorrect on multiple source
12	Saliva & blood	Mixture: saliva & blood	20000	possible mixture as unbalanced peaks seen										possible mixture present, no brown eye individual in mixture, unsure of hair colour due to mixture	Correct on source
13	Hair	Anagen hair extraction	76	very low sample, quite unbalanced	0.275	0.010	0.007	0.708	0.988	0.012	0.953	0.032	0.015	single source, light to dark blond hair, blue eyed individual	Correct on both
14	Touched item	Trace DNA: single source	50	low profile some unbalanced peaks	0.518	0.000	0.367	0.115	0.210	0.790	0.014	0.051	0.935	single source, dark brown to black hair, brown eyed individual	Correct on both
15	Touched item	Trace DNA: single source	10	hardly no profile, too little DNA											-
16	Mucus & saliva	Mixture: mucus & saliva	18700	possible mixture										unsure of both hair and eye as possible mixture	Correct on source
17	Touched item	Trace DNA: single source	130	good profile	0.491	0.002	0.165	0.343	0.615	0.385	0.252	0.165	0.583	single source, dark brown hair, brown eyed individual	Correct on both
18	Frozen blood	Intentional heme inhibition (2%)	20	good profile	0.248	0.004	0.050	0.698	0.931	0.069	0.953	0.032	0.015	single source, dark blond hair, blue eyed individual	Correct on both
19	Frozen blood	Intentional heme inhibition (10%)	10	very low profile or inhibition											-
20	Touched item	Arm swab: single source	58	low profile	0.224	0.011	0.020	0.745	0.969	0.031	0.953	0.032	0.015	single source, light to dark blond hair, blue eyed individual	Correct on both
21	Saliva	single source	21000	high input good profile	0.153	0.000	0.843	0.004	0.004	0.996	0.000	0.002	0.998	single source, black hair, brown eyed individual	Correct on both
22	Vaginal secretions & semen	Mixture: vaginal swab & semen	31000	possible mixture										unsure of both hair and eye as possible mixture	Correct on source
23	Saliva & semen	Mixture: saliva & semen	1360	possible mixture										unsure of both hair and eye as possible mixture	Correct on source
24	Vaginal swab & blood	Mixture: vaginal swab & blood	14360	good profile	0.119	0.137	0.005	0.739	0.996	0.004	0.963	0.027	0.010	single source, light blond hair, blue eyed individual	Incorrect on single source therefore profile incorrect
25	Touched item	Trace DNA: multiple source	10	unbalanced peaks, possible mixture										unsure of both hair and eye as possible mixture	Correct on source

Supplementary Table 2: Simulated casework samples examined with the HIrisPlex system in a blind proficiency testing where only DNA concentrations but no sample type or preparation information were provided to the analyst. Also included are comments by analyst, concordance accuracy with known HIrisPlex profiles and final phenotype predictions of sample donors.

HIrisPlex hair & eye colour prediction tool - enhanced version 1.0

	SNP ID	Minor allele	No.
1	N29insA	**A**	0
2	rs11547464	**A**	0
3	rs885479	**T**	0
4	rs1805008	**T**	0
5	rs1805005	**T**	0
6	rs1805006	**A**	0
7	rs1805007	**T**	0
8	rs1805009	**C**	0
9	Y152OCH	**A**	0
10	rs2228479	**A**	0
11	rs1110400	**C**	0
12	rs28777	**C**	0
13	rs16891982	**C**	0
14	rs12821256	**G**	0
15	rs4959270	**A**	0
16	rs12203592	**T**	0
17	rs1042602	**T**	0
18	rs1800407	**A**	0
19	rs2402130	**G**	0
20	rs12913832	**T**	0
21	rs2378249	**C**	0
22	rs12896399	**T**	0
23	rs1393350	**T**	0
24	rs683	**G**	0

HAIR	Prob
Brown	0.218
Red	0.002
Black	0.038
Blond	0.743

SHADE	Prob
Light	0.949
Dark	0.051

EYE	Prob
Blue	0.858
Int.	0.081
Brown	0.061

* Please insert the number of minor alleles:
e.g. If major allele = A, minor allele = B
If minor allele is not observed (AA) it shall remain 0
If minor allele is heterozygote (AB) insert 1
If minor allele is homozygote (BB) insert 2

This sheet is protected for incident modifications
No password is needed to unprotect it

Supplementary Table 3: Enhanced interactive macro version 1 for HIrisPlex eye and hair colour, including shade prediction probabilities based on the enlarged underlying genotype and phenotype databases for eye and hair colour.

id	Eye colour phenotype	Prediction	PBlue	Pintermediate	Pbrown	Hair colour phenotype	Prediction	PBlond	Pbrown	PRed	PBlack	PLight	PDark
1	brown	Brown	0.10146128	0.135691241	0.76284748	black	dark brown-black	0.35675963	0.46841555	0.00164754	0.17317728	0.65230817	0.34769183
2	blue	Blue	0.91733707	0.051904319	0.03075861	dark blond	dark blond	0.68461641	0.28207963	0.0052575	0.02804646	0.94387229	0.05612771
3	blue	Blue	0.88417827	0.082985421	0.03283631	brown	dark brown-black	0.24232875	0.62207995	0.00360981	0.1319815	0.60867913	0.39132087
4	blue	Blue	0.96344032	0.026617516	0.00994216	dark blond	light blond	0.77962196	0.17241281	0.01409412	0.03387112	0.96134928	0.03865072
5	blue	Blue	0.88417827	0.082985421	0.03283631	blond	dark brown-black	0.34352889	0.52142134	0.00689855	0.12815122	0.74620964	0.25379036
6	blue	Brown	0.14280904	0.154996761	0.7021942	dark blond	dark blond	0.49358585	0.4086209	0.00340309	0.09439016	0.83877681	0.16122319
7	hazel	Brown	0.10146128	0.135691241	0.76284748	dark blond	dark blond	0.49179201	0.41984999	0.06069028	0.02766772	0.95833421	0.04166579
8	hazel	Blue	0.88417827	0.082985421	0.03283631	dark blond	dark blond	0.34935182	0.27906553	0.31936178	0.05222086	0.88777649	0.11222351
9	blue	Brown	0.1871636	0.150600211	0.66223619	dark blond	dark blond	0.4481497	0.4332327	0.00486178	0.11375582	0.78624685	0.21375315
10	blue	Blue	0.95254206	0.032427348	0.01503059	dark blond	light blond	0.78447567	0.16817961	0.00461246	0.04273226	0.95695918	0.04304082
11	hazel	Brown	0.25900147	0.235215559	0.50578297	black/dark brown	dark brown-black	0.29510244	0.54856885	0.00342495	0.15290376	0.71686941	0.28313059
12	blue	Blue	0.91733707	0.051904319	0.03075861	blond	light blond	0.71677765	0.10989947	0.16378292	0.00953996	0.98674929	0.01325071
13	brown	Brown	0.14280904	0.154996761	0.7021942	dark blond	red	0.23801331	0.06950746	0.68500166	0.00747757	0.97292846	0.02707154
14	blue	Blue	0.96344032	0.026617516	0.00994216	dark blond	light blond	0.70062229	0.2448677	0.01078747	0.04372254	0.95784338	0.04215662
15	blue	Blue	0.85815415	0.080702531	0.06114332	dark blond	dark blond	0.66131611	0.30222612	0.00926104	0.02719672	0.94204941	0.05795059
16	blue	Blue	0.93646384	0.043001243	0.02053492	dark blond	light blond	0.80668698	0.1199015	0.05994988	0.01346164	0.9851998	0.0148002
17	blue	Blue	0.92390139	0.059001059	0.01709755	blond	light blond	0.82677347	0.13604893	0.0371773	2.94E-07	1	0
18	blue	Blue	0.95724337	0.035601469	0.00715516	dark blond	dark brown	0.31206577	0.57367029	0.00609331	0.10817064	0.71645203	0.28354797
19	blue	Blue	0.85815415	0.080702531	0.06114332	dark blond	light blond	0.72406631	0.23533141	0.00338787	0.03721441	0.94748719	0.05251281
20	blue	Blue	0.89053878	0.067965751	0.04149547	dark blond	light blond	0.72404594	0.24329502	0.00677045	0.02588859	0.97181585	0.02818415
21	hazel	Brown	0.15017182	0.226672878	0.6231553	dark blond	dark blond	0.52455507	0.43569681	0.00278593	0.03696219	0.89983332	0.10016668
22	green/hazel	Brown	0.10146128	0.135691241	0.76284748	dark blond	dark blond	0.46535922	0.41219448	0.09026047	0.03218584	0.94550922	0.05449078
23	green with thin hazel ring	Blue	0.93646384	0.043001243	0.02053492	blond	dark blond	0.64015452	0.32024269	0.00500774	0.03459505	0.92659356	0.07340644
24	blue	Blue	0.95254206	0.032427348	0.01503059	dark blond	light blond	0.71091278	0.2149372	0.00151779	0.07263223	0.92860826	0.07139174
25	hazel	Blue	0.92390139	0.059001059	0.01709755	blond	dark blond maybe hint of red	0.59616907	0.13281481	0.26927161	0.00174451	0.99716712	0.00283288
26	brown	Brown	0.00088557	0.02271794	0.97639649	black	dark brown-black	0.23946382	0.58382903	0.00202233	0.17468482	0.52072626	0.47927374
27	blue	Blue	0.89053878	0.067965751	0.04149547	brown	dark blond	0.5826631	0.36067702	0.00774787	0.04891201	0.88832353	0.11167647
28	blue	Blue	0.91733707	0.051904319	0.03075861	dark blond	blond red	0.57262271	0.18815373	0.23096615	0.00825741	0.98365312	0.01634688
29	blue	Blue	0.8526117	0.066550001	0.0808383	dark blond	dark brown-black	0.38859149	0.36150545	0.04416654	0.20573652	0.65614074	0.34385926
30	grey with lots of brown spots	Blue	0.88417827	0.082985421	0.03283631	dark blond	dark brown	0.31702978	0.5824591	0.00568753	0.09482358	0.77331297	0.22668703
31	hazel	Blue	0.91733707	0.051904319	0.03075861	brown	light blond	0.73428829	0.22311049	0.00272073	0.03988049	0.95002711	0.04997289
32	green	Blue	0.85815415	0.080702531	0.06114332	brown	dark blond	0.67215249	0.28287625	0.01968543	0.02528584	0.96661551	0.03338449
33	blue	Blue	0.95254206	0.032427348	0.01503059	blond	light blond	0.72936448	0.2192453	0.00304629	0.04834394	0.95071008	0.04928992
34	blue	Blue	0.91733707	0.051904319	0.03075861	dark blond	dark blond	0.64076042	0.32552999	0.01063573	0.02307386	0.96408632	0.03591368
35	blue	Blue	0.91733707	0.051904319	0.03075861	dark blond	light blond	0.72999924	0.24575175	0.00279858	0.02145043	0.96436762	0.03563238
36	grey with hazel ring	Brown	0.14280904	0.154996761	0.7021942	d-blond	dark blond	0.60590033	0.23685228	0.14158024	0.01566715	0.98550352	0.01449648
37	green	Blue	0.85356563	0.08193291	0.06450146	dark-brown	dark brown-black	0.08240732	0.72348133	0.00209999	0.19201136	0.24679348	0.75320652
38	hazel	Brown	0.14280904	0.154996761	0.7021942	brown	dark blond-brown	0.46927395	0.39074543	0.0016539	0.13832673	0.77819355	0.22180645
39	hazel	Brown	0.14280904	0.154996761	0.7021942	red blond	blond red	0.43814778	0.18302418	0.3788273	7.50E-07	0.99999868	1.32E-06

40	hazel	Brown-intern	0.27494811	0.336137845	0.38891404	dark brown	dark brown-black	0.02278865	0.89421031	0.02466645	0.05833459	0.26736834	0.73263166
41	blue one eye with brown spot	Blue	0.85815415	0.080702531	0.06114332	d-blond with red hint	dark blond maybe hint red	0.60634351	0.20904407	0.14874362	0.0358688	0.93443171	0.06556829
42	grey/hazel	Blue	0.85815415	0.080702531	0.06114332	black	light blond	0.81995447	0.16784789	0.00324987	0.00894776	0.98661061	0.01338939
43	blue	Blue	0.95254206	0.032427348	0.01503059	red blond	red	0.06673452	0.0181099	0.91507576	7.98E-05	0.99917945	0.00082055
44	brown/black	Brown	0.07545889	0.136120833	0.78842028	dark brown	dark brown	0.41081978	0.46628191	0.0014094	0.12148891	0.738603	0.261397
45	blue	Blue	0.89053878	0.067965751	0.04149547	blond	dark blond maybe hint red	0.66915611	0.14013768	0.1820496	0.00865661	0.99230375	0.00769625
46	brown	Brown	0.14280904	0.154996761	0.7021942	black	dark brown-black	0.39271907	0.40728841	0.00229501	0.19769751	0.73128489	0.26871511
47	blue	Blue	0.85815415	0.080702531	0.06114332	blond	dark blond	0.69561552	0.19583001	0.07777935	0.03077512	0.96442388	0.03557612
48	blue	Blue	0.41095974	0.245806847	0.34323342	blond	dark blond	0.54937902	0.37927158	0.00589768	0.06545172	0.91829538	0.08170462
49	brown	Brown	0.05208509	0.115774317	0.83214059	blond	dark blond-brown	0.49358585	0.4086209	0.00340309	0.09439016	0.83877681	0.16122319
50	brown	Brown	0.10146128	0.135691241	0.76284748	black	dark brown-black	0.30280866	0.51734497	0.00406624	0.17578013	0.57235384	0.42764616
51	brown	Brown	0.05208509	0.115774317	0.83214059	brown	dark blond	0.55046155	0.35830737	0.00318523	0.08804586	0.85383225	0.14616775
52	blue	Blue	0.95254206	0.032427348	0.01503059	blond	light blond	0.81995447	0.16784789	0.00324987	0.00894776	0.98661061	0.01338939
53	brown	Brown	0.00109733	0.034618306	0.96428436	brown	dark brown-black	0.03820918	0.54696861	0.00097474	0.41384746	0.08058759	0.91941241
54	hazel	Brown	0.15990633	0.213430875	0.6266628	black	dark brown-black	0.11262767	0.73427356	0.00342138	0.14967739	0.45697	0.54303
55	hazel	Brown	0.05208509	0.115774317	0.83214059	blond	dark brown	0.21219013	0.70626797	0.01938227	0.06215963	0.75804465	0.24195535
56	hazel	Brown	0.10146128	0.135691241	0.76284748	brown	dark blond-brown	0.48261963	0.46031889	0.00351711	0.05354437	0.90516927	0.09483073
57	green	Blue	0.85815415	0.080702531	0.06114332	blond	light blond	0.72404594	0.24329502	0.00677045	0.02588859	0.97181585	0.02818415
58	brown	Brown	0.00058189	0.01839379	0.98102432	brown	dark brown-black	0.10466127	0.62000715	0.00097062	0.27436096	0.18692046	0.81307954
59	blue	Blue	0.88417827	0.082985421	0.03283631	brown	brown	0.31468876	0.59640934	0.00493681	0.08396509	0.76069624	0.23930376
60	brown	Brown	0.10146128	0.135691241	0.76284748	black	brown	0.42973257	0.48493765	0.0028841	0.08244568	0.80732245	0.19267755
61	hazel	Brown	0.05208509	0.115774317	0.83214059	black	dark blond-brown	0.45207678	0.43821753	0.00219867	0.10750703	0.79313673	0.20686327
62	blue	Blue	0.91733707	0.051904319	0.03075861	blond	dark blond	0.64239425	0.32438065	0.01225714	0.02096796	0.97012588	0.02987412
63	blue	Blue	0.93646384	0.043001243	0.02053492	blond	light blond	0.75057464	0.15768614	0.07840921	0.01333	0.9820158	0.0179842
64	hazel	Brown	0.1871636	0.150600211	0.66223619	blond	dark blond-brown	0.45207678	0.43821753	0.00219867	0.10750703	0.79313673	0.20686327
65	blue	Blue	0.91733707	0.051904319	0.03075861	dark blond	dark blond	0.69236116	0.2554207	0.00261322	0.04960493	0.93451166	0.06548834
66	blue	Blue	0.80569977	0.077491592	0.11680863	blond	dark blond-brown	0.42338458	0.46238106	0.08015587	0.03407849	0.93696199	0.06303801
67	hazel	Blue	0.48913029	0.216898488	0.29397123	blond	dark blond brown	0.51046828	0.37312382	0.00153172	0.11487618	0.82036087	0.17963913
68	brown/black	Brown	0.00058189	0.01839379	0.98102432	black	brown	0.28691451	0.66407931	0.04900195	4.23E-06	0.99998465	1.53E-05
69	blue	Blue	0.95254206	0.032427348	0.01503059	dark blond	dark blond	0.66609321	0.28300934	0.00391067	0.04698679	0.94055391	0.05944609
70	blue	Blue	0.93646384	0.043001243	0.02053492	blond	light blond	0.78108087	0.20025517	0.00969612	0.00896784	0.98409542	0.01590458
71	blue	Blue	0.95950839	0.032599892	0.00789172	dark blond	dark blond-brown	0.37714054	0.50552579	0.08615166	0.031182	0.9484152	0.0515848
72	blue	Blue	0.85815415	0.080702531	0.06114332	blond	dark blond	0.65989889	0.30143579	0.01339847	0.02526685	0.96382244	0.03617756
73	blue	Blue	0.85815415	0.080702531	0.06114332	blond	light blond	0.75226371	0.21582218	0.0099256	0.02198851	0.9630935	0.0369065
74	hazel	Brown	0.15017182	0.226672878	0.6231553	brown	dark blond-brown	0.46457764	0.43852712	0.00553559	0.09135965	0.85414465	0.14585535
75	blue	Blue	0.93646384	0.043001243	0.02053492	blond	light blond	0.75440972	0.20940601	0.00500403	0.03118024	0.96075795	0.03924205
76	hazel	Brown	0.05208509	0.115774317	0.83214059	brown	dark blond-brown	0.41081978	0.46628191	0.0014094	0.12148891	0.738603	0.261397
77	blue	Blue	0.93646384	0.043001243	0.02053492	brown	light blond	0.72406631	0.23533141	0.00338787	0.03721441	0.94748719	0.05251281
78	blue	Blue	0.93646384	0.043001243	0.02053492	dark blond	light blond	0.82429696	0.10467476	0.07102804	2.38E-07	1	0
79	blue	Blue	0.95254206	0.032427348	0.01503059	dark blond	light blond	0.75735212	0.22722411	0.0060933	0.00933047	0.98401644	0.01598356

80	blue	Blue	0.95254206	0.032427348	0.01503059	blond	light blond	0.72936448	0.2192453	0.00304629	0.04834394	0.95071008	0.04928992
81	blue	Blue	0.85815415	0.080702531	0.06114332	brown	dark blond	0.67002559	0.28775631	0.00348979	0.03872831	0.93973913	0.06026087
82	green with hazel ring	Blue	0.89053878	0.067965751	0.04149547	dark blond	dark blond	0.69173059	0.27139319	0.00644969	0.03042653	0.95256885	0.04743115
83	hazel	Brown	0.07545889	0.136120833	0.78842028	brown	dark blond-brown	0.39426989	0.43232139	0.11133883	0.06206989	0.81626546	0.18373454
84	hazel	Brown	0.21523426	0.233140886	0.55162485	dark blond	dark brown-black	0.09231201	0.66073497	0.00113189	0.24582113	0.25830414	0.74169586
85	blue	Blue	0.91733707	0.051904319	0.03075861	dark blond	dark blond	0.69173059	0.27139319	0.00644969	0.03042653	0.95256885	0.04743115
86	hazel	Brown	0.00027503	0.014449425	0.98527555	brown	dark brown-black	0.1395345	0.54640316	0.00066979	0.31339255	0.28503572	0.71496428
87	blue	Blue	0.91733707	0.051904319	0.03075861	blond	dark blond	0.66404366	0.2855331	0.04210635	0.00831689	0.98261645	0.01738355
88	brown/black	Brown	0.34991621	0.228046708	0.42203709	dark blond	dark brown-black	0.09618656	0.74998248	0.00786054	0.14597042	0.51639012	0.48360988
89	blue/brown	Blue	0.91733707	0.051904319	0.03075861	dark blond	dark blond	0.56575179	0.22771098	0.20103295	0.00550428	0.98984966	0.01015034
90	blue	Blue	0.91733707	0.051904319	0.03075861	blond	light blond	0.72128732	0.21684234	0.02831318	0.03355716	0.95053849	0.04946151
91	hazel	Brown	0.00027503	0.014449425	0.98527555	blond	dark brown-black	0.12299472	0.59420984	0.00138889	0.28140655	0.32036382	0.67963618
92	blue	Blue	0.91733707	0.051904319	0.03075861	dark blond	light blond	0.78447567	0.16817961	0.00461246	0.04273226	0.95695918	0.04304082
93	brown/black	Brown	0.10146128	0.135691241	0.76284748	brown	brown	0.27128022	0.64815309	0.00613293	0.07443377	0.7531111	0.2468889
94	blue	Blue	0.95254206	0.032427348	0.01503059	dark blond	light blond	0.70664549	0.13685739	0.12038889	0.03610823	0.95617956	0.04382044
95	blue	Blue	0.90871411	0.069215709	0.02207019	dark blond	dark blond-brown	0.43158301	0.53419827	0.00828221	0.02593651	0.92968864	0.07031136
96	blue	Blue	0.93114493	0.052581378	0.01627369	black	brown	0.34060697	0.57328712	0.0077385	0.07836742	0.84753324	0.15246676
97	blue	Blue	0.9154071	0.056697904	0.02789499	dark blond	light blond	0.78867738	0.19308541	0.00840924	0.00982796	0.9894178	0.0105822
98	blue	Blue	0.93646384	0.043001243	0.02053492	dark blond	dark blond	0.59289339	0.36011484	0.00472452	0.04226725	0.90453363	0.09546637
99	green with hazel ring	Blue	0.93646384	0.043001243	0.02053492	dark blond	light blond	0.76713341	0.1388643	0.06621772	0.02778458	0.96999342	0.03000658
100	blue	Blue	0.91733707	0.051904319	0.03075861	dark blond	dark blond	0.69763153	0.21886315	0.02200132	0.061504	0.90720915	0.09279085
101	blue/brown	Blue	0.41095974	0.245806847	0.34323342	dark blond	dark blond	0.60211414	0.36000274	0.00196867	0.03591445	0.93000325	0.06999675
102	hazel	Brown	0.07545889	0.136120833	0.78842028	dark blond	dark brown	0.41607645	0.42064001	0.00149375	0.16178978	0.72478295	0.27521705
103	hazel	Brown	0.00217225	0.046628423	0.95119933	brown	brown	0.2780205	0.59549788	0.02051006	0.10597156	0.79144793	0.20855207
104	blue	Blue	0.91733707	0.051904319	0.03075861	dark blond	dark blond	0.6751464	0.20039411	0.11575114	0.00870834	0.99101074	0.00898926
105	hazel	Brown	0.14280904	0.154996761	0.7021942	black	dark brown-black	0.2696382	0.57715267	0.00326081	0.14994833	0.65720907	0.34279093
106	green with hazel ring	Brown	0.3317714	0.244522149	0.42370645	brown	dark blond-brown	0.50034617	0.3912037	0.00189589	0.10655425	0.81253102	0.18746898
107	blue	Blue	0.91733707	0.051904319	0.03075861	blond	dark blond	0.66769293	0.27323983	0.00725184	0.05181539	0.9436448	0.0563552
108	blue	Brown	0.1871636	0.150600211	0.66223619	blond	dark blond	0.64126987	0.29505023	0.06367859	1.31E-06	0.99999834	1.66E-06
109	blue	Blue	0.90871411	0.069215709	0.02207019	black	dark brown-black	0.32354259	0.59134114	0.01235855	0.07275772	0.78933387	0.21066613
110	brown/black	Brown	0.10146128	0.135691241	0.76284748	black	dark blond	0.66021681	0.26707483	0.05180131	0.02090706	0.96406515	0.03593485
111	brown/black	Brown	0.20248093	0.248033697	0.54948537	black	dark brown-black	0.38765824	0.50162728	0.00145652	0.10925797	0.72376861	0.27623139
112	blue	Blue	0.91733707	0.051904319	0.03075861	black	dark blond	0.64577076	0.17518083	0.16976011	0.0092883	0.9829597	0.0170403
113	blue	Brown	0.15017182	0.226672878	0.6231553	black	brown	0.39036583	0.55540276	0.00724106	0.04699035	0.87573144	0.12426856
114	hazel	Brown	0.07545889	0.136120833	0.78842028	black	dark brown-black	0.32051724	0.45940856	0.00184435	0.21822985	0.58291424	0.41708576
115	blue	Blue	0.91733707	0.051904319	0.03075861	black	dark blond	0.55250294	0.39796067	0.01469914	0.03483725	0.9344104	0.0655896
116	blue	Blue	0.89053878	0.067965751	0.04149547	black	dark blond	0.69236116	0.2554207	0.00261322	0.04960493	0.93451166	0.06548834
117	hazel	Brown	0.14280904	0.154996761	0.7021942	black	dark brown-black	0.33524923	0.45155016	0.00252014	0.21068046	0.65971843	0.34028157
118	blue	Blue	0.75120096	0.097454119	0.15134492	black	dark blond	0.6475061	0.29024399	0.02380117	0.03844874	0.95086062	0.04913938
119	blue	Blue	0.91733707	0.051904319	0.03075861	black	dark blond-brown	0.51126047	0.44343389	0.02587265	0.01943299	0.9477928	0.0522072

Supplementary Table 4: Test set for accuracy of enhanced HIrisPlex model (Supplementary Table 3) (n = 119).

Chapter 3.2

Bringing colour back after 70 years: Predicting eye and hair colour from skeletal remains of World War II victims using the HIrisPlex system

Forensic Science International: Genetics 26 (2016) 48 – 57

Abstract

Retrieving information about externally visible characteristics from DNA can provide investigative leads to find unknown perpetrators, and can also help in disaster victim and other missing person identification cases. Aiming for the application to both types of forensic casework, we previously developed and forensically validated the HIrisPlex test system enabling parallel DNA prediction of eye and hair colour. Although a recent proof-of-principle study demonstrated the general suitability of the HIrisPlex system for successfully analyzing DNA from bones and teeth of various storage times and conditions, practical case applications to human remains are scarce. In this study, we applied the HIrisPlex system to 49 DNA samples obtained from bones or teeth of World War II victims excavated at six sites, mostly mass graves, in Slovenia. PCR-based DNA quantification ranged from 4 pg/µl to 313 pg/µl and on an average was 41 pg/µl across all samples. All 49 samples generated complete HIrisPlex profiles with the exception of one *MC1R* DNA marker (N29insA) missing in 83.7% of the samples. In 44 of the 49 samples (89.8%) complete 15-loci autosomal STR (plus amelogenin) profiles were obtained. Of 5 pairs of skeletal remains for which STR profiling suggested an origin in the same individuals, respectively, 4 showed the same HIrisPlex profiles and predicted eye and hair colours, respectively. Discrepancies in one pair (sample 26 and 43) are likely to be explained by DNA quantity and quality issues observed in sample 43. Sample 43 had the lowest DNA concentration of only 4 pg/µl, producing least reliable STR results and could be misleading in concluding that samples 43 and 26 originate from the same individual. The HIrisPlex-predicted eye and hair colours from two skeletal samples, suggested to derive from two brothers via STR profiling together with a living sister, were confirmed by the living sister's report. Overall, we demonstrate that after more than 70 years, HIrisPlex-based eye and hair colour prediction from skeletal remains is feasible with high success rate. Our results further encourage the use of the HIrisPlex system in missing person/disaster victim identification to aid the identification process in cases where ante-mortem samples or putative relatives are not directly available, and DNA predicted eye and hair colour information provides leads for locating them, allowing STR based individual identification.

Introduction

DNA analysis from human remains is of immense relevance in missing persons identification and disaster victim identification (DVI) [1–4]. Being a resilient molecule, DNA degrades gradually in hard tissues, such as bones and teeth, thereby empowering the extraction of DNA from remains under favourable storage conditions [5]. Temperature, humidity, pH, geochemical properties of the soil, the presence of micro-organisms, storage time and other factors all affect the preservation of DNA in skeletal remains [6,7]. Prevention and detection of contamination with exogenous DNA are critical in forensic DNA studies of human remains, and shall be guided by experiences in the field of ancient DNA analysis typically dealing with much older remains where consequently problems of degradation, inhibition, and contamination typically increase [8,9]. Success of DNA typing of skeletal remains is further influenced by the types of bone specimen available, the characteristics of the soil and levels of humidity in which they are found, the post-mortem taphonomic process and the time they have been buried and other external and internal factors [10,11]. Recent advances in massively parallel sequencing (MPS), also referred to as next generation sequencing (NGS), provides some advantages over previously used technologies to analyse DNA from human remains, particularly ancient ones [12–14]. It is true that the use of NGS would augment the analysis time and reduce the consumption of reagents and samples. However, the overall cost is too high and may not be affordable for routine use by the forensic labs. Further, the lack of population data on a lot of STRs and SNPs analyzed with NGS kits is an disadvantage for its routine applications. Often in ancient DNA NGS studies, the overall genome coverage is quite low, therefore any particular SNP needed for the prediction could also be low. Hence, though NGS seems to be feasible, it comes with certain disadvantages.

In disaster victim identification (DVI) and other missing person identification cases, STR (or SNP) profiling from human remains is typically applied for matching with ante-mortem samples of the missing individual and with DNA from putative relatives that can be used for relationship testing based on STR/SNP profile similarity [2,15–17]. However, the more time has passed since the disaster event or the time when persons got missing, the less likely it is that ante-mortem samples and putative relatives can be located to make them available for STR/SNP profiling, with World War II (WWII) victims serving as suitable example. In such cases, including those from WWII [18–21] lineage markers from

mitochondrial DNA or from the non-recombining part of the Y-chromosome (in case of male individuals) provide useful information to find distant relatives on the maternal and paternal side, respectively. Moreover, the recent establishment of Forensic DNA Phenotyping has made it viable to predict externally visible characteristics (EVCs) of an individual from DNA [22,23]. Besides its application to crime scene stains aiming to find unknown perpetrators not identifiable via comparative STR (likewise SNP) profiling, FDP is also suitable for missing person and DVI, particularly in cases where ante-mortem samples or putative relatives are not directly available but FDP outcomes allow guiding to locate them via dedicated investigation to make them available for STR/SNP-based identification of the missing.

Recently, the HIrisPlex test system for parallel eye and hair colour prediction from DNA using 24 DNA markers and two prediction models, one for eye colour and one for hair colour based on reference datasets of thousands of Europeans was developed, improved and forensically validated [24,25]. Although having more EVCs predictable from DNA, as currently is underway [23], would certainly enhance the use of FDP for unknown perpetrator and missing person identification, eye and hair colour serve as a start in case of European persons involved, and can already provide valuable guidance as it already alone reduces the number of possible persons. In a proof-of-principle study, the general suitability of the HIrisPlex test system to successfully analyse DNA from bones and teeth of various ages stored under various conditions was previously demonstrated [17]. Recently, the HIrisPlex markers analysed via dedicated ancient DNA technologies in the skeletal remains attributed by multiple lines of evidence to King Richard III of England (1452 to 1485), the last king of England killed on the battlefield more than 500 years ago. HIrisPlex DNA markers and prediction modelling revealed a blue eye probability of 96% and a blond hair probability of 77% [26]. Based on these findings the authors concluded that the King's eye colour most likely was blue and his most likely hair colour was blond or dark blond/brown, the latter in case he changed the hair colour during adolescence, which HIrisPlex (or any other hair colour DNA test currently available) cannot find out. However, studies that investigate the suitability of the HIrisPlex system, including its genotyping assay, to human remains for the purpose of missing person or disaster victim identification are scarce as of yet.

In the present study, we applied the HIrisPlex system to DNA samples obtained from 49 bones or teeth from WWII victims excavated at several sites in Slovenia, mostly mass graves, for which we also established forensic STR data. Aims of our study were to

further investigate the suitability of the HIrisPlex system for DVI purposes in general, and to potentially aid the identification process of these victims in particular. Although 70 years have passed since the end of WWII, identification of WWII victims is still relevant as a large number of individuals killed during WWII are still missing and remain to be identified. In Slovenia for instance, the Commission on Concealed Mass Graves registered almost 600 mass graves with approximately 100,000 victims [27].

Materials and methods

Bone and teeth samples

The skeletal remains were excavated from six different sites within Slovenia, five mass graves from WWII - Konfin I and Konfin II (both located near Grčarice pri Ribnici), Storžič, Bodovlje Gorge (Bodoveljska Grapa), and Mozelj, as well as one individual grave - Nova Gorica where mass executions took place at the end of WWII. Locations of the grave sites are indicated in Figure 1.

Figure 1: Map of Slovenia with indicated locations of the WWII grave sites investigated.

The following bone and teeth specimens were analysed from the 6 grave sites: 15 femurs and 8 tibias from Konfin I, 10 teeth (molars 1, 2 and 3 from upper and lower jawbones) and 7 femurs from Konfin II, 2 femurs and 1 molar (molar 2) from Storžič, 3 femurs from Bodovlje Gorge, 2 femurs from Mozelj and 1 femur from the individual grave at Nova Gorica. A total of 38 bone samples (30 femurs and 8 tibias) and 11 teeth samples (five molars 3, four molars 2 and two molars 1) were genotyped in this study. For genetic investigations, 8 to 10 cm bone fragment was taken from the compact cortical diaphysis of each femur and tibia. Molars were removed from the upper and lower jawbones and whole teeth were used for grinding. Fragments of bones and teeth sampled for DNA analyses were frozen at -20°C until the DNA extraction occurred.

Attempts to minimize contamination

Special procedures to reduce the risk of contamination with exogenous DNA are imperative when studying old human remains. In order to reduce contamination from previous handling, we performed washing in bi-distilled water, detergent and ethanol [28]; radiation with UV light and removing the bone surface by drilling and acquiring the bone material directly from the inside of the specimen [29]. Extractions were performed with strict precautions, including protective clothing (sterile disposable coat, cap, mask, double latex gloves); equipments and surfaces were treated with bleach and irradiated with UV light [8]. All sample manipulations were performed in laminar flow cabinets equipped with HEPA filters and UV lights, using dedicated pipettes, disposable sterile filter tips and sterile tubes. Between each sample, all the tools for drilling, cutting, and grinding were cleaned by washing with bleach, sterile bi-distilled water and 80% ethanol, sterilised and UV irradiated for 72 hours [30]. Pre and post-PCR work were carried out in separate rooms [31]. Within the pre-PCR laboratory, separate working localites are confined for each step in the bone typing procedure [32]. Cleaning and grinding the bones took place in a closed safety cabinet MC 3 (Iskra Pio, Šentjernej, Slovenia, EU). DNA-based contamination monitoring was applied in all stages using blank controls. Extraction-negative controls were processed in every batch of extraction and PCR-negative controls in every amplification reaction to verify the purity of the extraction and amplification reagents and plastics [33]. Since contamination of the endogenous DNA of bones and teeth with modern DNA can occur during exhumation, storage of the skeletal remains, anthropological investigations, and molecular genetic typing, we created an

elimination database for each WWII mass grave. Elimination DNA databases, containing autosomal STR profiles of the individuals that participated in the exhumation, subsequent handling of the remains, DNA extraction and final DNA genotyping were used to check the authenticity of the obtained DNA profiles. Bone and tooth sample analyses were physically and temporally separated from analyses of samples for the elimination database.

DNA extraction

A full demineralisation protocol employing 0.5 M EDTA was used for extraction of DNA from skeletal remains. Bone and tooth sample cleaning and powdering was performed according to Zupanič Pajnič et al., (2016) [34] and DNA was extracted from powder as described elsewhere by Zupanič Pajnič (2016) [35]. In brief, genomic DNA was obtained from 0.5 g of bone or tooth powder incubated in 10 ml of 0.5 M EDTA (Promega, Madison, WI, USA), pH 8.0 overnight at 37°C in a Thermomixer comfort (Eppendorf, Hamburg, Germany, EU), shaken at 750 rpm. After centrifugation at 2,800 rpm for 15 minutes in a Megafuge 1.0 centrifuge (Heraeus, Hanau, Germany) the supernatant was discarded. Ten ml of ultrapure distilled water (Gibco, Carlsbad, CA, USA) was added to the precipitate and vortexed at high speed for 10 seconds. The centrifugation step was repeated and the supernatant discarded. Twenty µl freshly prepared 1 M DTT (Sigma-Aldrich), 100 µl G2 buffer and 60 µl proteinase K (both from the EZ1 DNA Investigator Kit (Qiagen, Hilden, Germany) [36] was added to the precipitate and incubated at 56°C in a Thermomixer comfort (Eppendorf) for 2 hours, shaken at 750 rpm. There was no undigested bone or tooth powder left after the demineralisation and lysis stage. Lysis was followed by centrifugation at 1,900 rpm for 2 minutes in a Megafuge 1.0 centrifuge (Heraeus). Up to 400 µl of the supernatant was transferred to the sample tube and 1 µl of cRNA (EZ1 DNA Investigator Kit, Qiagen) was added. The DNA was purified in a Biorobot EZ1 (Qiagen) device using the EZ1 DNA Investigator Card and EZ1 DNA Investigator Kit (Qiagen). Following the manufacturer's instructions (Qiagen Companies, 2014), the Biorobot EZ1 was used to obtain genomic DNA from decalcified bone or tooth precipitate using the trace protocol and DNA from the buccal swab samples collected for the elimination database was obtained using the "tip dance" protocol. The final volume of 50 µl of the colourless bone and tooth extracts was obtained. The extraction negative controls were included in the extraction process to verify the purity of the extraction reagents and plastics.

DNA quantification

In old and degraded samples, it is critical to assess the quantity and quality of human DNA that is available for genetic analysis. In order to determine the quantity of nuclear DNA and the presence of inhibitors of the PCR reaction, the DNA extracts from bone and teeth samples were quantified by real-time PCR using the Quantifiler Human DNA Quantification Kit (Applied Biosystems, Foster City, CA, USA), which includes an internal PCR control to monitor inhibition of the PCR in the extracts. The reactions were carried out using the 7500 Real Time PCR System (Applied Biosystems) with the HID Real-Time PCR Analysis Software, version 1.1 (Applied Biosystems) according to the manufacturer's recommended procedures (Applied Biosystems, 2003). Negative and positive controls were included. To determine the relative level of inhibitors present in the extracts, shifts in the Ct value for the internal PCR control (IPC) in the real-time PCR assay were checked.

Autosomal STR typing

Genotyping of the same 15 autosomal STRs and amelogenin was performed using two commercially available kits, the Investigator ESSplex Plus Kit (Qiagen) [37] or the AmpFlSTR NGM PCR Amplification Kit (Applied Biosystems). The amplification protocols and the thermal cycling conditions were according to the manufacturer's instructions (Applied Biosystems, 2009; Qiagen Companies, 2013). The Nexus Master Cycler (Eppendorf) was used for PCR amplification of DNA. Simultaneously with the forensic samples, we amplified the positive control (Control DNA 007 with NGM and 9948 with ESSplex kit) and negative PCR control, as well as the extraction negative controls where the maximum volume of extracts was used for amplification. The fluorescent-labelled PCR products were separated using a 10 seconds injection time and a 3 kV injection voltage on an automatic ABI PRISM 3130 Genetic Analyser (Applied Biosystems) using the 3130 Performance Optimized Polymer 4 (Applied Biosystems) and the GeneScan-500 LIZ (Applied Biosystems) internal size standard for NGM amplification products and DNA size standard 550 (BTO) (Qiagen) for ESSplex Plus amplification products. The genetic profiles were determined using the Data Collection v 3.0 and GeneMapper ID v 3.2 software (Applied Biosystems) with a 50 relative fluorescence units (RFU) peak amplitude threshold for all dyes. The genotypes were confirmed with previously obtained profiles [38] for all of the bones and teeth from Mozelj, Storžič, Bodovlje Gorge, Konfin I and Konfin II mass graves. Since all of the

previously extracted 50 µl were used-up in the previous study [38], we performed new DNA extractions from bone and tooth powder. The same DNA extracts were used for STR and HIrisPlex genotyping. STR typing of the same loci was also carried out in DNA of persons included in the elimination databases using the AmpFLSTR Identifiler PCR Amplification Kit (Applied Biosystems) or AmpFlSTR NGM PCR Amplification Kit (Applied Biosystems) kit according to the manufacturer's instructions. All STR profiles obtained from the skeletal remains were compared to those collected in the elimination databases to monitor DNA contamination (see below).

HIrisPlex genotyping

The robust, single multiplex HIrisPlex assay targets 24 DNA markers, 6 of which are used for eye colour prediction overlapping with the previously established IrisPlex assay: rs12913832 (*HERC2*), rs1800407 (*OCA2*), rs16891982 (*SLC45A2 or MATP*), rs12203592 (*IRF4*), rs12896399 (*SLC24A4*), and rs1393350 (*TYR*), while the former 4 plus the following 18 are used for hair colour and hair shade prediction: Y152OCH, N29insA (INDEL), rs1805006, rs11547464, rs1805007, rs1805008, rs1805009, rs1805005, rs2228479, rs1110400 and rs885479 all from the *MC1R* gene, rs1042602 (*TYR*), rs4959270 (*EXOC2*), rs28777 (*SLC45A2 or MATP*), rs683 (*TYRP1*), rs2402130 (*SLC24A4*), rs12821256 (*KITLG*), and rs2378249 (*PIGU*). HIrisPlex genotyping includes a multiplex PCR amplification of all 24 DNA markers simultaneously, followed by a multiplex single base extension reaction using the SNaPshot chemistry (Applied Biosystems) and the product are finally analysed via capillary electrophoresis using the ABI 3130xl DNA Analyzer (Applied Biosystems). All HIrisPlex marker details, primer sequences, the multiplex protocol, as well as the forensic validation of the assay are described in detail elsewhere [24,25]. All DNA samples were multiplex genotyped at least three times. As the marker N29insA failed to give results in the multiplex reaction in many samples, a singleplex reaction was done with this marker for all these samples. Additionally, markers N29insA (1), rs1805008 (4), rs1805007 (7), rs4959270 (15), rs12913832 (20) and rs683 (24) were run in a singleplex for the samples 14, 15, 18, 20, 26, 29, 37, 42, 43, 44 and 45 due to marker problems observed with multiplex analysis.

Model-based HIrisPlex eye and hair colour prediction

Individual probability estimation for the three eye colour categories blue, intermediate and brown was done using the genotype data from 6 IrisPlex SNPs included in the HIrisPlex assay as described elsewhere [24,39] by applying the enhanced IrisPlex eye colour prediction model based on genotype and phenotype data of 9,188 Europeans as previously described [25], which is available via the publically accessible Website http://hirisplex.erasmusmc.nl/. Final eye colour statement was done based on the category with the highest prediction probability as recommended previously [25,39]. As previously emphasized [39–41] the IrisPlex DNA markers and prediction model are highly accurate in predicting blue and brown eye colour but much less accurate for non-blue and non-brown colours typically grouped into the intermediate eye colour category.

Individual probability estimation for the four hair colour categories red, blond, brown and black as well as for the hair colour shade categories light and dark was done using the 22 hair colour predictive DNA variants included in the HIrisPlex assay as described elsewhere [24], and applying the enhanced HIrisPlex hair colour prediction model based on genotype and phenotype data of 1601 Europeans as previously described [25] and is available via the website http://hirisplex.erasmusmc.nl/. Final hair colour statements were achieved with the prediction guide as described previously [24]. As emphasized before [24,25], the HIrisPlex system (or any other currently available DNA test for hair colour) is unable to detect age-dependent hair colour changes; which practically means that if blond hair is concluded from a DNA sample of an adult individual with high enough probability, this individual most likely was blond as child but not necessarily is blond anymore as adult in case considerable darkening had occurred during adolescence that has led to a change in hair colour category, most often seen from blond to light brown or brown.

Results and discussion

Quantifiler-based DNA quantity and STR-based DNA quality of the samples

DNA was extracted from a total of 49 bone and teeth samples collected from five WWII mass graves and an individual grave where mass executions took place at the end of WWII (see methods). The ability to remove PCR inhibitors from sample extracts was monitored by the Ct values for the internal PCR control (IPC); no inhibition was detected in any of the

49 DNA extracts analysed. The difference between the IPC Ct of the samples and of NTCs was within +/- 1 Ct unit in all the quantification plates, indicating that the PCR inhibitors were effectively removed during the extractions.

Results of the PCR-based human autosomal DNA quantification using the Quantifiler Human DNA Quantification Kit (Applied Biosystems) are summarized in Table 1. The detection limit of the Quantifiler assay used is reported by the manufacturer with 23 pg/µl (Applied Biosystems, 2003). We measured DNA quantities above 23 pg/µl DNA in 27 of the 49 DNA extracts (Table 1). Seven bones and one tooth samples showed DNA quantification values of only 10 picograms (pg) or less (Table 1).

Effective measures were taken to prevent contamination during excavations and genotypings (see methods). In no case a STR match between a bone/teeth DNA sample and a person collected in the elimination databases was obtained. From 44 of the 49 DNA extracts tested (89.8%), we obtained a complete 16-locus STR/amelogenin profile using the Investigator ESSplex Plus Kit (Qiagen) [37] or the AmpFlSTR NGM PCR Amplification Kit (Applied Biosystems), the remaining 5 samples revealed partial STR profiles (Table 1). The STR results of the analysed bones and teeth were compared to those of the same skeletal remains obtained from a previously performed efficiency study of three autosomal amplification kits; NGM (Applied Biosystems), PowerPlex ESX 17 (Promega) and Investigator ESSPlex (Qiagen) [38] in order to check for drop-outs and drop-ins. In five samples drop-outs were observed (there were up to five drop-out alleles noted), and in two samples drop-ins were seen (Table 1). Notably, no drop-ins were found in the extraction negative control or in the PCR negative controls included on the analysis plates.

Allele and locus drop-outs and drop-ins appeared correlated with low DNA quantities, but not completely so. On one hand, four of the five samples with observed partial STR profiles had Quantifiler-measured DNA quantities of 10 pg/µl or lower. On the other hand, four other samples with 10 pg/µl or lower DNA concentration showed complete 16 loci STR/amelogenin profiles without signs of drop-ins/drop-outs. Of the twenty two DNA samples measured below 23 pg/µl, which is reported by the manufacturer as detection limit of the PCR-based DNA quantification system used, eighteen (82%) revealed a complete 16-loci STR/amelogenin profile.

The STR genotyping results suggests that some of the bones or teeth from which DNA samples were extracted belonged to the same individuals, respectively, which is not

unexpected in mass grave exhumations where the particular situation under which the bones were excavated did not necessarily allow to perform a bone–individual matching from the bone location alone. Out of five pairs of skeletal elements analysed, in two pairs no discrepancies were noted, two pairs showed identical profile with allele drop-out at one locus and one pair identical profile with allele drop-outs at two loci. From Konfin I mass grave, sample 20 and sample 14 showed identical 16-loci STR/amelogenin profiles. Although Investigator ESSPlex Plus STR typing of sample 20 revealed three drop-outs (D21S11, D8S1179 and D2S1338), no drop-outs were seen when additionally typing this sample with the PowerPlex ESX 17 (Promega) kit, providing the identical 16-locus STR/amelogenin profile as seen in sample 14. Identical Y-STR and mtDNA profiles obtained in 2009 when the identification of victims from Konfin I mass grave was performed [18] suggested that both samples originate from the same paternal and maternal lineage, and the autosomal STR-typing results refine this to the same individual. Furthermore, sample 18 and sample 15 from the Konfin I mass grave showed identical ESSPlex Plus STR profiles except for one allele drop-out observed for sample 18 at D2S1338. Identical Y-STR and mtDNA profiles [18] suggest that sample 18 and 15 originate from the same paternal and maternal lineage, which together with the obtained autosomal STR results suggests that both samples originate from the same individual.

Besides these two pairs of bones from Konfin I mass grave, there were another three pairs of bones and teeth from Konfin II mass grave with identical STR/amelogenin results, respectively. The victims of Konfin II mass grave were not identified yet and there are no results of Y-STR and mtDNA typing available for comparison. The same STR profile was obtained with NGM (sample 45) and ESSPlex Plus (sample 29) STR kits for sample 45 and sample 29 suggesting that both samples likely belong to the same individual. Sample 42 and sample 49 showed identical NGM STR profiles except for two allele drop-outs observed for sample 42 at D18S51 and D12S391 suggesting that both samples likely belong to the same individual. The last pair from the Konfin II mass grave was samples 43 and 26. Sample 43 after typing with Investigator ESSPlex Plus STR kit showed five drop-outs (vWA, D8S1179, D2S1338, D22S1045, and FGA) of which four were not seen when genotyping the sample with the NGM STR kit in an additional analysis. Besides this one drop-out (D8S1179), the STR profile was identical to that of sample 26 where no drop-outs were observed after ESSPlex Plus STR typing. Notably, sample 43 was the most problematic and challenging

of all DNA samples analysed in this study. The extractions were repeated 6 times and only 2 extracts delivered autosomal STR profiles, and with its measured 4 pg/µl DNA concentration, sample 43 belongs to the two lowest concentrated DNA samples analysed here.

Sample No	Skeletal material	Grave site	DNA Quantity [pg/µl]*	STR summary results #
Sample 1- brother 1	Femur	Konfin 1	20	16/0/0
Sample 2	Femur	Konfin 1	43	16/0/0
Sample 3	Femur	Konfin 1	25	16/0/0
Sample 4- brother 2	Femur	Konfin 1	24	16/0/0
Sample 5	Femur	Konfin 1	25	16/0/0
Sample 6	Femur	Konfin 1	28	16/0/0
Sample 7	Femur	Konfin 1	47	16/0/0
Sample 8	Femur	Konfin 1	25	16/0/0
Sample 9	Femur	Konfin 1	26	16/0/0
Sample 10	Femur	Konfin 1	18	16/0/0
Sample 11	Femur	Konfin 1	26	16/0/0
Sample 12	Femur	Konfin 1	27	16/0/0
Sample 13	Femur	Konfin 1	51	16/0/0
Sample 14	Tibia	Konfin 1	84	16/0/0
Sample 15	Tibia	Konfin 1	13	16/0/0
Sample 16	Tibia	Konfin 1	38	16/0/0
Sample 17	Tibia	Konfin 1	21	16/0/0
Sample 18	Tibia	Konfin 1	39	15/1/0
Sample 19	Tibia	Konfin 1	13	16/0/0
Sample 20	Tibia	Konfin 1	10	13/3/0
Sample 21	Femur	Storžič	10	15/1/0
Sample 22	Femur	Storžič	13	16/0/0
Sample 23	tooth- RM2	Storžič	313	16/0/0
Sample 24	tooth- LM3	Konfin 2	26	16/0/0
Sample 25	tooth- RM2	Konfin 2	136	16/0/0
Sample 26	tooth- RM3	Konfin 2	132	16/0/0
Sample 27	tooth- LM1	Konfin 2	14	16/0/0
Sample 28	tooth- RM3	Konfin 2	31	16/0/0
Sample 29	tooth- LM1	Konfin 2	9	16/0/0
Sample 30	tooth- RM2	Konfin 2	30	16/0/0
Sample 31	tooth- LM3	Konfin 2	77	16/0/0
Sample 32	tooth- LM2	Konfin 2	241	16/0/0
Sample 33	Femur	Bodovlje	38	16/0/0
Sample 34	Femur	Bodovlje	20	16/0/0
Sample 35	Femur	Bodovlje	14	16/0/0
Sample 36	Tibia	Konfin 1	16	16/0/0
Sample 37	Femur	Konfin 1	30	16/0/0
Sample 38	Femur	Mozelj	17	16/0/0
Sample 39	Femur	Mozelj	18	16/0/0
Sample 40	Femur	Konfin 2	30	16/0/0
Sample 41	Femur	Konfin 2	10	16/0/0
Sample 42	Femur	Konfin 2	10	14/1/1 (drop-in 2x)
Sample 43	Femur	Konfin 2	4	11/4/1
Sample 44	Femur	Konfin 2	4	16/0/0 (drop-in 2x)
Sample 45	Femur	Konfin 2	70	16/0/0
Sample 46	Femur	Konfin 2	15	16/0/0
Sample 47	Femur	Nova Gorica	10	16/0/0
Sample 48	Femur	Konfin 1	50	16/0/0
Sample 49	tooth- RM3	Konfin 2	20	16/0/0

Table 1: Characteristics, DNA quantity, and efficiency of autosomal STR typing of 49 DNA samples extracted from bones or teeth of WWII victims excavated in Slovenia.
*From Human Quantifiler DNA Quantification kit, # 16 autosomal STRs using the Investigator ESSplex Plus (Qiagen) or NGM (Applied Biosystems) kit: Number of loci with complete results / number of loci with partial results / number of loci with complete locus drop-outs.

HIrisPlex genotyping and eye / hair colour prediction

All the 49 bones and teeth samples were genotyped using the HIrisPlex assay including 24 DNA markers informative for eye and hair colour [24,25]. All the samples were HIrisPlex-genotyped multiple times in a multiplex, as well as singleplex for some markers for some samples (see methods) to confirm the obtained multiplex genotyping and consequent eye/hair colour prediction results. Table 2 shows the genotypes obtained for all the 24 HIrisPlex DNA markers for all the 49 samples tested, and their HIrisPlex-predicted eye and hair colours are reported in Table 3. We obtained complete HIrisPlex profiles for all 49 samples except for the *MC1R* DNA marker N29insA, which dropped out in 41 (83.7%) samples. All the DNA markers in the HIrisPlex assay have a certain degree of influence on the final eye and hair colour prediction as described previously [24,25]. For instance, each of the 11 MC1R DNA markers (SNP 1 to 11) impacts on red hair colour. The *MC1R* INDEL polymorphism N29insA was the most sensitive, of all the 24 DNA markers in the HIrisPlex assay, to suboptimal quality or quantity of template DNA [23,24]. Given the quality and quantity of the samples, in none of the 49 samples used it was possible to determine the N29insA genotype via the HIrisPlex multiplex assay, despite multiple attempts (data not shown). However, with a singleplex analysis, it was possible to reveal the genotype for N29insA for 8 samples i.e., sample 2, 18, 22, 23, 25, 26, 29 and 32. Except for sample 26, N29insA did not give any result. However, the markers 8 (rs1805009) in sample 26, 9 (Y152OCH) in sample 37 and the markers 2 (rs11547464) and 15 (rs4959270) in samples 42 and 49 contributed to the red hair colour. Considering all HIrisPlex DNA markers in the prediction modelling, samples 26, 37, 42 and 49 were predicted to have red hair colour.

Sample ID	Genotype*																							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
Sample 1	-	G	C	CT	G	C	C	G	C	G	T	A	G	GA	CA	C	GT	GA	A	T	CT	T	CT	T
Sample 2	C	G	C	C	T	C	C	G	C	G	T	A	G	GA	CA	C	G	GA	GA	C	T	G	C	T
Sample 3	-	G	C	C	G	C	C	G	C	GA	T	A	G	A	CA	C	GT	G	A	C	CT	G	CT	T
Sample 4	-	G	C	CT	G	C	C	G	C	G	T	A	G	A	CA	C	G	G	GA	C	CT	GT	CT	T
Sample 5	-	G	CT	C	G	C	C	G	C	G	T	A	G	A	CA	C	GT	G	A	T	T	GT	T	T
Sample 6	-	G	C	C	G	C	CT	G	C	G	T	A	G	A	CA	C	T	G	A	CT	CT	G	C	GT
Sample 7	-	G	C	CT	G	C	C	G	C	G	T	A	G	A	C	C	GT	G	GA	CT	CT	GT	C	GT
Sample 8	-	G	C	C	G	C	C	G	C	GA	T	CA	GC	GA	C	C	GT	G	A	C	CT	GT	CT	T
Sample 9	-	G	C	C	G	C	C	G	C	G	T	A	G	A	C	C	G	GA	A	CT	T	GT	CT	T
Sample 10	-	G	C	C	G	C	C	G	C	A	T	A	G	GA	CA	C	G	GA	GA	CT	CT	G	CT	T
Sample 11	-	G	C	C	G	C	C	G	C	G	T	A	G	A	A	C	GT	G	GA	CT	C	G	CT	T
Sample 12	-	G	C	C	G	C	C	G	C	G	T	A	G	GA	C	C	T	GA	A	CT	T	T	C	GT
Sample 13	-	G	C	CT	G	C	C	G	C	G	T	A	G	A	A	C	G	G	A	C	CT	G	CT	GT
Sample 14	-	G	C	C	G	C	C	G	C	GA	T	A	G	A	A	C	G	G	A	CT	T	G	CT	GT
Sample 15	-	G	C	C	G	C	C	G	C	GA	T	A	G	A	C	C	GT	G	A	C	CT	GT	CT	GT
Sample 16	-	G	C	C	G	C	C	G	C	G	T	A	GC	GA	C	C	GT	G	A	C	T	GT	C	T
Sample 17	-	G	C	C	G	C	C	G	C	G	T	A	G	A	A	C	GT	G	GA	CT	CT	GT	C	T
Sample 18	C	G	C	C	G	C	C	G	C	GA	T	A	G	A	C	C	GT	G	A	C	CT	GT	CT	GT
Sample 19	-	G	C	C	G	C	C	G	C	A	T	A	G	A	CA	C	T	GA	GA	C	CT	G	C	T
Sample 20	-	G	C	C	G	C	C	G	C	GA	T	A	G	A	A	C	G	G	A	CT	T	G	CT	GT
Sample 21	-	G	C	C	G	C	C	G	C	G	T	A	G	GA	CA	C	GT	G	A	T	T	T	C	G
Sample 22	C	G	C	C	G	C	C	G	C	GA	T	A	GC	A	A	C	G	G	GA	CT	T	GT	C	G
Sample 23	C	G	C	C	G	C	C	G	C	G	T	A	G	A	CA	C	G	G	A	CT	T	GT	CT	T
Sample 24	-	G	C	C	G	C	C	G	C	G	T	CA	GC	A	CA	C	G	A	A	T	T	G	C	T
Sample 25	C	G	CT	C	G	C	C	G	C	G	T	A	G	A	A	T	G	G	A	CT	T	GT	CT	T
Sample 26	C	G	C	C	G	C	C	GC	C	G	T	A	G	A	CA	C	G	G	A	C	CT	T	T	GT
Sample 27	-	G	C	C	G	C	C	G	C	G	T	A	G	A	CA	CT	GT	G	A	CT	CT	GT	C	T
Sample 28	-	G	C	C	GT	C	C	G	C	G	T	A	G	A	CA	C	GT	G	A	CT	T	T	C	GT
Sample 29	C	G	C	C	G	C	C	G	C	G	T	A	G	A	A	C	GT	G	A	C	CT	GT	CT	T
Sample 30	-	G	C	C	G	C	C	G	C	G	T	A	G	A	CA	C	T	G	A	T	T	G	C	GT
Sample 31	-	G	C	C	GT	C	C	G	C	G	T	A	G	A	CA	C	G	G	A	CT	T	T	CT	G

Sample 32	C	G	C	C	GT	C	C	G	C	G	T	A	G	A	CA	C	G	G	A	CT	CT	GT	C	T
Sample 33	-	G	C	CT	G	C	C	G	C	GA	T	A	G	A	C	C	G	G	GA	CT	CT	G	C	GT
Sample 34	-	G	C	C	G	C	C	G	C	GA	T	A	GC	GA	CA	C	G	G	GA	T	T	GT	C	T
Sample 35	-	G	CT	C	G	C	C	G	C	G	T	A	G	A	A	C	GT	G	G	CT	T	G	CT	T
Sample 36	-	G	C	C	G	C	C	G	C	G	T	A	G	A	C	C	G	A	A	T	T	GT	C	G
Sample 37	-	G	C	C	G	C	C	G	CA	G	T	A	G	A	CA	C	T	G	GA	C	T	GT	C	GT
Sample 38	-	G	C	C	G	C	C	G	C	G	T	A	G	A	A	C	G	G	A	CT	T	GT	C	GT
Sample 39	-	G	C	C	G	C	C	G	C	G	T	A	G	A	C	CT	GT	G	GA	CT	T	T	CT	T
Sample 40	-	G	C	CT	G	C	C	G	C	G	T	A	G	A	CA	CT	G	G	GA	CT	T	G	C	GT
Sample 41	-	G	C	C	G	C	CT	G	C	G	T	A	G	A	C	C	G	G	A	C	CT	GT	CT	G
Sample 42	-	GA	C	C	G	C	CT	G	C	G	T	A	G	GA	CA	C	G	G	A	C	T	G	T	GT
Sample 43	-	G	C	C	G	C	C	G	C	G	T	A	G	A	CA	C	G	G	A	C	CT	T	T	G
Sample 44	-	G	C	C	G	C	C	G	C	G	T	A	G	A	A	C	T	G	GA	CT	CT	GT	C	GT
Sample 45	-	G	C	C	G	C	C	G	C	G	T	A	G	A	A	C	GT	G	A	C	CT	GT	CT	T
Sample 46	-	G	C	C	G	C	CT	G	C	G	T	A	G	A	C	C	GT	G	A	C	T	T	C	T
Sample 47	-	G	CT	C	G	C	C	G	C	G	T	A	G	A	CA	C	GT	G	A	C	T	T	C	G
Sample 48	-	G	C	CT	G	C	C	G	C	G	T	A	G	GA	CA	C	GT	G	GA	CT	T	GT	CT	T
Sample 49	-	GA	C	C	G	C	CT	G	C	G	T	A	G	GA	CA	C	G	G	A	C	T	G	T	GT

Table 2: HIrisPlex genotyping results for the 49 DNA samples extracted from WWII skeletal remains.

* The numbers 1 – 24 represents the 24 DNA variants included in the HIrisPlex assay with ordering according to their position in the electropherogram from lowest fragment size (1) to largest fragment size (24): 1- N29insA, 2- rs11547464, 3- rs885479, 4- rs1805008, 5- rs1805005, 6- rs1805006, 7- rs1805007, 8- rs1805009, 9- Y152OCH, 10- rs2228479, 11- rs1110400, 12- rs28777, 13- rs16891982, 14- rs12821256, 15- rs4959270, 16- rs12203592, 17- rs1042602, 18- rs1800407, 19- rs2402130,20- rs12913832, 21- rs2378249, 22- rs12896399, 23- rs1393350 and 24- rs683.

Sample ID	Probability Values Hair				Shade		Most likely predicted hair colour category	Probability Values Eye			Most likely predicted eye colour category
	Blond	Brown	Red	Black	Light	Dark		Blue	Intermediate	Brown	
Sample 1	0.409	0.481	0.056	0.053	0.842	0.158	DarkBrown	0.007	0.072	0.921	Brown
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 2	0.754	0.225	0.015	0.006	0.994	0.006	Blond	0.924	0.059	0.017	Blue
AUC Loss	0	0	0	0	0	0		0	0	0	
Sample 3	0.734	0.221	0.013	0.032	0.952	0.048	Blond	0.891	0.068	0.041	Blue
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 4	0.684	0.160	0.142	0.014	0.975	0.025	DarkBlond	0.936	0.043	0.021	Blue
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 5	0.129	0.431	0.001	0.440	0.270	0.730	Dark brown/Black	0.001	0.028	0.971	Brown
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 6	0.354	0.440	0.189	0.017	0.955	0.045	Brown	0.052	0.116	0.832	Brown
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 7	0.373	0.479	0.074	0.073	0.766	0.234	DarkBrown-Black	0.101	0.136	0.763	Brown
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 8	0.508	0.413	0.006	0.074	0.832	0.168	Brown	0.853	0.067	0.081	Blue
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 9	0.521	0.367	0.001	0.111	0.818	0.182	Brown	0.332	0.245	0.424	Brown
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 10	0.615	0.300	0.010	0.075	0.880	0.120	Brown	0.202	0.248	0.549	Brown
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 11	0.351	0.551	0.015	0.083	0.716	0.284	DarkBrown	0.075	0.136	0.788	Brown
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 12	0.532	0.434	0.002	0.032	0.915	0.085	DarkBlond/Light Brown	0.406	0.222	0.373	Blue
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 13	0.696	0.130	0.162	0.012	0.982	0.018	DarkBlond/Light Brown	0.891	0.068	0.041	Blue
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 14[♦]	0.464	0.341	0.004	0.191	0.758	0.242	DarkBlond/Light Brown	0.075	0.136	0.788	Brown
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 15[♣]	0.659	0.284	0.01	0.047	0.917	0.083	DarkBlond/Light Brown	0.936	0.043	0.021	Blue
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	

Sample 16	0.613	0.342	0.001	0.044	0.938	0.062	DarkBlond/Light Brown	0.806	0.077	0.117	Blue
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 17	0.391	0.508	0.008	0.093	0.753	0.247	DarkBrown	0.101	0.136	0.763	Brown
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 18➶	0.659	0.284	0.01	0.047	0.917	0.083	DarkBlond/Light Brown	0.936	0.043	0.021	Blue
AUC Loss	0	0	0	0	0	0		0	0	0	
Sample 19	0.736	0.208	0.031	0.025	0.960	0.040	Blond	0.924	0.059	0.017	Blue
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 20♦	0.464	0.341	0.004	0.191	0.758	0.242	DarkBlond/Light Brown	0.075	0.136	0.788	Brown
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 21	0.188	0.628	0.001	0.183	0.443	0.557	DarkBrown-Black	0.001	0.023	0.975	Brown
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 22	0.097	0.361	0.000	0.541	0.202	0.798	Black	0.028	0.064	0.909	Brown
AUC Loss	0	0	0	0	0	0		0	0	0	
Sample 23	0.469	0.391	0.002	0.138	0.778	0.222	Brown/Dark Brown	0.143	0.155	0.702	Brown
AUC Loss	0	0	0	0	0	0		0	0	0	
Sample 24	0.087	0.411	0.000	0.502	0.155	0.845	Black	0.001	0.038	0.961	Brown
AUC Loss	0.003	0	0	0	0	0		0	0	0	
Sample 25	0.026	0.472	0.001	0.502	0.079	0.921	Black	0.293	0.316	0.391	Brown
AUC Loss	0	0	0	0	0	0		0	0	0	
Sample 26◬	0.302	0.235	0.45	0.013	0.959	0.041	Red	0.972	0.022	0.007	Blue
AUC Loss	0	0	0	0	0	0		0	0	0	
Sample 27	0.119	0.698	0.003	0.181	0.338	0.662	DarkBrown	0.160	0.213	0.627	Brown
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 28	0.349	0.551	0.004	0.095	0.797	0.203	DarkBrown/Brown	0.187	0.151	0.662	Brown
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 29●	0.742	0.224	0.015	0.02	0.969	0.031	Blond	0.936	0.043	0.021	Blue
AUC Loss	0	0	0	0	0	0		0	0	0	
Sample 30	0.136	0.618	0.001	0.246	0.303	0.697	DarkBrown-Black	0.000	0.014	0.985	Brown
AUC Loss	0	0	0	0	0	0		0	0	0	

Sample 31	0.322	0.531	0.003	0.144	0.729	0.271	DarkBrown	0.252	0.165	0.583	Brown	
AUC Loss	0.003	0	0.008	0	0	0		0	0	0		
Sample 32	0.381	0.515	0.007	0.097	0.797	0.203	DarkBrown	0.101	0.136	0.763	Brown	
AUC Loss	0	0	0	0	0	0		0	0	0		
Sample 33	0.400	0.378	0.093	0.128	0.725	0.275	Brown	0.052	0.116	0.832	Brown	
AUC Loss	0.003	0	0.008	0	0	0		0	0	0		
Sample 34	0.083	0.416	0.000	0.501	0.183	0.817	Black	0.000	0.007	0.993	Brown	
AUC Loss	0.003	0	0.008	0	0	0		0	0	0		
Sample 35	0.328	0.473	0.003	0.195	0.648	0.352	DarkBrown/Brown	0.075	0.136	0.788	Brown	
AUC Loss	0.003	0	0.008	0	0	0		0	0	0		
Sample 36	0.199	0.497	0.000	0.304	0.378	0.622	DarkBrown-Black	0.008	0.113	0.880	Brown	
AUC Loss	0.003	0	0.008	0	0	0		0	0	0		
Sample 37	0.00	0.00	1.00	0.00	0.942	0.058	Red	0.917	0.052	0.031	Blue	
AUC Loss	0.003	0	0.008	0	0	0		0	0	0		
Sample 38	0.458	0.396	0.002	0.143	0.781	0.219	Brown	0.101	0.136	0.763	Brown	
AUC Loss	0.003	0	0.008	0	0	0		0	0	0		
Sample 39	0.093	0.686	0.001	0.221	0.245	0.755	DarkBrown-Black	0.350	0.228	0.422	Brown	
AUC Loss	0.003	0	0.008	0	0	0		0	0	0		
Sample 40	0.156	0.610	0.036	0.198	0.443	0.557	DarkBrown	0.087	0.192	0.721	Brown	
AUC Loss	0.003	0	0.008	0	0	0		0	0	0		
Sample 41	0.618	0.243	0.128	0.010	0.986	0.014	Dark Blond/Light Brown	0.936	0.043	0.021	Blue	
AUC Loss	0.003	0	0.008	0	0	0		0	0	0		
Sample 42[♦]	0.379	0.028	0.593	0	1	0	Red	0.915	0.057	0.028	Blue	
AUC Loss	0.003	0	0.008	0	0	0		0	0	0		
Sample 43[△¬]	0.652	0.294	0.008	0.047	0.922	0.078	DarkBlond	0.972	0.022	0.007	Blue	
AUC Loss	0.003	0	0.008	0	0	0		0	0	0		
Sample 44	0.321	0.589	0.009	0.081	0.715	0.285	DarkBrown	0.101	0.136	0.763	Brown	
AUC Loss	0.003	0	0.008	0	0	0		0	0	0		
Sample 45[●]	0.742	0.224	0.015	0.02	0.969	0.031	Blond	0.936	0.043	0.021	Blue	
AUC Loss	0.003	0	0.008	0	0	0		0	0	0		

Sample 46	0.718	0.184	0.092	0.006	0.994	0.006	Blond	0.953	0.032	0.015	Blue
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 47	0.640	0.284	0.003	0.074	0.916	0.084	DarkBlond/Light Brown	0.953	0.032	0.015	Blue
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 48	0.600	0.315	0.061	0.024	0.950	0.050	DarkBlond/Light Brown	0.143	0.155	0.702	Brown
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	
Sample 49◆	0.379	0.028	0.593	0	1	0	Red	0.915	0.057	0.028	Blue
AUC Loss	0.003	0	0.008	0	0	0		0	0	0	

Table 3: HIrisPlex-based eye and hair colour prediction results for the 49 DNA samples extracted from WWII skeletal remains.

*Probability estimation and final prediction of eye colour was done according to Walsh et al. 2011, 2014. Probability estimation and final prediction for hair colour and hair shade was done according to Walsh et al. 2014. and ♦, ●, ↗, ❖ and △ indicates pairs of DNA samples that were extracted from bones or teeth most likely belonging to the same individual, respectively, as indicated by autosomal STR data combined with previously obtained Y-chromosomal and mtDNA data (Zupanič Pajnič et al., 2010). ⇥ indicates the sample pair 26 and 43 for which the same autosomal STR profile was obtained, as well as same DNA-predicted eye colour, while predicted hair colours were different, see text for explanation. Eye and hair colour phenotypes of samples 1 and 4 (brothers) were confirmed by witness statement of the identified living sister.

All the DNA markers in the HIrisPlex assay have a certain degree of influence on the final eye and hair colour prediction as described previously [24, 25]. For instance, each of the 11 *MC1R* DNA markers (SNP 1 to 11) impacts on red hair colour. The *MC1R* INDEL polymorphism N29insA was the most sensitive, of all the 24 DNA markers in the HIrisPlex assay, to suboptimal quality or quantity of template DNA [23, 24]. Given the quality and quantity of the samples, in none of the 49 samples used it was possible to determine the N29insA genotype via the HIrisPlex multiplex assay, despite multiple attempts (data not shown). However, with a singleplex analysis, it was possible to reveal the genotype for N29insA for 8 samples i.e., sample 2, 18, 22, 23, 25, 26, 29 and 32. Except for sample 26, N29insA did not give any result. However, the markers 8 (rs1805009) in sample 26, 9 (Y152OCH) in sample 37 and the markers 2 (rs11547464) and 15 (rs4959270) in samples 42 and 49 contributed to the red hair colour. Considering all HIrisPlex DNA markers in the prediction modelling, samples 26, 37, 42 and 49 were predicted to have red hair colour.

Allelic drop-outs and drop-ins at HIrisPlex markers other than N29insA were observed in samples 14, 15, 18, 20, 26, 29, 37, 42, 43, 44 and 45. Hence, to confirm the results obtained for these samples, in addition to multiple genotypings (4 times), singleplex reaction for the markers 1, 4, 7 15, 20 and 24 was done, as the allelic drop-outs and drop-ins noted at these markers would affect the final eye and hair colour predictions of these samples. For all these samples, a consensus approach was followed when reporting the final genotype and phenotype prediction results as shown in Table 2 and 3.

The results from STR typing, supported by Y-chromosome and mtDNA analysis previously performed in some of the bones [18], suggests that five pairs of DNA samples were extracted from bones or teeth that belong to the same individual, respectively: sample 14 and 20, sample 29 and 45, sample 15 and 18, sample 42 and 49, as well as sample 26 and 43. Notably, this information was not available to the people who performed HIrisPlex genotyping and eye/hair colour prediction analysis. The HIrisPlex genotypes, obtained eye and hair colour probabilities, and finally concluded eye and hair colour phenotypes were the same for samples 14 and 20 both being dark blond/light brown and brown eyes; for samples 15 and 18 both being dark blond/light brown hair and blue eyes; for samples 29 and 45 both being blond hair and blue eyes; and for samples 42 and 49 both being red hair and blue eyes (Table 2 and 3), as expected under the assumption that these sample pairs belong to the same individual, respectively. However, for samples 26 and 43, although the eye colour was blue in

both with the same eye colour probabilities and the same genotypes at all underlying 6 SNPs, their hair colours were concluded differently with different probabilities based on different HIrisPlex profiles. For sample 26 we concluded red hair from obtaining probabilities of 45% for red, 30.2% for blond, 23.5% for brown, and 1.3% for black. For sample 43 however, the conclusion was dark blond with probabilities of 65.2% for blond, 29.4% for brown, 4.7% for black, and only 0.8% for red (Table 2 and 3). Multiple genotypings and several singleplex reactions produced the same HIrisPlex results, which were partly different in both samples as noted. At rs1805009 (SNP 8), the genotype was GC for sample 26 and GG for sample 43, and at rs683 (SNP 24), the genotype for sample 26 was GT and GG for sample 43. Currently, there is no data evidence other than the obtained STR results that the samples 26 and 43 indeed originate from the same individual's skeleton. Skeletal remains from the 62 victims in the Konfin II mass grave were excavated as determined according to the anthropological investigation and identification of victims was not performed yet, and there are no results from Y-STR and mtDNA typing available as of yet. Notably, these skeletal remains could not be excavated in anatomical position since the bodies were thrown into a karst cave in 1945 and water run-off mixed the remains during time passed. However, as mentioned above, sample 43 was the most difficult one of all DNA samples tested and had the lowest DNA concentration measured with only 4 pg/µl (together with sample 44). This resulted in drop-outs at 5 of the 15 STRs in the Investigator ESSPlex Plus STR kit, although four of them were not seen with the NGM STR kit. Given that sample 43 had the lowest DNA concentration and consequently the least reliable STR results, it may either be that the obtained STR results are misleading in concluding that samples 43 and 26 originate from the same individual, but in fact are from different but closely related individuals, which could explain the different hair colour results obtained with HIrisPlex. Alternatively, the obtained HIrisPlex genotyping results of sample 43 are not reliable for the hair color predicting DNA markers, or both types of genetic data are unreliable. Based on the currently available data, no definite conclusion can be drawn on these two samples, which illustrates the problems that can occur when analysing DNA from skeletal remains stored under unfavourable conditions.

Among the analysed bones there were two femurs from Konfin I mass grave (samples1 and 4) that belonged to two brothers as identified via autosomal STR similarities with a living sister also analysed in the run of the investigation. For sample 1, HIrisPlex analysis suggests dark brown hair and brown eyes, and for sample 4 dark blond hair and

blue eyes were determined. The living sister of the analysed two brothers has confirmed the HIrisPlex-predicted eye and hair colours of both of her brothers.

Conclusions

Overall, our results provide increased data evidence for the suitability of the HIrisPlex system to successfully retrieve eye and hair colour information from DNA of human remains for the purpose of missing person/disaster victim identification. Based on the promising results from the present study together with those from a previous study [17], we envision that the HIrisPlex system is applied to many more disaster victim and missing person identification cases where ante-mortem samples and putative relatives are not directly available, so that the DNA-predicted eye and hair colour can assist locating them to make them available for STR profiling for final identification of the missing. We furthermore envision that in addition to eye and hair colour, other externally visible characteristics become predictable from DNA of human remains in the future, allowing to describe a missing person's appearance from his or her skeletal remains more and more detailed, improving to guide the search for ante-mortem samples or putative relatives.

Acknowledgements

Barbara Gornjak Pogorelc and Katja Vodopivec Mohorčič are greatly acknowledged for their technical assistance in sample preparation and STR typing. The authors like to thank the Governmental Commission on Concealed Mass Graves of the Republic of Slovenia for their support in excavations of Second World War victims. This study was financially supported by Erasmus MC University Medical Center Rotterdam.

References

[1] Biesecker LG, Bailey-Wilson JE, Ballantyne J, Baum H, Bieber FR, Brenner C, Budowle B, Butler JM, Carmody G, Conneally PM, Duceman B, Eisenberg A, Forman L, Kidd KK, Leclair B, Niezgoda S, Parsons TJ, Pugh E, Shaler R, Sherry ST, Sozer A, Walsh A (2005) Epidemiology. DNA identifications after the 9/11 World Trade Center attack. Science 310 (5751):1122-3.

[2] Prinz M, Carracedo A, Mayr WR, Morling N, Parsons TJ, Sajantila A, Scheithauer R, Schmitter H, Schneider PM; International Society for Forensic Genetics (2007) DNA Commission of the International Society for Forensic Genetics (ISFG): recommendations regarding the role of forensic genetics for disaster victim identification (DVI). Forensic Sci Int Genet 1(1):3-12.

[3] Hartman D, Drummer O, Eckhoff C, Scheffer JW, Stringer P (2011) The contribution of DNA to the disaster victim identification (DVI) effort. Forensic Sci Int 205(1-3):52-8.

[4] Montellius K and Lindblom B (2012) DNA analysis in disaster victim identification. Forensic Sci Med Pathol 8(2):140-7.

[5] Hofreiter M, Serre D, Poinar HN, Kuch M, Pääbo S. Ancient DNA. (2001) Nat Rev Genet 2:353–9.

[6] Burger J, Hummel S, Hermann B, Henke W (1999) DNA preservation: a microsatellite-DNA study on ancient skeletal remains. Electrophoresis 20:1722-1728.

[7] Higgins D and Austin JJ (2013) Teeth as a source of DNA for forensic identification of human remains:a review. Sci Justice 53(4):433-41.

[8] Pääbo S, Poinar H, Serre D, Jaenicke-Despres V, Hebler J, Rohland N, et al (2004) Genetic analyses from ancient DNA. Annu Rev Genet 38:645-679.

[9] Gilbert MTP, Bandelt H-J, Hofreiter M, Barnes I. (2005) Assessing ancient DNA studies. Trends Ecol Evol. 20:541–4.

[10] Romanini C, Catelli ML, Borosky A, Pereira R, Romero M, Salado Puerto M, et al (2012) Typing short amplicon binary polymorphisms: Supplementary SNP and Indel genetic information in the analysis of highly degraded skeletal remains. Forensic Sci Int Genet 6:469-476.

[11] Campos PF, Craig OE, Turner-Walker G, Peacock E, Willerslev E, Gilbert MT (2013) DNA in ancient bones – where is it located and how should we extract it. Ann Anat 194(1):7-16.

[12] Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R, Bertalan M, Nielsen K, Gilbert MT, Wang Y, Raghavan M, Campos PF, Kamp HM, Wilson AS, Gledhill A, Tridico S, Bunce M, Lorenzen ED, Binladen J, Guo X, Zhao J, Zhang X, Zhang H, Li Z, Chen M, Orlando L, Kristiansen K, Bak M, Tommerup N, Bendixen C, Pierre TL, Grønnow B, Meldgaard M, Andreasen C, Fedorova SA, Osipova LP, Higham TF, Ramsey CB, Hansen TV, Nielsen FC, Crawford MH, Brunak S, Sicheritz-Pontén T, Villems R, Nielsen R, Krogh A, Wang J, Willerslev E (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature. 463(7282):757-62.

[13] Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Leidinger P, Backes C, Khairat R, Forster M, Stade B, Franke A, Mayer J, Spangler J, McLaughlin S, Shah M, Lee C, Harkins TT, Sartori A, Moreno-Estrada A, Henn B, Sikora M, Semino O, Chiaroni J, Rootsi S, Myres NM, Cabrera VM, Underhill PA, Bustamante CD, Vigl EE, Samadelli M, Cipollini G, Haas J, Katus H, O'Connor BD, Carlson MR, Meder B, Blin N, Meese E, Pusch CM, Zink A (2003) New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. Nat Commun. 28;3:698.

[14] Knapp M, Lalueza-Fox C and Hofreiter M (2015) Re-inventing ancient human DNA. Investig Genet 6:4.

[15] Brenner CH, Weir BS (2003) Issues and strategies in the DNA identification of World Trade Center victims.Theor Popul Biol. 63:173–178.

[16] Ge J, Budowle B, Chakraborty R (2010) DNA identification by pedigree likelihood ratio accommodating population substructure and mutations. Investig Genet. 1(1):8.

[17] Draus-Barini J, Walsh S, Pospiech E, Kupiec T, Glab H, Branicki W, et al (2013) Bona fide colour: DNA prediction of human eye and hair colour from ancient and contemporary skeletal remains. Investig Genet 4:3.

[18] Zupanič-Pajnič I, Gornjak-Pogorelc B, Balažic J (2010) Molecular genetic identification of skeletal remains from the Second world war Konfin I mass grave in Slovenia. Int J Legal Med 124:307-317.

[19] Zupanič Pajnič I (2013) Genetic identification of Second World War victim's skeletal remains. Lap Lambert Academic Publishing, Saarbrücken.

[20] Ossowski A, Kuš M, Brzezinski P, Pruffer J, Piatek J, Zielinska G, et al (2013) Example of human individual identification from World War II gravesite. Forensic Sci Int 233:179-192.

[21] Morild I, Hamre SS, Huel R, Parsons TJ (2015) Identification of missing Norwegian World War II soldiers, in Karelia Russia. J Forensic Sci 60:1104-1110.

[22] Kayser M and de Knijff P (2011) Improving human forensics through advances in genetics, genomics and molecular biology. Nat Rev Genet 12(3):179-92.

[23] Kayser M (2015) Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. Forensic Sci Int Genet 18:33-48.

[24] Walsh S, Liu F, Wollstein A, Kovatsi L, Ralf A, Kosiniak-Kamysz A et al (2013) The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. Forensic Sci Int Genet 7:98-115.

[25] Walsh S, Chaitanya L, Clarisse L, Wirken L, Draus-Barini J, Kovatsi L et al (2014) Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage. Forensic Sci Int Genet 9:150-61.

[26] King TE, Gonzalez Fortes G, Balaresque P, Thomas MG, Balding D, Maisano Delser P, et al (2014) Identification of the remains of King Richard III. Nature Communications 5:5631.

[27] Ferenc M (2008) Topografija evidentiranih grobišč (Topography of documented mass graves). In: Dežman J (ed) Poročilo Komisije Vlade Republike Slovenije za reševanje vprašanj prikritih grobišč 2005-2008. Družina, Ljubljana, pp 7-27.

[28] Wilson MR, DiZinno JA, Polanskey D, Replogle J, Budowle B (1995) Validation of mitochondrial DNA sequencing for forensic casework analysis. Int J Legal Med 108:68-74.

[29] Amory S, Huel R, Bilić A, Loreille O, Parsons TJ (2012) Automatable full demineralization DNA extraction procedure from degraded skeletal remains. Forensic Sci Int Genet 6:398-406.

[30] Rohland N, Hofreiter M (2007) Ancient DNA extraction from bones and teeth. Nat Protoc 2:1756-1762.

[31] Caputo M, Irisarri M, Alechine E, Corach D (2013) A DNA extraction method of small quantities of bone for high-quality genotyping. Forensic Sci Int Genet 7:488-493.

[32] Alonso A, Andelinović Š, Martin P (2001) DNA typing from skeletal remains: evaluation of multiplex and megaplex STR systems on DNA isolated from bone and teeth samples. Croat Med J 42:260-266.

[33] Parson W, Gusmao L, Hares DR, Irwin JA, Mayr WR, Morling N, et al (2014) DNA Commission of the International Society for Forensic Genetics: Revised and extended guidelines for mitochondrial DNA typing. Forensic Sci Int Genet 13:134-142.

[34] Zupanič-Pajnič I, Debska M, Gornjak-Pogorelc B, Vodopivec Mohorčič K, Balažic J, Zupanc T et al (2016) Highly efficient automated extraction of DNA from old and contemporary skeletal remains. J Forensic Legal Med 37:78-86

[35] Zupanič Pajnič I (2016) Extraction of DNA from human skeletal material, in: Goodwin W (Ed), Forensic DNA Typing Protocols, Methods in Molecular Biology, vol 1420, pp 89-108.

[36] Qiagen Companies (2014) EZ1 DNA Investigator Handbook. Hilden.

[37] Qiagen Companies (2013) Investigator ESSplex Plus handbook. Hilden.

[38] Zupanič Pajnič I, Gornjak Pogorelc B, Balažic J, Zupanc T, Štefanič B (2012) Highly efficient nuclear DNA typing of the World War II skeletal remains using three new autosomal short tandem repeat amplification kits with the extended European Standard Set of loci. Croat Med J 53:17-23.

[39] Walsh S, Liu F, Ballantyne KN, van Oven M, Lao O, Kayser M (2011) IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic Sci Int Genet 5:170-80.

[40] Liu F, van Duijn K, Vingerling JR, Hofman A, Uitterlinden AG, Janssens AC, Kayser M (2009) Eye colour and the prediction of complex phenotypes from genotypes. Curr Biol 19(5):R192-3.

[41] Walsh S, Wollstein A, Liu F, Chakravarthy U, Rahu M, Seland JH, Soubrane G, Tomazzoli L, Topouzis F, Vingerling JR, Vioque J, Fletcher AE, Ballantyne KN, Kayser M. (2012) DNA-based eye colour prediction across Europe with the IrisPlex system. Forensic Sci Int Genet 6:330-40.

Chapter 4

The HIrisPlex-S system and its forensic validation: The first tool for simultaneous prediction of eye, hair and skin colour from DNA

Manuscript in preparation

Abstract

Forensic DNA Phenotyping (FDP) i.e. the prediction of human externally visible characteristics from DNA has become a fast growing topic in forensic research and in some countries with legal allowance of FDP it has already entered forensic practice. Previously, we developed and forensically validated the IrisPlex system for eye color prediction and the HIrisPlex system for combined eye and hair color prediction from DNA. With this study, we present the HIrisPlex-S system (S for skin) for simultaneous prediction of eye, hair, and skin color from DNA as well as its successful forensic developmental validation. The HIrisPlex-S system consists of two multiplex assays, the previously developed HIrisPlex assay targeting 24 SNPs and a novel multiplex assay targeting additional 17 SNPs. It further includes three statistical prediction models, the previous IrisPlex model for eye color prediction based on 6 SNPs, the previous HIrisPlex model for hair color prediction based on 22 SNPs, and the novel HIrisPlex-S model for skin color prediction based on 36 SNPs targeted by both multiplex genotyping assays. Besides describing and introducing the HIrisPlex-S DNA test system, we also demonstrate the forensic developmental validation of the HIrisPlex-S genotyping assay. We illustrate that the newly developed 17-plex assay performs in full agreement with the Scientific Working Group on DNA Analysis Methods (SWGDAM) guidelines as previously demonstrated for the 24-plex HIrisPlex assay, the second multiplex assay of the HIrisPlex-S system. Complete profiles were obtained from only 63 pg of DNA with both assays. Rigorous testing of simulated forensic casework samples such as blood, semen, saliva stains, and low quantity touch (trace) DNA samples proved the robustness and the efficiency of the assay. The results from artificially degraded DNA and from simulated casework samples of degraded DNA show that the two multiplex assays are suitable for analysing degraded samples. With this study, we present the first forensically validated test system for parallel eye, hair and skin colour from DNA available for forensic and other practical usage.

Introduction

When the standard DNA profiling based on short tandem repeat (STR) analysis fails to identify the donor of a human biological sample found at a crime-scene, i.e., because his/her DNA profile is not yet known to the investigators, alternative approaches are needed to aid the police investigations to find the unknown perpetrators. Forensic DNA phenotyping (FDP), which is the ability to predict externally visible characteristics (EVC) from forensic DNA samples, is a promising alternative to reduce the potential number of suspects and to provide a direction to police investigations for finding unknown perpetrators [1-3]. Besides its use on all types of forensic stains found at crime scenes, FDP is also useful for describing EVCs from skeletal remains of deceased individuals [4–6], such as in missing person cases or mass disaster victim identification (DVI), where it can assist the investigation to locate putative relatives or ante-mortem samples. The application of large-scale genome-wide association studies (GWAS) to EVCs have identified numerous SNPs that can serve as informative DNA predictors for EVCs for those appearance traits where the available genetic knowledge already explains enough of the phenotypic EVC information.

Human pigmentation traits, i.e., the colour variation of the eyes (iris), hair, and skin represent EVCs where genetic knowledge is currently most advanced, Consequently, DNA prediction of categorical eye and hair colour is now achievable with the forensically validated IrisPlex and HIrisPlex that use model-based probability prediction [7-11] and are extensively tested and applied [12-14], as well as alternative tools that largely overlap with Iris/HIrisPlex in the SNP predictors used [14-18]. Skin colour is genetically more complex than eye and hair colour given its worldwide variation maximized between continental groups, in contrast to eye and hair color variation being almost exclusively restricted to people of European (and nearby) ancestry, providing challenges for identifying skin colour genes. The differences in the type, amount, and distribution of melanin pigment present in the skin melanocytes and the consequent variation in skin type is determined by the UV sensitivity [19]. It is broadly assumed that skin colour evolved in humans through adaptation to sunlight due to the Out-of-Africa migration of modern humans [20,21]. On the global scale, skin colour varies broadly from very pale to very dark colour. Because of the necessary prerequisite of GWAS regarding genetically homogenous study populations, compared to eye and hair colour, the knowledge about the genes that determine the skin colour variation is less advanced due to the global

distribution of the skin colour variation [22]. GWAS on human skin colour separately done in Europeans [23] and South Asians [19] delivered various genes, but because of the limited skin color variation within continental groups as oppose to between continental groups, the genetic knowledge on skin colour is currently less complete. Therefore, only a few studies are available on the DNA prediction of skin colour, and the level of detail at which DNA-based skin color prediction was previously achieved as well as the prediction accuracies obtained are limited [22]. Recently, a comprehensive survey investigating the skin colour predictive value of 59 candidate SNPs was published [18]. In this study, a subset of 10 SNPs were selected for predicting skin colour and the naïve Bayesian classification on these markers reported prevalence adjusted prediction average accuracies expressed as area under the operating receivers characteristics curve (AUC) values of 0.999 for white, 0.996 for black and 0.803 for intermediate skin colour [18]. However, there is a need for DNA-based skin colour prediction to cover more subtle differences in skin colour such as observed within the European and the Asian populations, respectively [24]. Moreover, the predictive approach of Maroñas et al. [18] is based on a limited number of 285 individuals unable to provide robust prediction outcomes, and no genotyping assay is made available by these authors.

Aiming to overcome current limitation in predicting skin colour from DNA in forensic and anthropological applications, in this study, we introduce the HIrisPlex-S system for simultaneous prediction of eye, hair, and skin color from DNA. HIrisPlex-S consists of two multiplex assays, the previously developed HIrisPlex assay [10,11] and a novel second multiplex assay. It further consists of three statistical prediction models, the previous IrisPlex model for eye color prediction [7–9], the previous HIrisPlex model for hair color prediction [10,11], and the novel HIrisPlex-S model for skin color prediction recently developed by Walsh et al. (in preparation). Here we described the development of the second multiplex assay targeting 17 skin colour informative SNPs, which together with 19 SNPs targeted with the HIrisPlex assay allow the prediction of very pale, pale, intermediate, dark and dark/black skin color categories with AUCs of 0.75 for very pale, 0.73 for pale, 0.75 for intermediate, 0.84 for dark and 0.98 for dark/black using the HIrisPlex-S model currently based on 1204. worldwide individuals (Walsh et al. in preparation). In addition, we performed the forensic developmental validation of the HIrisPlex-S system following the strict guidelines of the Scientific Working Group on DNA Analysis Methods (SWGDAM) [25] as essential prerequisite for future forensic application.

Materials and methods

For the developmental validation, a selection of body fluid and tissue samples were collected from individuals with informed consent. Test samples included single and multiple source samples and simulated casework samples (saliva, blood, semen, vaginal swabs and touched items) were generated in-house. DNA was extracted from all the samples using the QIAamp DNA Mini kit (Qiagen, Hagen, Germany) according to the manufacturer's guidelines or an in-house protocol (unpublished). All the extracted DNA samples were quantified using the Quantifiler Human DNA Quantification kit (Applied Biosystems Inc., Foster City, USA) following the manufacturer's guidelines.

Multiplex design and protocol

The HIrisPlex-S genotyping assay includes 17 SNPs from 7 genes: rs3114908 (*MC1R*), rs1800414 (*OCA2*), rs10756819, rs2238289 (*HERC2*), rs17128291 (*SLC24A4*), rs6497292 (*HERC2*), rs1129038 (*HERC2*), rs1667394 (*HERC2*), rs1126809 (*TYR*), rs1470608 (*OCA2*), rs1426654 (*SLC24A5*), rs6119471 (*ASIP*), rs1545397 (*OCA2*), rs6059655 (*ASIP*), rs12441727 (*HERC2*), rs3212355 (*MC1R*) and rs8051733 (*MC1R*). Primer3Plus, free web-based design software was used to design the 17 primer pairs and their respective single-base extension (SBE) primers, using the default parameters of the software program [27]. For each SNP, both forward and reverse SBE primers were designed. The selection of the final primers was based on the suitability of the primers for the multiplex and the genotype of the resultant product permitting complete multiplexing. To ensure complete capillary separation between the PCR products, poly-T tails of differing lengths were added to the 5' end of the SBE primers. In order to cater for degraded samples, the PCR fragments were designed to be short as much possible and PCR fragment size was limited to less than 150 bp. All the primer sequences were analysed using AutoDimer software program [28] to avoid primer-primer interactions. All the marker details, primer sequences and primer concentrations can be found in Table 1. PCR amplification was performed in a single multiplex PCR assay in a total volume of 10 µl, containing PCR primers in specified concentrations, 1 µl genomic DNA extract (varying concentrations), 1X PCR buffer (Applied Biosystems), 2.5 mM MgCl2 (Applied Biosystems), 220 µM of each dNTP (Roche, Mannheim, Germany) and 1.75 U AmpliTaq Gold DNA polymerase (Applied Biosystems). A GeneAmp PCR System thermocycler (Applied Biosystems) was used for all amplifications, with the

following thermocyclic parameters: 95°C for 10 minutes, 33 cycles of 95°C for 30 seconds and 62°C for 30 seconds, 62°C for 5 minutes. The amplified PCR products were purified with ExoProStar – S (GE Healthcare Europe GmbH, Eindhoven, The Netherlands) and incubated at 37°C for 45 minutes and 80°C for 15 minutes. SBE multiplex reaction was carried out using 2 µl of the PCR product and 1 µl of the SNaPshot Ready Reaction Mix with the following thermocycler parameters: 96°C for 2 minutes; 25 cycles of 96°C for 10 seconds, 50°C for 5 seconds, 60°C for 30 seconds. The extended products were then purified using 1 µl of shrimp alkaline phosphatase (SAP) (USB Corporation) and incubated at 37°C for 45 minutes and 75°C for 15 minutes. Finally, the purified extended products were analysed on the ABI 3130xl Genetic Analyser (Applied Biosystems) with POP-7 polymer on a 36 cm capillary length, with an injection voltage of 2.5kV for 10 seconds and run time of 500 seconds at 60°C. Gene Mapper v3.7 software program (Applied Biosystems) was used for the allele calling and analysis of the results.

Sensitivity Study

The sensitivity of the HIrisPlex-S system was evaluated to determine the minimum concentration of DNA needed to achieve accurate and reliable results. DNA samples from 3 individuals were genotyped at 500 pg, 250 pg, 125 pg, 63 pg and 32 pg. DNA samples were prepared by serial dilution from 500 pg and all the dilutions were quantified in duplicates with the Quantifiler Human DNA Quantification kit (Applied Biosystems) following the manufacturer's guidelines. A threshold of 50 relative fluorescence units (RFU) was used to call the allele peaks.

Mixture Studies

DNA samples from two individuals of known genotype, at genomic DNA concentration of 500 pg was mixed in the ratios of 1:1, 1:5 and 1:10. DNA mixtures were used to assess the ability of the HIrisPlex-S system to predict the skin colour in case of multiple sample donors.

Stability studies

DNA from an individual with a concentration of 250 pg was exposed to ultraviolet light at intervals of 0 seconds, 30 seconds, 60 seconds, 5 minutes, 10 minutes, 15 minutes, 20 minutes and 30 minutes using a Bio-Link (Vilber Lourmat) at a strength of 50 J/cm2 and genotyped to assess the robustness of the assay when exposed to such conditions.

Assay Position	PCR Primers	Bases	Sequence	Concentration	SBE Primers	Bases	Sequence	Concentration
1	rs3114908_F	20	CAGAACACAGCCACACCCTA	0.4 µm	rs3114908_R	29	TTTTTTTTTTAGAGAAGGGTCAAGCACTT	0.12 µm
	rs3114908_R	20	CATAAAGGGGTCACCAGCAA	0.4 µm				
2	rs1800414_F	20	GCTGCAGGAGTCAGAAGGTT	0.4 µm	rs1800414_R	43	TTT TTT TTT TTC AGA ATC CCG TCA GAT ATC CTA	0.2 µm
	rs1800414_R	20	GGGACAAACGAATTGAGGAA	0.4 µm				
3	rs10756819_F	20	AAAGCAAGCTCATGTTTCCA	0.4 µm	rs10756819_F	35	TTTTTTTTTTTTTGGACCAGTTATTTGGGTTTGGA	1.7 µm
	rs10756819_R	23	CGTCATGACTAGAAAAACACCAA	0.4 µm				
4	rs2238289_F	21	GGAACATGAAGATTTCCCAGT	0.4 µm	rs2238289_F	53	TTT TTT TTT TTT TTT TTT TTG AGA TTG GAA GAT TGG AGC C	0.5 µm
	rs2238289_R	22	CTGATTCAGGTCTGCTGTCACT	0.4 µm				
5	rs17128291_F	20	CCAGCACTGCCAAAATAACA	0.4 µm	rs17128291_R	58	TTT TTT TTT TTT TTT TTT TTT CAA TGT GCA CTG GAT TAA AAG TC	1 µm
	rs17128291_R	20	CTCTTTGGACCCATCACCTC	0.4 µm				
6	rs6497292_F	21	TCTGCTGTAGAACCAATGTCC	0.4 µm	rs6497292_R	61	TTT TTT TTT TTT TTT TTT TTT TTT TTG TCT CCT GTG TCT TCA TCC T	0.2 µm
	rs6497292_R	21	GAATTGCACCTGTAGCTCCAT	0.4 µm				
7	rs1129038_F	20	ATGTCGACTCCTTTGCTTCG	0.4 µm	rs1129038_F	70	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT GAG CCA GGC AGC AGA GC	0.4 µm
	rs1129038_R	20	ACACCAGGCAGCCTACAGTC	0.4 µm				
8	rs1667394_F	24	CAGCTGTAGAGAGAGACTTTGAGG	0.4 µm	rs1667394_R	73	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT GCA GCA ATT CAA AAC GTG CAT A	0.2 µm
	rs1667394_R	20	CACCATTAAGACGCAGCAAT	0.4 µm				
9	rs1126809_F	27	TGTTTCTTAGTCTGAATAACCTTTTCC	0.4 µm	rs1126809_F	77	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TGT ATT TTT GAG CAG TGG CTC C	0.05 µm
	rs1126809_R	20	GGTGCATTGGCTTCTGGATA	0.4 µm				
10	rs1470608_F	26	TTTCTTGTGTTAACTGTCCTTACAAA	0.4 µm	rs1470608_F	62	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTCATTCTCTCTTAAAAATATTAATTTGCACC	4 µm
	rs1470608_R	23	GGAAAATATGTTAGGGTTGATGG	0.4 µm				
11	rs1426654_F	20	TTCAGCCCTTGGATTGTCTC	0.4 µm	rs1426654_F	86	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TGT CTC AGG ATG TTG CAG GC	0.16 µm
	rs1426654_R	26	TGAGTAAGCAAGAAGTATAAGGAGCA	0.4 µm				
12	rs6119471_F	20	GCAGGAGAATTGCTGGAACT	0.4 µm	rs6119471_R	91	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TGA AGG AAG AGT GAA AAT GCG TAA	1 µm
	rs6119471_R	21	AACCCGAAGGAAGAGTGAAAA	0.4 µm				
13	rs1545397_F	24	GGTATAGGATTATTTGGGGAATGA	0.4 µm	rs1545397_F	97	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT GTA CAA CTT TGT GAA TAT ACT AAA ATA C	1 µm
	rs1545397_R	28	TGGAGATATAGAATTCACACAACATAAA	0.4 µm				
14	rs6059655_F	20	GTGAGGAAATCGAGGCTCAG	0.4 µm	rs6059655_R	76	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGCTGATGCCCTGAGCA	2 µm
	rs6059655_R	20	AGGAGAAAGCTGCAGATCCA	0.4 µm				
15	rs12441727_F	20	GGGAAGAGACAGCTCCATGT	0.4 µm	rs12441727_F	106	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TGG CTC AGT GTG GCC TT	0.5 µm
	rs12441727_R	20	ACAATCCTGGGAGGTACACG	0.4 µm				
16	rs3212355_F	20	GAGTGAACCCAGGAAGATGC	0.4 µm	rs3212355_R	113	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TCC GAA GCC CAG CAG G	1.5 µm
	rs3212355_R	20	CATCAAAGGCAGACCTCTCG	0.4 µm				
17	rs8051733_F	19	AGGCGGTGGTCTCTCTCTC	0.4 µm	rs8051733_R	115	TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTC ACC CTG CCT GTC TCG	1.6 µm
	rs8051733_R	20	TTGCAACAGGAGGGTCTAGG	0.4 µm				

Table 1: Details of the 17 DNA variants of the HIrisPlex - S assay, including the PCR and SBE primer sequences and concentrations.

Simulated case samples

Simulated casework samples that include fresh blood, dried blood, semen, vaginal swab, saliva, mucus, inhibitor (haematin), mixtures and trace DNA were genotyped using the HIrisPlex-S assay under a blind proficiency test (Table 2). All the samples were quantified using the Quantifiler Human DNA Quantification kit (Applied Biosystems) as per the manufacturer's guidelines.

Results and discussion

Multiplex design and protocol

The HIrisPlex–S includes two assays (24-plex and 17-plex). In addition to the 24 SNPs in the HIrisPlex assay, these additional 17 SNPs (rs3114908 (*MC1R*), rs1800414 (*OCA2*), rs10756819 (*BNC2*), rs2238289 (*HERC2*), rs17128291 (*SLC24A4*), rs6497292 (*HERC2*), rs1129038 (*HERC2*), rs1667394 (*HERC2*), rs1126809 (*TYR*), rs1470608 (*OCA2*), rs1426654 (*SLC24A5*),rs6119471 (*ASIP*), rs1545397 (*OCA2*), rs6059655 (*ASIP*), rs12441727 (*HERC2*), rs3212355 (*MC1R*) and rs8051733 (*MC1R*) in the newly developed multiplex assay together with 19 SNPs from the HIrisPlex assay, i.e., 36 SNPs in total, predict the skin colour as belonging to one of the five categories: very pale, pale, intermediate, dark and dark-black (Walsh et al., in preparation). The SNPs N29insA (*MC1R*), rs1805005 (*MC1R*), rs1805009 (*MC1R*), Y152OCH (*MC1R*) and rs4959270 (*EXOC2*) from the HIrisPlex assay did not contribute to the skin color prediction as previous modelling had revealed and are thus not considered in the HIrisPlex-S model for skin color prediction (Walsh et al. in preparation). All the PCR and SBE primers in the HIrisPlex-S assay were designed considering the application of this assay to degraded forensic samples as well, where short amplicons are advantageous. PCR amplicon sizes of less than 150 bp were considered for all the 17 newly added SNPs as applied before to the 24 SNPs of the HIrisPlex assay [10]. To cater to the complete separation between the products, poly-T tails of differing lengths were added to the SBE primers. Both the PCR and SBE multiplexes were optimized to achieve balanced SNP alleles with similar peak intensities to ascertain that accurate results are obtained across all the SNPs in a wide range of samples with different DNA quantities. The SNPs rs1470608 (*OCA2*) (SNP 10) and rs6059655 (*ASIP*) (SNP 14) were the most difficult markers to optimise and hence had undergone several rounds of re-designing. The peak height of rs6059655

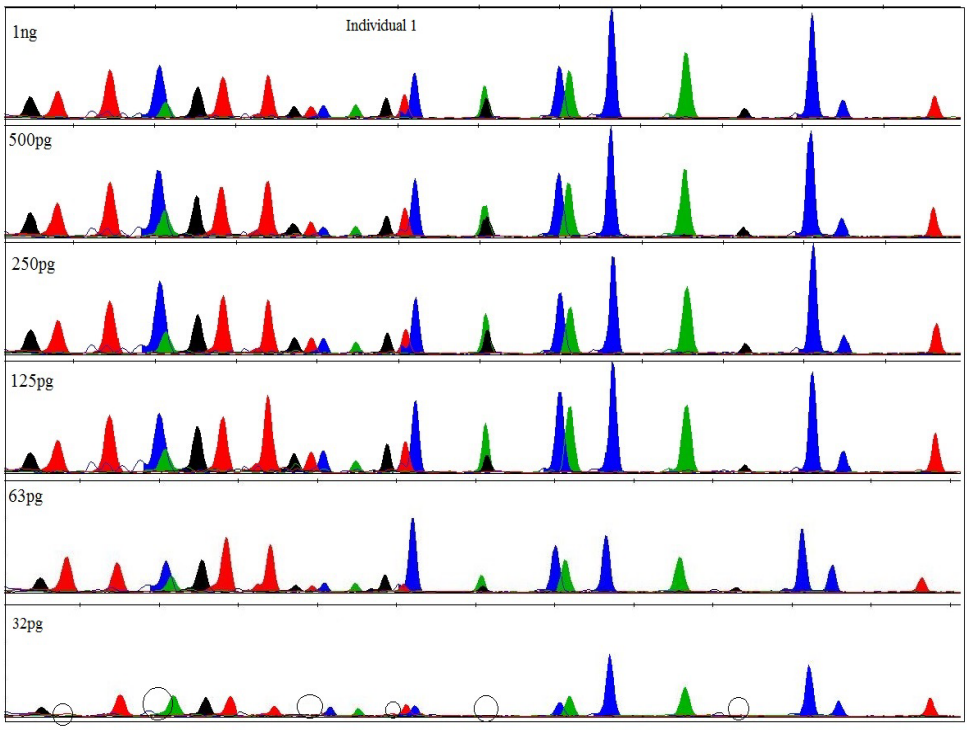
(*ASIP*) was lower as compared to the other 16 SNPs in this multiplex assay. Hence there are chances that rs6059655 may be one of the first to drop-out in case of low quantity or quality DNA. However, in case allelic drop-outs are observed in a sample, the SNPs shall be genotyped in a singleplex.

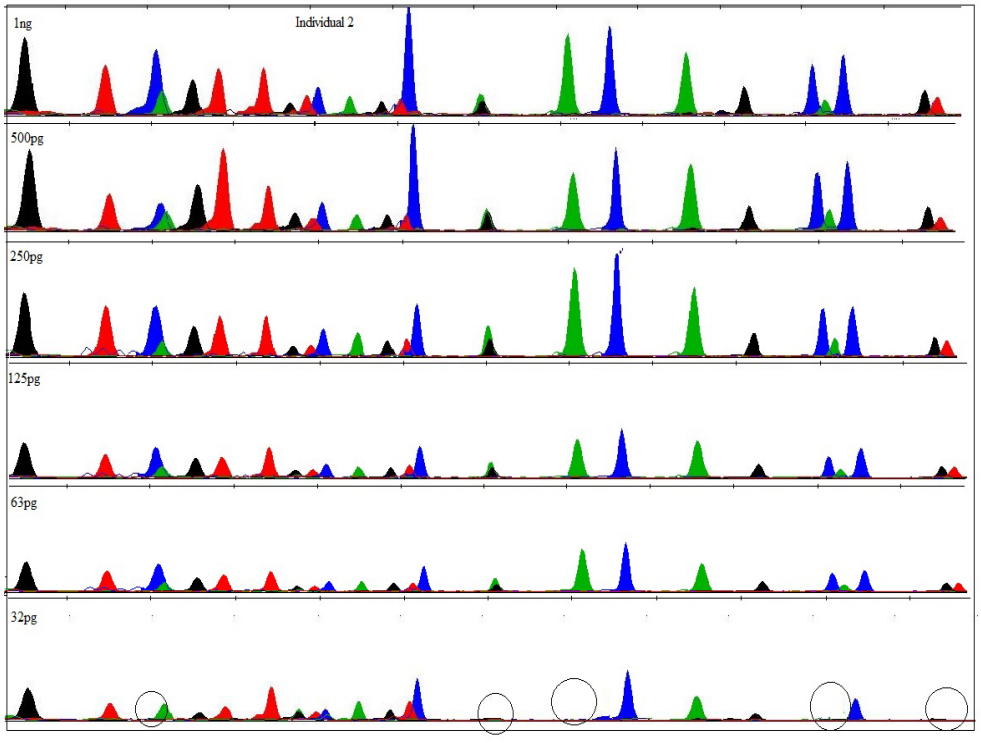
An Excel macro for statistical prediction of the eye, hair and skin color simultaneously was developed (Walsh et al., in preparation). Similar to the previously published IrisPlex and HIrisPlex assays [7-11], when the minor alleles are input into the HIrisPlex-S prediction tool (Table 3), they are used to predict the eye, hair and skin colour of the individual using the HIrisPlex-S prediction model (Walsh et al., in preparation). From the three eye color categories of blue, intermediate and brown; from the four hair color categories of blond, brown, red and black; and from the five skin color categories of very pale, pale, intermediate, dark and dark/black, the highest probability value is indicative of the predicted eye, hair and skin color (Table 3). However, the use of prediction guide is essential to predict the hair colour [10,11]. An online web-based tool is available at www.erasmusmc.nl/genetic_identification/resources/Irisplex_HIrisPlex/ for the prediction of eye and hair color from full profiles as well partial genotype profiles, which will be updated soon towards skin color.

As described elsewhere (Walsh et al. in preparation), the skin color prediction model was developed using samples from several populations with varying levels of skin color - 294 (24%) Ireland, 104 (9%) Greece, 684 (57%) Poland, individuals from HGDP set that include 20 (1.5%) Senegal, 20 (1.5%) Nigeria, 10 (1%) Kenya, 11 (1%) Papua New Guinea, 23 (2%) Japan, 38 (3%) China (Han). Notably, no skin color phenotypes are available directly for HGDP-CEPH samples but for this study they were assumed from the geographic origin of the samples together with their genetic evidence of non-admixture as established previously. Of these samples, a subset of 80% was used to train the model based on the Multinomial Logistic Regression (MLR), as previously published by Liu et al. [26]. The remaining 20% of the dataset was used to validate the model and to assess its prediction accuracy. The AUC values achieved with the HIrisPlex-S skin colour prediction model were 0.75 for very pale, 0.73 for pale, 0.75 for intermediate, 0.84 for dark and 0.98 for dark/black (Walsh et al., in preparation).

Sensitivity studies

Three individuals with different skin colors were serially diluted and genotyped with the HIrisPlex–S assay. The genomic DNA amounts ranged from 500 pg to 32 pg. A threshold of 50 RFU was used for allele calling in the capillary electrophoresis (CE) analysis using an ABI 3130xl instrument. As shown in Figure 1, it was possible to obtain full HIrisPlex–S profiles in samples at 63 pg of DNA. Allelic dropouts were observed at 32 pg in several amplifications in all three individuals as indicated by the black circles in Figure 1. In individual 1, dropouts were observed at SNPs rs3114908 (*MC1R*) (allele T from the heterozygous CT), rs10756819 (*BNC2*) (allele G from the heterozygous GA), rs6497292 (*HERC2*) (heterozygous alleles CT) rs1667394 (*HERC2*) (allele C of heterozygous CT), rs1470608 (*OCA2*) (heterozygous alleles CA) and rs6059655 (*ASIP*) (homozygous allele C). In individual 2, dropouts were observed at rs10756819 (*BNC2*) (Allele G of heterozygous GA), rs1470608 (*OCA2*) (heterozygous alleles CA), rs1426654 (*SLC24A5*) (homozygous allele A), rs6059655 (*ASIP*) (homozygous allele C), rs12441727 (*HERC2*) (heterozygous alleles GA) and rs8051733 (*MC1R*) (heterozygous alleles CT). In individual 3, dropouts were observed at rs1129038 (*HERC2*) (homozygous allele A), rs1126809 (*TYR*) (allele A of heterozygous GA) and rs6059655 (*ASIP*) (homozygous allele C). With the decreasing DNA concentration, heterozygote peak imbalance was observed. However, with such difficult samples, confirmation via singleplex analysis is always recommended. Notably, the sensitivity of the HIrisPlex–S was similar to the HIrisPlex assay [10,11], but lower than the IrisPlex assay that had only 6 SNPs [7,8]. The HIrisPlex-S assay demonstrated sensitivity higher than the SNPforID 52-SNP multiplex [29] and commercially available STR multiplexes such as AmpFlSTR NGM Select from Thermo Fisher Scientific [30].


Individual 1
1ng
500pg
250pg
125pg
63pg
32pg


Individual 2
1ng
500pg
250pg
125pg
63pg
32pg

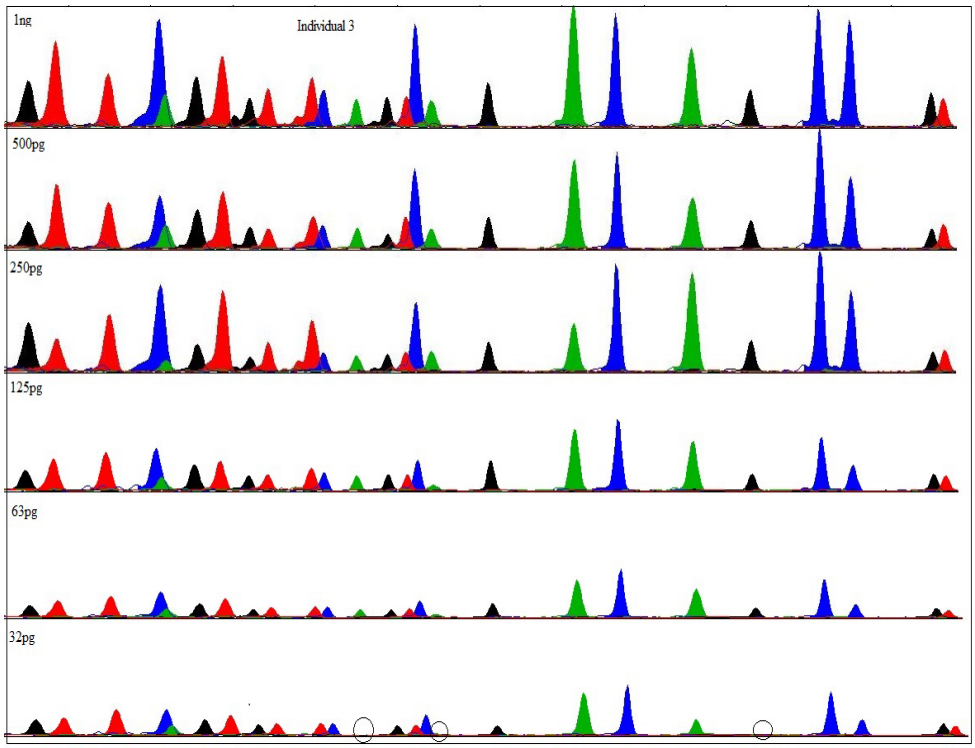


Figure1: Sensitivity testing of the HIrisPlex-S assay using three individuals with different skin colour phenotypes and different HIrisPlex-S genotypes using DNA samples ranging from 1ng to 32 pg DNA input. Individual 1 displayed phenotypic dark skin colour, individual 2 displayed phenotypic intermediate skin colour, and individual 3 displayed phenotypic pale skin colour. Alleles that have dropped out depending on DNA input amounts are indicated by black circles.

Mixture Studies

DNA from two individuals carrying different pigmentation phenotypes and genotypes, at a DNA input level of 500 pg, were mixed in the ratios of 1:1, 1:5 and 1:10. The results generated from the 17-Plex are illustrated in Figure 2 (for results of the HIrisPlex assay, see [11]). It can be observed that in a mixture ratio of 1:1 there appears a peak balance in the SNP assay for all 17 SNPs except at SNP 17 (rs8051733 (*MC1R*)). It is difficult to deduce if the profile at a mixture ratio of 1:1 is a true mixture or a sample with heterozygous alleles. However, at mixture ratios of 1:5 and 1:10 dropouts, peak imbalance was observed at SNPs 4, 6, 8 and 13. Peak imbalance could indicate a probable presence of mixtures, however, a dropout may not be considered true unless it was compared to a single sample profile. The samples used in this study belonged to different eye, hair and skin colour categories. Individual 1 reported a phenotype of brown eye, black hair and dark skin color. Individual 2 reported a phenotype of

blue eyes, blond hair and pale skin color category. As with eye and hair color and using the Iris/HIrisPlex system, there is a considerable chance that when individuals belonging to the same skin colour categories contribute to a DNA mixture, they may have the same alleles and such a sample would leave the mixture undetectable. This is the disadvantage, in general, for all the autosomal SNP assays [8,11] and not specifically for the Iris/HIris/HIrisPlex–S assays. It is to be noted that in routine forensic casework applications, STR is always the method of choice for mixture interpretation and we illustrated the performance of HIrisPlex–S assay in case of mixtures for additional information.

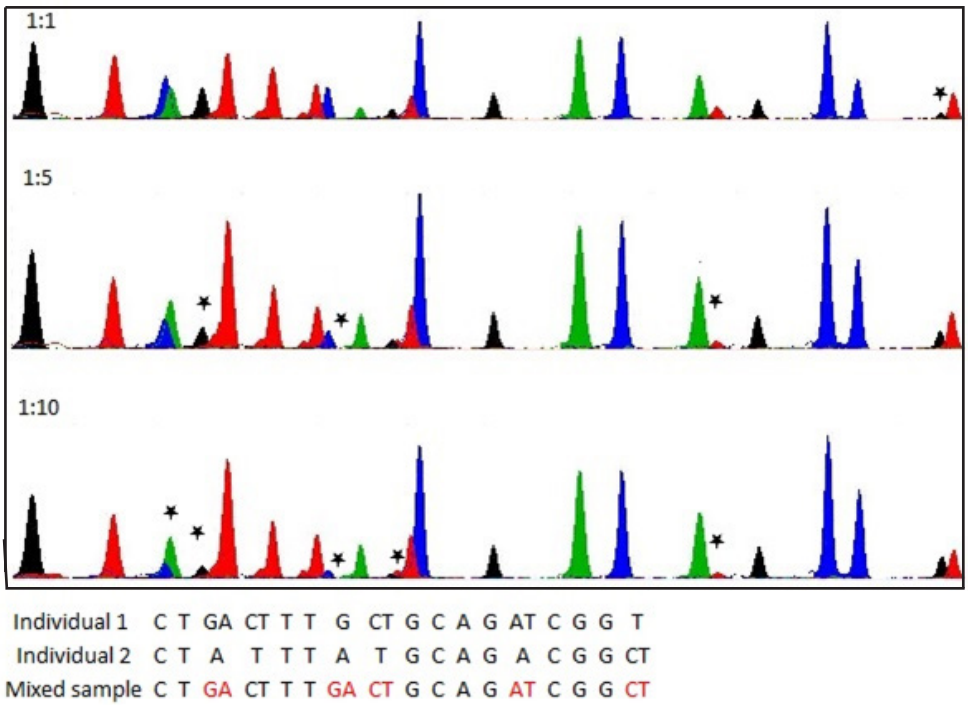


Figure 2: HIrisPlex-S mixture analysis of 1:1, 1:5 and 1:10 from DNA samples of two individuals with different skin colour phenotypes and different HIrisPlex-S genotypes. Individual 1 had phenotypic dark skin colour, individual 2 had phenotypic pale skin colour. The total amount of DNA used was approximately 500 pg. The asterisks represent the alleles present in the 1:1 mixture that may display unusual peak balance in the 1:5 and 1:10 mixtures indicating the presence of a mixture.

Degradation studies

In order to test the robustness of the HIrisPlex–S assay, a single source DNA sample at an input of 250 pg was subjected to artificial degradation by exposing to ultraviolet light at varying time intervals of 0 seconds, 30 seconds, 1 minute, 5 minutes, 10 minutes, 20 minutes, and 30 minutes. The samples were then genotyped with the HIrisPlex–S assay. Full profiles were obtained for the samples up to 10m of continuous exposure to ultraviolet light indicating the potential of HIrisPlex-S to deal with degraded DNA samples. After 20 minutes of UV light exposure however, dropout was observed at rs1470608 (*OCA2*) and rs10756819 (*BNC2*). It is always advantageous to amplify a sample several times when dealing with low quality DNA samples. However, the ability to obtain full profiles up to 10 minutes of UV light exposure indicates the robustness of the assay.

Simulated casework studies

To test the ability of the HIrisPlex–S assay to genotype and reliably predict the phenotype from different forensically relevant substrate, twenty-five simulated casework samples were prepared from blood, semen, saliva, trace DNA, mixtures and high and low concentrations of haematin inhibited samples. The samples were genotyped with the two multiplex assays. All the samples were blind-tested and the details are presented in Table 2. Single source samples were correctly genotyped and predicted, when there was sufficient DNA quantities. All mixtures were correctly identified; however, as not unexpected the prediction of the correct hair, eye and skin colour was not possible in the five mixture samples. Sample 21, a trace DNA sample from multiple sources, was wrongly interpreted as a single source and hence the phenotype interpretation was also incorrect. Very low peaks were observed in sample 21 making it difficult to make the correct interpretations. Also, it was not possible to obtain full profiles with high concentrations of haematin inhibitor. However, the HIrisPlex-S system was effective in producing results across different substrates and highly accurate for a number of casework type scenarios, particularly in single source samples from varying sources typically found at crime scenes. A concordance study of the HIrisPlex-S system involving 4 forensic laboratories with varying experiences is currently underway and will be added before finalizing this chapter for publication as scientific article.

Sample	Type	Simulated casework scenario	DNA quantity (pg)	Profile quality	Brown	Red	Black	Blond	Light	Dark	Blue	Int.	Brown	Very Pale	Pale	Int.	Dark	Dark-Black	Analyst interpretation	Correct/incorrect interpretation?
1	Dried blood	Direct on glass: single source	290	Full profile but low peaks	0.153	0.000	0.843	0.004	0.004	0.996	0.000	0.002	0.998	0.000	0.001	0.040	0.941	0.018	Black hair, brown eyes, dark skin colour	Correct
2	Saliva	Single source	3070	good profile	0.626	0.003	0.221	0.150	0.396	0.604	0.455	0.378	0.167	0.214	0.638	0.128	0.000	0.000	DarkBrown-Black hair, blue to intermediate eyes, pale skin colour	Correct
3	Frozen blood	Intentional heme inhibition (2%)	70	very low peaks but full profile	0.248	0.004	0.650	0.698	0.931	0.069	0.838	0.081	0.061	0.017	0.486	0.497	0.001	0.000	Blond hair, blue eyes, pale to intermediate skin colour	Correct
4	Saliva	single source	10530	good profile	0.193	0.003	0.032	0.772	0.962	0.038	0.917	0.052	0.031	0.012	0.505	0.462	0.001	0.000	Blond hair, blue eyes, pale to intermediate skin colour	Correct
5	Frozen blood	Single source	3070	good profile	0.193	0.003	0.032	0.772	0.962	0.038	0.917	0.052	0.031	0.012	0.505	0.482	0.001	0.000	Blond hair, blue eyes, pale to intermediate skin colour	Correct
6	Semen & blood	Mixture: semen & blood	8910	unbalanced peaks,possibly mixture															could not interpret the possibly mixture sample	-
7	Saliva	Direct on glass: single source	720	Full profile but low peaks	0.290	0.007	0.025	0.678	0.959	0.041	0.917	0.052	0.031	0.021	0.445	0.533	0.002	0.000	Blond hair, blue eyes, pale to intermediate skin colour	Correct
8	Touched item	Arm swab: single source	140	very low peaks	0.216	0.000	0.778	0.005	0.006	0.994	0.000	0.002	0.998	0.000	0.001	0.058	0.940	0.002	Black hair, brown eyes, dark skin colour	Correct
9	Vaginal secretions & semen	Mixture: vaginal swab & semen	14150	unbalanced peaks,possibly mixture															could not interpret the possibly mixture sample	-
10	Vaginal swab	single source	1840	good profile	0.404	0.003	0.193	0.400	0.652	0.348	0.143	0.155	0.702	0.054	0.621	0.325	0.000	0.000	Brown hair, brown eyes, pale to intermediate skin color	Correct
11	Saliva	Single source	21040	good profile	0.532	0.000	0.393	0.075	0.125	0.875	0.074	0.051	0.935	0.000	0.000	0.778	0.493	0.729	DarkBrown-Black hair, brown eyes, dark skin color	Correct
12	Saliva	Single source	6630	good profile	0.158	0.000	0.838	0.004	0.006	0.994	0.000	0.004	0.996	0.000	0.000	0.000	0.311	0.689	Black hair, brown eyes, dark-black skin color	Correct
13	Touched item	Trace DNA: single source	30	Full profile but very low peaks	0.097	0.000	0.900	0.002	0.003	0.997	0.000	0.002	0.998	0.000	0.000	0.748	0.252	0.000	Black hair, brown eyes, intermediate skin color	Correct
14	Touched item	Trace DNA: single source	110	good profile	0.464	0.001	0.168	0.367	0.648	0.352	0.052	0.116	0.832	0.084	0.360	0.633	0.001	0.001	DarkBrown-Black hair, brown eyes, intermediate skin color	Correct
15	Saliva & semen	Mixture: saliva & semen	11730	unbalanced peaks,possibly mixture	0.125	0.860	0.005	0.010	0.723	0.277	0.362	0.290	0.348	0.614	0.355	0.031	0.000	0.000	could not interpret the possibly mixture sample	-
16	Frozen blood	Single source	5980	good profile	0.153	0.000	0.843	0.004	0.004	0.996	0.000	0.002	0.998	0.000	0.001	0.040	0.941	0.018	Black hair, brown eyes, dark skin colour	Correct
17	Saliva & blood	Mixture: saliva & blood	5830	unbalanced peaks,possibly mixture															could not interpret the possibly mixture sample	-
18	Mucus & saliva	Mixture: mucus & saliva	10600	possible mixture															could not interpret the possibly mixture sample	Correct
19	Semen	Single source	4120	good profile	0.248	0.004	0.650	0.698	0.931	0.069	0.838	0.081	0.061	0.017	0.486	0.497	0.001	0.000	Blond hair, blue eyes, pale to intermediate skin colour	Correct
20	Touched item	Trace DNA: single source	90	Full profile but very low peaks	0.466	0.000	0.419	0.114	0.215	0.785	0.001	0.018	0.981	0.001	0.034	0.572	0.378	0.015	DarkBrown-Black hair, brown eyes, intermediate to dark skin color	Correct
21	Touched item	Trace DNA: multiple source	20	very low peaks but single source	0.477	0.003	0.130	0.391	0.746	0.254	0.052	0.116	0.832	0.007	0.498	0.500	0.004	0.000	DarkBrown-Black hair, brown eyes, pale to intermediate skin color	incorrect on single source and hence the phenotype interpretation
22	Vaginal secretions	Single source	11790	good profile	0.784	0.033	0.070	0.113	0.627	0.373	0.259	0.235	0.506	0.084	0.769	0.147	0.000	0.000	DarkBrown hair, brown eyes, pale skin color	Correct
23	Mucus	Single source	4160	good profile	0.489	0.001	0.417	0.094	0.185	0.815	0.028	0.064	0.909	0.007	0.158	0.764	0.036	0.035	DarkBrown-Black hair, brown eyes, intermediate skin color	Correct
24	Touched item	Trace DNA: single source	60	too many drop-outs, very low peaks in partial profile															could not interpret due to many drop-outs and hardly any profile	-
25	Frozen blood	Intentional heme inhibition (10%)	30	too many drop-outs, very low peaks in partial profile															could not interpret due to many drop-outs and hardly any profile	-

Table 2: Simulated casework samples prepared in house and tested where only the concentrations were known, the sample type and the simulated casework scenario were not provided.

HIrisPlex-S eye, hair & skin colour predictor

	SNP ID	Minor allele	No.
1	N29insA	A	0
2	rs11547464	A	0
3	rs885479	T	0
4	rs1805008	T	0
5	rs1805005	T	0
6	rs1805006	A	0
7	rs1805007	T	0
8	rs1805009	C	0
9	Y152OCH	A	0
10	rs2228479	A	0
11	rs1110400	C	0
12	rs28777	C	0
13	rs16891982	C	0
14	rs12821256	G	0
15	rs4959270	A	0
16	rs12203592	T	0
17	rs1042602	T	0
18	rs1800407	A	0
19	rs2402130	G	0
20	rs12913832	T	0
21	rs2378249	C	0
22	rs12896399	T	0
23	rs1393350	T	0
24	rs683	G	0
25	rs3114908	T	0
26	rs1800414	C	0
27	rs10756819	G	0
28	rs2238289	C	0
29	rs17128291	C	0
30	rs6497292	C	0
31	rs1129038	G	0
32	rs1667394	C	0
33	rs1126809	A	0
34	rs1470608	A	0
35	rs1426654	G	0
36	rs6119471	C	0
37	rs1545397	T	0
38	rs6059655	T	0
39	rs12441727	A	0
40	rs3212355	A	0
41	rs8051733	C	0

HAIR	Prob
Brown	0.218
Red	0.002
Black	0.038
Blond	0.743

SHADE	Prob
Light	0.949
Dark	0.051

EYE	Prob
Blue	0.858
Int.	0.081
Brown	0.061

* Please insert the number of minor alleles:
e.g. If major allele = A, minor allele = B
If minor allele is not observed (AA) it shall remain 0
If minor allele is heterozygote (AB) insert 1
If minor allele is homozygote (BB) insert 2

This sheet is protected for incident modifications
No password is needed to unprotect it

SKIN	Prob
Very Pale	0.009
Pale	0.392
Int.	0.598
Dark	0.001
Dark-Black	0.000

Table 3: Interactive macro version for HIrisPlex-S eye, hair and skin colour, based on the underlying genotype and phenotype databases for eye, hair and skin colour.

Conclusions

We hereby introduce the forensically validated HIrisPlex-S system that is capable of simultaneously predicting the eye, hair and skin color phenotypes from DNA using two multiplex assays and three statistical prediction models. A novel multiplex assay targeting 17 skin color informative SNPs was developed as an extension to the previously developed and validated HIrisPlex assay targeting 24 SNPs, of which 19 are additionally used for skin color prediction with the HIrisPlex-S model. The newly developed 17-plex assay is capable of coping with degraded DNA samples due to the PCR fragment sizes of less than 150 bp as was previously established also for the 24-plex HIrisPlex assay, the second multiplex of the HIrisPlex-S system. The developmental validation of the two multiplex assays of the HIrisPlex-S system fulfills SWGDAM guidelines in terms of sensitivity, stability, mixtures and simulated casework samples studies. Most notably, the HIrisPlex-S system is capable of generating full profiles with a minimum input DNA quantity of 63 pg. Accurate and reliable genotyping results were produced from simulated casework samples that included blood, semen, saliva, hair, and low quantity trace DNA samples. Mixture interpretation is difficult as with all FDP tools. The presented validation results demonstrate the robustness and reliability of the HIrisPlex-S system and may encourage the use of this tool for the prediction of eye, hair and skin color from DNA in criminal cases, missing person cases as well as outside the forensic arena such as those in anthropological or evolutionary applications to bringing back eye, hair and skin color of deceased persons from analyzing their skeletal remains. Future efforts will be placed in enhancing the HIrisPlex-S prediction model for skin color.

Acknowledgements

We are grateful to all volunteers who provided biological samples and pigmentation phenotype information to this study. MK and LC are supported by Erasmus MC University Medical Center Rotterdam.

References

[1] Kayser M, Schneider PM, DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations, Forensic Sci Int Genet. 3 (2009) 154–161.

[2] Kayser M, de Knijff P, Improving human forensics through advances in genetics, genomics and molecular biology, Nat. Rev. Genet. 12 (2011) 179–192.

[3] Kayser M, Forensic DNA Phenotyping: DNA testing for externally visible characteristics, in: J.A. Siegel, P.J. Saukko (Eds.), Encyclopedia of Forensic Sciences, Academic Press, Waltham, 2013, pp. 369–374.

[4] Draus-Barini J, Walsh S, Pośpiech E, Kupiec T, Głąb H, Branicki W, Kayser M. Bona fide colour: DNA prediction of human eye and hair colour from ancient and contemporary skeletal remains. Investig Genet. 2013 Jan 14;4(1):3.

[5] Chaitanya L, Zupanič Pajnič I, Walsh S, Balažic J, Zupanc T, Kayser M, Bringing colour back after 70 years: Predicting eye and hair colour from skeletal remains of World War II victims using the HIrisPlex system. Forensic Sci Int: Genet. 2016, Available online 6 October

[6] King TE, Fortes GG, Balaresque P, Thomas MG, Balding D, Maisano Delser P, Neumann R, Parson W, Knapp M, Walsh S, Tonasso L, Holt J, Kayser M, Appleby J, Forster P, Ekserdjian D, Hofreiter M, Schürer K. Identification of the remains of King Richard III. Nat Commun. 2014 Dec 2;5:5631.

[7] Walsh S, Liu F, Ballantyne KN, 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 Jun;5(3):170-80.

[8]. Walsh S, Lindenbergh A, Zuniga SB, Sijen T, de Knijff P, Kayser M, Ballantyne KN. Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence. Forensic Sci Int Genet. 2011 Nov;5(5):464-71.

[9] Walsh S, Wollstein A, Liu F, Chakravarthy U, Rahu M, Seland JH, Soubrane G, Tomazzoli L, Topouzis F, Vingerling JR, Vioque J, Fletcher AE, Ballantyne KN, Kayser M. DNA-based eye colour prediction across Europe with the IrisPlex system. Forensic Sci Int Genet. 2012 May;6(3):330-40.

[10] 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 Jan;7(1):98-115.

[11] Walsh S, Chaitanya L, Clarisse L, Wirken L, Draus-Barini J, Kovatsi L, Maeda H, Ishikawa T, Sijen T, de Knijff P, Branicki W, Liu F, Kayser M. Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage. Forensic Sci Int Genet. 2014 Mar;9:150-61.

[12] Kastelic V, Pośpiech E, Draus-Barini J, Branicki W, Drobnič K. Prediction of eye color in the Slovenian population using the IrisPlex SNPs. Croat Med J. 2013 Aug;54(4):381-6.

[13] Chaitanya L, Walsh S, Andersen JD, Ansell R, Ballantyne K, Ballard D, Banemann R, Bauer CM, Bento AM, Brisighelli F, Capal T, Clarisse L, Gross TE, Haas C, Hoff-Olsen P, Hollard C, Keyser C, Kiesler KM, Kohler P, Kupiec T, Linacre A, Minawi A, Morling N, Nilsson H, Norén L, Ottens R, Palo JU, Parson W, Pascali VL, Phillips C, Porto MJ, Sajantila A, Schneider PM, Sijen T, Söchtig J, Syndercombe-Court D, Tillmar A, Turanska M, Vallone PM, Zatkalíková L, Zidkova A, Branicki W, Kayser M. Collaborative EDNAP exercise on the IrisPlex system for DNA-based prediction of human eye colour. Forensic Sci Int Genet. 2014 Jul;11:241-51.

[14] Valenzuela RK, Henderson MS, Walsh MH, Garrison NA, Kelch JT, Cohen-Barak O, Erickson DT, John Meaney F, Bruce Walsh J, Cheng KC, Ito S, Wakamatsu K, Frudakis T, Thomas M, Brilliant MH. Predicting phenotype from genotype: normal pigmentation. J Forensic Sci. 2010 Mar 1;55(2):315-22.

[15] Hart KL, Kimura SL, Mushailov V, Budimlija ZM, Prinz M, Wurmbach E. Improved eye- and skin-color prediction based on 8 SNPs.Croat Med J. 2013 Jun;54(3):248-56.

[16] Spichenok O, Budimlija ZM, Mitchell AA, 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 Nov;5(5):472-8.

[17] Allwood JS, Harbison S. SNP model development for the prediction of eye colour in New Zealand. Forensic Sci Int Genet. 2013 Jul;7(4):444-52.

[18] Maroñas O, Phillips C, Söchtig J, Gomez-Tato A, Cruz R, Alvarez-Dios J, de Cal MC, Ruiz Y, Fondevila M, Carracedo Á, Lareu MV. Development of a forensic skin colour predictive test. Forensic Sci Int Genet. 2014 Nov;13:34-44.

[19] Sturm RA. Molecular genetics of human pigmentation diversity. Hum Mol Genet. 2009 Apr 15;18(R1):R9-17.

[20] Jablonski NG, Chaplin G. The evolution of human skin coloration. J Hum Evol. 2000 Jul;39(1):57-106.

[21] Relethford JH. Hemispheric difference in human skin color. Am J Phys Anthropol. 1997 Dec;104(4):449-57.

[22]. Kayser M. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. Forensic Sci Int Genet. 2015 Sep;18:33-48.

[23]. Liu F, Visser M, Duffy DL, Hysi PG, Jacobs LC, Lao O, Zhong K, Walsh S, Chaitanya L, Wollstein A, Zhu G, Montgomery GW, Henders AK, Mangino M, Glass D, Bataille V, Sturm RA, Rivadeneira F, Hofman A, van IJcken WF, Uitterlinden AG, Palstra RJ, Spector TD, Martin NG, Nijsten TE, Kayser M. Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up. Hum Genet. 2015 Aug;134(8):823-35.

[24]. Branicki Wojciech and Manfred Kayser. Prediction of Human Pigmentation Traits from DNA Polymorphisms. In: eLS. John Wiley & Sons, Ltd: Chichester. 2015 Oct.

[25] Scientific Working Group on DNA Analysis Methods. SWGDAM Validation Guidelines, 2012 Available at: http://swgdam.org/SWGDAM_Validation_Guidelines_APPROVED_Dec_2012.pdf.

[26]. Liu F, van Duijn K, Vingerling JR, Hofman A, Uitterlinden AG, Janssens AC, Kayser M. Eye color and the prediction of complex phenotypes from genotypes. Curr Biol. 2009 Mar 10;19(5):R192-3

[27] Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA. Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W71-4.

[28] Vallone PM, Butler JM. AutoDimer: a screening tool for primer-dimer and hairpin structures. Biotechniques. 2004 Aug;37(2):226-31.

[29] Phillips C, Salas A, Sánchez JJ, Fondevila M, Gómez-Tato A, Alvarez-Dios J, Calaza M, de Cal MC, Ballard D, Lareu MV, Carracedo A; SNPforID Consortium. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genet. 2007 Dec;1(3-4):273-80.

[30] Green RL, Lagacé RE, Oldroyd NJ, Hennessy LK, Mulero JJ. Developmental validation of the AmpFℓSTR® NGM SElect™ PCR Amplification Kit: A next-generation STR multiplex with the SE33 locus. Forensic Sci Int Genet. 2013 Jan;7(1):41-51.

Chapter 5

Biogeographic matrilineal ancestry inference

Chapter 5.1

Developmental validation of mitochondrial DNA genotyping assays for adept matrilineal inference of biogeographic ancestry at a continental level

Forensic Science International: Genetics 11 (2014) 39 – 51

Abstract

Mitochondrial DNA (mtDNA) can be used for matrilineal biogeographic ancestry prediction and can thus provide investigative leads towards identifying unknown suspects, when conventional autosomal short tandem repeat (STR) profiling fails to provide a match. Recently, six multiplex genotyping assays targeting 62 ancestry-informative mitochondrial single nucleotide polymorphisms (mt-SNPs) were developed. This hierarchical system of assays allows detection of the major haplogroups present in Africa, America, Western Eurasia, Eastern Eurasia, Australia and Oceania, thus revealing the broad geographic region of matrilineal origin of a DNA donor. Here, we provide a forensic developmental validation study of five multiplex assays targeting all the 62 ancestry-informative mt-SNPs following the Scientific Working Group on DNA Analysis Methods (SWGDAM) guidelines. We demonstrate that the assays are highly sensitive; being able to produce full profiles at input DNA amounts of as little as 1 pg. The assays were shown to be highly robust and efficient in providing information from degraded samples and from simulated casework samples of different substrates such as blood, semen, hair, saliva and trace DNA samples. Reproducible results were successfully achieved from concordance testing across three independent laboratories depicting the ease and reliability of these assays. Overall, our results demonstrate the suitability of these five mt-SNP assays for application to forensic casework and other purposes aiming to establish an individual's matrilineal genetic ancestry. With this validated tool, it is now possible to determine the matrilineal biogeographic origin of unknown individuals on the level of continental resolution from forensic DNA samples to provide investigative leads in criminal and missing person cases where autosomal STR profiling is uninformative.

Introduction

The successful forensic application of DNA analysis methods as tools to exonerate innocents and to convict criminals has been expanding progressively, inculcating the recent technological advancements into practice to decipher the conundrum of human identity. Currently, autosomal short tandem repeats (STRs) are extensively used for human identification and thus aid in solving crimes with the available biological evidence. However, there is a necessity to employ additional DNA markers when conventional STR typing practically fails or does not yield a match to a known person including those whose profile is present in a criminal DNA database. The attractive features of mitochondrial DNA (mtDNA), i.e. its molecular stability, small size (16570 bp), high cellular copy number, maternal inheritance and haploidity, have made it an indispensable tool in forensic practice. Initially applied as identification tool for maternal lineages by use of DNA sequence information from the first hypervariable segment (HVS-I) (sometimes in combination with HVS-II) [1], particular mtDNA single nucleotide polymorphisms (mt-SNPs) are starting to be used for inference of matrilineal bio-geographic origin. Substantiating the matrilineal biogeographic origin of an unknown individual provides an impetus for investigations where an STR profile is unattainable from limited or degraded samples or in the absence of hits with known suspects including those represented in criminal DNA databases [1,2]. The traditional practice of Sanger sequencing of hypervariable portions of the mtDNA control region is not only laborious, it also does not always allow reliable haplogroup inference due to the high level of homoplasy (phylogenetically recurrent mutations) in this region, which can obscure phylogenetic signatures [3]. In contrast, relatively stable haplogroupdefining mt-SNPs and small insertion/deletion polymorphisms, mostly located within the coding region of the mt genome, allow straight-forward haplogroup inference and, due to the highly sensitive multiplex genotyping technologies available, less DNA is required as compared to sequencing [1]. Although sets of ancestryinformative autosomal SNPs have been suggested recently [4–7], the resolution at which biogeographic origin can be obtained with these SNP sets does not reach even the continental level, as geographic regions in close proximity such as Europe, Western and Central Asia and Northern Africa are usually poorly differentiated with these markers. However, decades of intensive population genetic research into the worldwide distribution of mtDNA genetic variation has yielded a large number of

ancestry-informative mt-SNPs including those to differentiate more close geographic regions [3,8–11]. Taking advantage of this knowledge, six SNaPshot multiplex assays, which collectively targeted 62 different mt-SNPs, were previously developed to discern 70 major mtDNA haplogroups typical for African, American, Western Eurasian, Eastern Eurasian, Australian and Oceanian populations [1,2]. These assays permit inference of the biogeographic region of maternal lineage origin to at least the continental, and sometimes subcontinental, level. The present study provides the developmental validation of five assays targeting all the 62 mt-SNPs based on the Scientific Working Group on DNA Analysis Methods (SWGDAM) guidelines [12]. When a new method is aimed to be introduced into the forensic community, the standardisations of the method and quality assurance are of key importance for their applicability in laboratories worldwide [13]. The SWGDAM developmental validation guidelines emphasize on the accuracy, precision, reproducibility and concordance of a novel procedure or a technique prior to its routine implementation in all the forensic laboratories [12]. With this study, we show that the five mt-SNP multiplex assays for inferring matrilineal biogeographic origin on the continental level fully meet the usual requirements for DNA test systems to be applied to forensic casework.

Materials and methods

Simulated casework samples were generated in-house and the majority of the DNA extractions were done using the QIAamp DNA Mini kit (Qiagen, Hagen, Germany) following the manufacturer's guidelines, except for the saliva samples that were extracted using an in-house protocol (unpublished), and vaginal swab DNA was extracted using Chelex (Sigma–Aldrich). See Supplementary Table 2 for sample and simulated casework details. The Ethnic Diversity DNA panel (EDP-1), manufactured by the European Collection of Cell Cultures (ECACC) and distributed by Sigma, consisting of purified genomic DNA from 92 individuals of diverse ethnic backgrounds ('African', 'Oriental', 'Australian Aborigine', 'Thai', 'Italian', 'French', 'South American Indian', 'Japanese' and 'Ashkenazi Jewish'), was used for the validation studies. Non-human DNA samples of cat, dog, mouse, pig, rat, chicken and bovine were obtained from Novagen, Inc. (Madison, WI); the chimpanzee sample was described elsewhere [14].

Multiplex design and protocol

Previously, we introduced six multiplexes targeting 62 mt-SNPs that delineate 70 matrilineal haplogroups associated with regions of Africa, Western Eurasia, Eastern Eurasia, America and Oceania (Australia, Near Oceania and Remote Oceania), respectively [1,2]. The three multiplexes (Multiplexes 1–3) we previously described for worldwide maternal ancestry inference except Oceania [1] were used here according to the previous publication with a few modifications to the protocol. The previously published multiplexes for inferring genetic ancestry within Oceania (Multiplexes 4–6) [2] were reduced to two multiplexes by reallocating the mt-SNPs of the previous Multiplex 5 to the revised Multiplexes 4 and 6, respectively, thereby increasing the productivity and decreasing the sample and reagent consumption. For all the final five multiplexes, PCR amplification was performed in a total reaction volume of 6 µl, containing PCR primers in specified concentrations (Multiplexes 1–3) [1] and containing PCR primers in specified concentrations for Multiplexes 4 and 6 (shown in Table 1), 1X GeneAmp PCR Gold Buffer (Applied Biosystems, Foster City, USA), 4.5 mM MgCl2 (Applied Biosystems), 100 µM of each dNTP (Roche, Mannheim, Germany), 0.3 µL of 5 U/ml AmpliTaq Gold DNA Polymerase (Applied Biosystems), and 1 µL of DNA extract containing varying amounts from 1 pg to 1000 pg. All the amplifications were performed on a GeneAmp PCR System 9700 thermocycler (Applied Biosystems) in 384-well reaction plates (Applied Biosystems), with the following thermocyclic parameters: 95°C for 10 minutes; 30 cycles of 94°C for 15 seconds, 60°C for 45 seconds; 60°C for 5 minutes. The PCR products were purified with 2 mL of ExoSAP-IT (USB Corporation, OH, USA) and incubated at 37°C for 45 minutes and 80°C for 15 minutes. Multiplex single-base primer extension reactions were carried out with a GeneAmp PCR System 9700 thermocycler (Applied Biosystems) in a 384-well reaction plates (Applied Biosystems), in a total reaction volume of 6 µL. The reaction contained extension primers at the specified concentrations ([1], Table 1), 1 µL of the SNaPshot Ready Reaction Mix and 1 µL of the purified PCR product, with the following thermo-cyclic parameters: 96°C for 2 minutes; 25 cycles of 96°C for 10 seconds, 50°C for 5 seconds, 60°C for 30 seconds. The extended products were cleaned with 1 µL of Shrimp Alkaline Phosphatase (SAP) (USB Corporation) by incubating at 37°C for 45 minutes and 75°C for 15 minutes. The cleaned extension products were electrophoretically separated and detected on a 3130xl Genetic Analyzer

(Applied Biosystems) with POP-7 polymer, by running a mixture of 1 µL extension product, 8.5 µL HiDi formamide (Applied Biosystems) and 0.5 µL of LIZ-120 (Applied Biosystems), with a 23 seconds injection time at 1.2 kV. Results were analyzed using the Gene Mapper v3.7 software (Applied Biosystems).

Sensitivity study

The sensitivity of the multiplexes was assessed to determine the minimum concentration of DNA with which accurate and reliable results could be obtained. The sensitivity of Multiplexes 1–3 were previously reported to be 4 pg of total DNA (including genomic and mtDNA) [1]. However, the current study repeated this test with different samples due to the changes applied to the previously published protocol [1]. Variable amounts of DNA input ranging from 1000 pg to as little as 1 pg were run and their profiles evaluated. Five DNA samples from different populations – 'French' (ECACC ID: HO104); 'Australian Aboriginal' (ECACC ID: WON,I); 'Italian' (ECACC ID: J0528239); 'Black African' (ECACC ID: MP0008) and 'South American Indian' (ECACC ID: GRC-138) were genotyped at 500 pg, 250 pg, 125 pg, 63 pg, 32 pg, 4 pg, and 1 pg DNA input. All the DNA dilutions were quantified and confirmed in duplicates with the Quantifiler® Human DNA Quantification kit (Applied Biosystems) following the manufacturer's guidelines. It is to be noted that detecting extremely low concentrations are below the Quantifiler® Human DNA Quantification kit's detection limit. The lower concentrations were prepared, in duplicates, by careful dilution from reliably measured (Quantifiler1 Human DNA Quantification kit) samples of higher concentration.

Electrophoretic sizing precision and accuracy

To assess the ability of the analytical instrument ABI 3130xl to produce reliable data and demonstrate sizing accuracy, five samples from different populations – 'French' (ECACC ID: HO104); 'Australian Aboriginal' (ECACC ID: WON,I); 'Italian' (ECACC ID: J0528239); 'Black African' (ECACC ID: MP0008) and 'South American Indian' (ECACC ID: GRC-138) were genotyped five times at a genomic DNA (gDNA) input amount of 125 pg. The average and the standard deviation for each marker peak position were calculated across all five multiplexes.

Mixture Studies

DNA from individuals of known genotype, at a gDNA concentration of 500 pg/µl was mixed in the ratios of 1:1, 1:5 and 1:10, respectively. Individuals from different haplogroups were combined to determine the ability of the assay to decipher the different haplogroups from mixed samples.

Reproducibility

SWGDAM emphasizes on producing concordant results between different laboratories. Our concordance study was performed by testing 48 DNA samples of differing gDNA concentrations by three independent labs. Only one of the three labs had previous experiences with these particular assays, whereas the other two labs have general experience with using the SNaPshot technology for SNP typing. Depending on the laboratory, fragments were separated and genotyped on an ABI 3100 machine using POP-4 or on an ABI 3130xl machine using POP-7 polymer, and results were analyzed using GeneMarker (Soft Genetics) or GeneMapper v 4.0 (Applied Biosystems) softwares.

Specificity and stability

There are always chances of contamination from other biological sources at a crime scene. To assess the human specificity of the assay, a variety of non-human species were tested. Commercially available DNA samples of dog, pig, bovine, mouse, rat, and chicken at measured concentrations of 50 ng/ml and chimpanzee at 1 ng/ml were used. The robustness of the assays to type degraded samples was tested by exposing DNA from two individuals (a 'Japanese' with haplogroup M80/D and a 'South American Indian' with haplogroup A2*(xA2a,A2b)), to ultraviolet light at intervals of 0 seconds, 30 seconds, 60 seconds, 90 seconds, 5 minutes, 15 minutes, 30 minutes, and 60 minutes, using a Bio-Link (Vilber Lourmat) at a strength of 50 J/cm2, and subsequently genotyped to determine the stability when exposed to such conditions.

Population study

All the samples of the Ethnic Diversity ECACC Panel, representing 92 individuals from nine (as defined by the manufacturer) different population groups, were genotyped using all five multiplex assays

Multiplex	PCR Amplification					Single-base extension				
	Site		Primer Sequences(5'--3')	Conc	Amplicon	Primer Sequences(5'--3') (5' aspecific tail in lowercase italics)	Conc	Length (nt)	Orientation	Alleles (dye)
4	8404	F	TTTACAGTGAAATGCCCCAAC	0.4	93	t(gact)$_9$g GTGTAAGGAGTATGGGGGT	0.5	57	R	T (green), C (blue)
		R	TTTTTAGTTGGGTGATGAGGAA							
4	9123	F	CCACCCTAGCAATATCAACCA	0.5	147	(gact)$_{13}$ga TTATCATCTTCACAATTCTAATTCTACT	0.27	82	F	G (blue), A (green)
		R	TGTTGTCGTGCAGGTAGAGG							
4	9140	F	Same amplicon as 9123			act(gact)$_{12}$g ATTCTACTGACTATCCTAGAAATCG	0.7	77	F	C (yellow), T (red)
		R								
4	9156	F	Same amplicon as 9123			(gact)$_{11}$ga TAGAAGTGTGAAAACGTAGGC	0.3	67	R	A (red), G (yellow
		R								
4	10214	F	GAAAAATCCACCCCTTACGA	2	129	(gact)$_{19}$ CCGCGTCCCTTTCTC	2.4	91	F	C (yellow), T (red)
		R	GGGTAAAAGGAGGGCAATTT							
4	10400	F	GCCCTAAGTCTGGCCTATGA	0.25	90	(gact)$_5$gac CGTTTTGTTTAAACTATATACCAATTC	0.15	50	R	C (blue), T (green)
		R	TGAGTCGAAATCATTCGTTTT							
4	11251	F	TGAACGCAGGCACATACTTC	0.05	92	(gact)$_3$gac CCCCTACTCATCGCACT	0.1	32	F	A (green), G (blue
		R	TGAGCCTAGGGTGTTGTGAG							
4	12705	F	CCCAAACATTAATCAGTTCTTCAA	0.7	141	act(gact)$_5$g TTAATCAGTTCTTCAAATATCTACTCAT	0.3	52	F	T (red), C (yellow)
		R	CATCAACTGATGAGCAAGAAGG							
4	12771	F	Same amplicon as 12705			ct(gact)$_{12}$g GAAGGATATAATTCCTACGCC	0.5	72	R	G (yellow), A (red)
		R								
4	12940	F	CAACACAGCAGCCATTCAAG	0.3	157	act(gact)$_3$gac CATGAGACCCACAACAAATA	0.175	38	F	G (blue), A (green)
		R	GAGGCCTAGTAGTGGGGTGA							
4	13500	F	ATTGGCAGCCTAGCATTAGC	0.15	68	t(gact)$_{10}$g GGAATACCTTTCCTCACAGG	0.1	62	F	T (red), C (yellow)
		R	TCGATGATGTGGTCTTTGGA							
4	14025	F	ACCGCACAATCCCCTATCTA	0.25	124	act(gact)$_{17}$g TAACCTGACTAGAAAAGCTATTACC	0.15	97	F	T (red), C (yellow)
		R	ATGGAGGTGGAGATTTGGTG							
4	15301	F	CCACCCTCACACGATTCTTT	0.25	119	(gact)$_5$ ATTCTTTACCTTTCACTTCATCTT	0.15	44	F	A (green), G (blue)
		R	GGTGATTCCTAGGGGGTTGT							
4	15607	F	AATTCTCCGATCCGTCCCTA	0.7	131	ctgactga CGATCCGTCCCTAACAA	0.8	25	F	A (green), G (blue
		R	ATTGGCTTAGTGGGCGAAAT							

6	3645	F	CCTCAACCTAGGCCTCCTATTT	0.25	155	$(gact)_7$ga CACCTCTAGCCTAGCCGT	0.15	48	F	T (red), C (yellow)
		R	TGATGGCTAGGGTGACTTCAT							
6	5899	F	CAATGCTTCACTCAGCCATTT	0.25	112/111	ct$(gact)_5$ GCGAACATCAGTGGGG	0.15	38	R	a (blue), d (red)a
		R	TCATGCGCCGAATAATAGGT							
6	6077	F	GGTAACGACCACATCTACAACG	0.25	87	act$(gact)_3$g CAAATGCATGGGCTGT	0.2	32	R	C (blue), T (green)
		R	GCCTCCGATTATGATGGGTAT							
6	6719	F	ATCCTACCAGGCTTCGGAAT	2	134	$(gact)_{10}$g AAGAACCATTTGGATACATAGG	1	63	F	T (red), C (yellow)
		R	TGGTGTGCTCACACGATAAA							
6	11061	F	CTCCGACCCCCTAACAAC	1.2	179	act$(gact)_{17}$ga AAAAAAACTCTACCTCTCTATACTAATCT	1.2	92	F	C (yellow), T (red)
		R	TGATTAGTTCTGTGGCTGTGAA							
6	11288	F	CAACACCCTAGGCTCACTAAACA	0.25	138	act$(gact)_8$ga ACCCTAGGCTCACTAAACATT	0.15	58	F	C (yellow), T (red)
		R	AAGTGGAGTCCGTAAAGAGGTA							
6	11641	F	CAAACAGACCTAAAATCGCTCA	0.4	120	ct$(gact)_7$ga ATACTCTTCAATCAGCCACAT	0.4	53	F	A (green), G (blue)
		R	GGCGATTATGAGAATGACTGC							
6	12121	F	CTATCCCTCAACCCCGACAT	0.5	126	$(gact)_6$ga CTCAACCCCGACATCAT	0.5	43	F	T (red), C (yellow)
		R	TTCTTGTGAGCTTTCTCGGTAA							
6	14022	F	ACCGCACAATCCCCTATCTA	0.5	124	t$(gact)_{14}$ GTGCTGTGAAATTGTTTTAGG	0.5	78	R	A (red), G (yellow)
		R	ATGGAGGTGGAGATTTGGTG							
6	14890	F	TGAAACTTCGGCTCACTCCT	0.25	131	ct$(gact)_{11}$ AGTAGTGCATGGCTAGGAATAG	0.5	68	R	A (red), G (yellow)
		R	TTTACGTCTCGAGTGATGTGG							
6	15607	F	AATTCTCCGATCCGTCCCTA	0.7	131	act$(gact)_{16}$ga CGATCCGTCCCTAACAA	0.6	86	F	A (green), G (blue)
		R	ATTGGCTTAGTGGGCGAAAT							
6	15172	F	GAGGCCAAATATCATTCTGAGG	1	149	t$(gact)_{18}$ GCCAAATATCATTCTGAGG	0.25	102	F	G (blue), A (green)
		R	GATGAAGTGAAAGGTAAAGAATCG							
6	16247	F	CCCATGCTTACAAGCAAGTACA	2	108	t$(gact)_{13}$gac GAGRGGTGGCTTTGGAG	1.05	73	R	A (red), G (yellow)
		R	AAGGGTGGGTAGGTTTGTTG							

Table 1: Genotyping details of the different mt-SNPs detectable by Multiplexes 4 and 6 (for Multiplexes 1 - 3 see [1]).

Results and discussion

Multiplex re-design

Several modifications were applied to the multiplex design from our previous publications [1,2]. The total number of multiplexes was reduced from six to five, thus reducing the amount of sample needed, reagents used, and labour/time spent. In particular, previous Multiplex 5 that was used to resolve haplogroup Q into its sub- haplogroups Q1, Q2, Q2a and Q3 using the respective markers 12940, 14025, 11061, 10214 and 15172 [2], was split and re- allocated to the previous Multiplexes 4 and 6. To avoid primer dimer formation all the primers in Multiplexes 4–6 were screened using AutoDimer [15]. Based on the primer compatibility we decided to allocate markers 10214 (Q2a) and 14025 (Q1) to Multiplex 4. It is to be noted that marker 12940, diagnostic for basal haplogroup Q, was already included in Multiplex 4 and hence did not need to be re- allocated. The remaining two markers (11061, Q2, and 15172, Q3) were allocated to Multiplex 6. The 50 aspecific tail lengths of the extension primers were re-adjusted to ensure their electrophoretic separation in the revised multiplexes. The markers, alleles detected and the predicted haplogroups after the modifications of Multiplexes 4 and 6 are summarized in Table 1. Some of the marker phylogenies as published previously [1,2] were updated in light of the latest worldwide mtDNA tree (PhyloTree Build 15, http://www.phylotree.org; [16]). For example, marker 4883 included in Multiplex 1 is no longer solely defining haplogroup D but now also defines haplogroup M80 [17]. Supplementary Figure 1 provides the updated marker phylogenies for each of the five multiplexes. Figure 1 illustrates the targeted mt-SNPs and their respective geographic haplogroup classification [1,2] as detected by the five multiplex assays.

In addition to the above-mentioned alteration, the primer concentrations of Multiplexes 4 and 6 were re-adjusted to achieve a highly optimized assay after the addition of the above-mentioned primers. Another change that was applied to all the five multiplex assays was that after the addition of ExoSAP-IT (USB Corporation, OH, USA), incubation time at 37^{o}C was increased to 45 minutes (was: 15 minutes) to enhance primer and dNTP removal thereby resolving some artefacts. Finally, the amount of AmpliTaq used in all the five multiplex assays was increased from 0.07 mL to 0.3 mL per reaction, thereby enhancing the sensitivity, thus enabling the detection of very low templates.

The multiplexes were designed in a hierarchical fashion but also provide flexibility so that any prior knowledge or expectation about the sample under investigation can be used to select the most appropriate multiplex to start with. For instance, if there is a reasonable expectation that the sample comes from an individual of European origin then it makes sense to start with Multiplex 2, because the major European haplogroups (H, U, K, J and T) are all targeted by Multiplex 2. Hence, there is no need for testing a sample with all five multiplexes. Consider another instance where the unknown sample is suspected to be of Oceanian matrilineal origin, and then the starting assay should be Multiplex 4 followed by Multiplex 6 for further resolution if needed.

A recent publication [18] reported a phylogenetic instability of the marker 16247, included in Multiplex 6, which appears to have undergone multiple back mutations to its ancestral state. This A to G transition denotes haplogroup B4a1a1a (at the HVS-I level also known as the Polynesian Motif) and its sub-lineages (as shown in Supplementary Figure 1). It is now considered highly unstable within B4a1a1 and hence not reliable in inferring haplogroup B4a1a1a status, which is especially frequent in Remote Oceania [19,20]. However, in Multiplex 6 we also target position 14022, an A to G transition defining the directly ancestral haplogroup B4a1a1 (as shown in Supplementary Fig. 1), representing the immediate predecessor of B4a1a1a [21]. Hence, we can be sure that with Multiplex 6 haplogroup B4a1a1 can be determined reliably. The 16247 results can then either be ignored or interpreted with the necessary caution as outlined elsewhere [18]. In addition to 16247, there are some more recurrent markers. For every recurrent marker, a stable upstream marker has been included in each of the multiplexes as a safeguard. For instance, in Multiplex 4 the marker 15607 (A–G transition) can point to either haplogroup T membership or haplogroup P membership. In order to distinguish between the two possibilities, an intermediate marker 11251 (defining JT) has been included which is present in derived state in haplogroup T but not in P. Another marker, 8281–8289 in Multiplex 2, is mildly recurrent throughout the mtDNA phylogeny [16]. However, no other marker was available as diagnostic for haplogroup B405 so we were left with no other choice than to target 8281–8289 for this haplogroup.

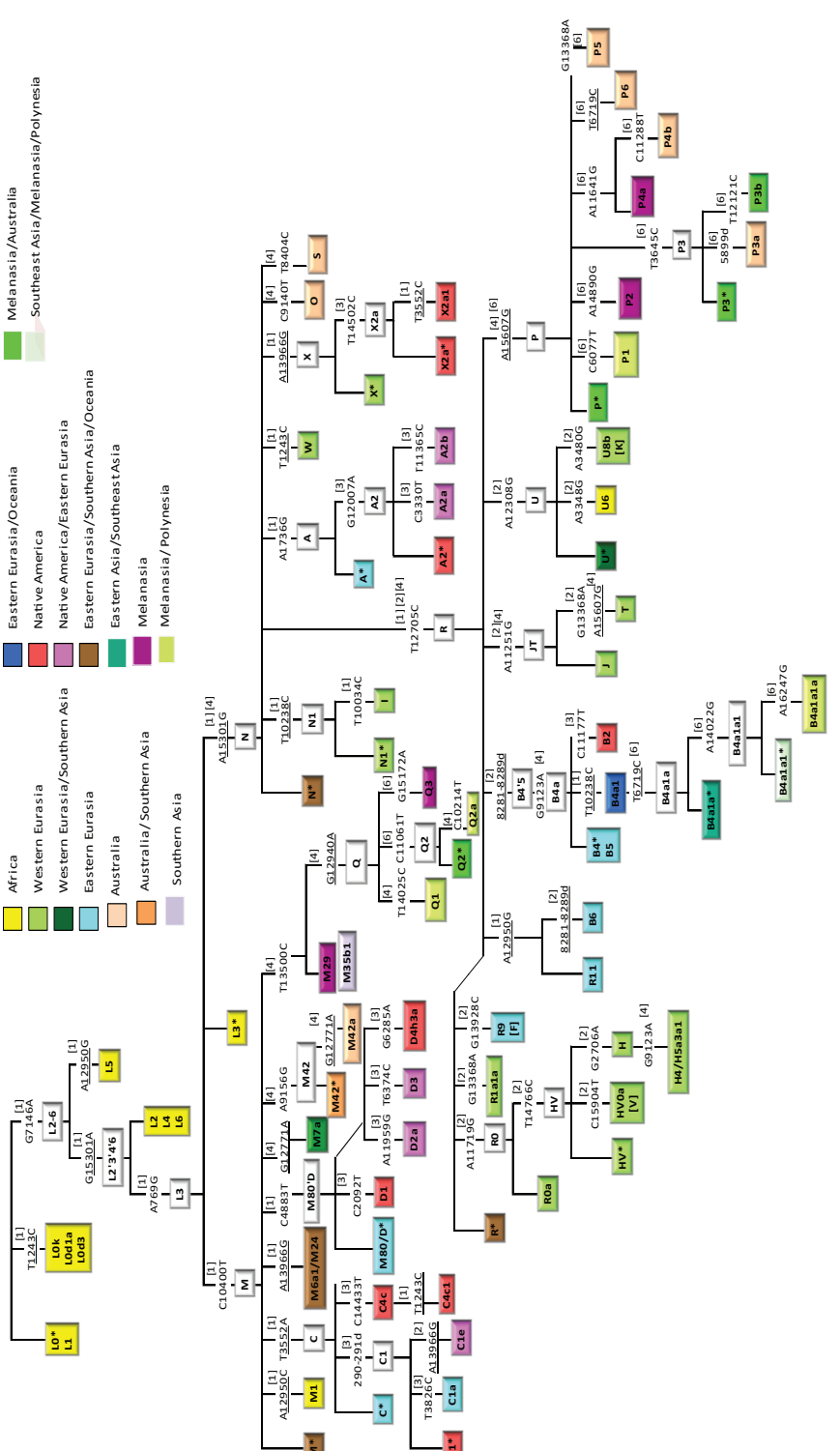


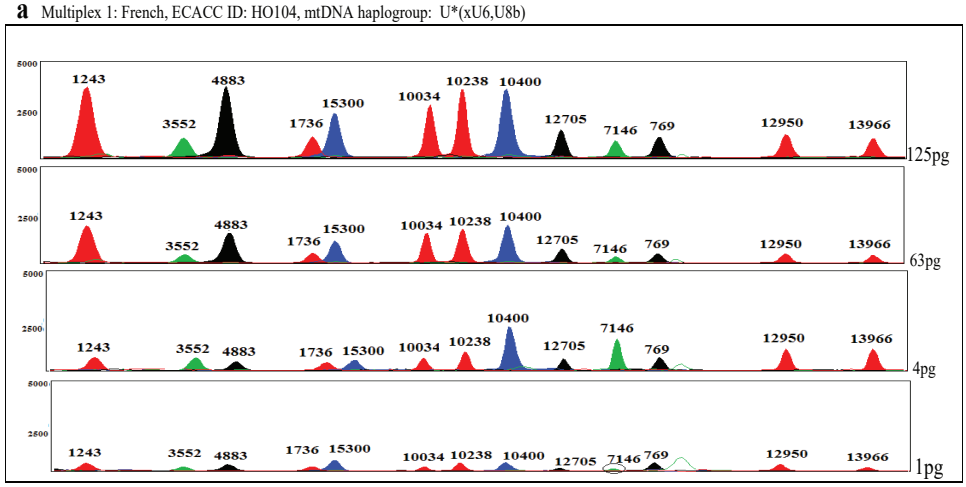
Figure 1: Phylogenetic relationships of the targeted mt-SNPs and the geographic haplogroup classification. The combined use of the five multiplexes enables the assignment of any individual's mtDNA to one of the colour-coded haplogroups. Each colour corresponds to the most probable geographic origin of that haplogroup. SNP position numbers are relative to the Cambridge Reference Sequence (rCRS) and the numbers within square brackets indicates the multiplex in which the respective SNP is included. Recurrent SNPs are underlined.

Sensitivity studies

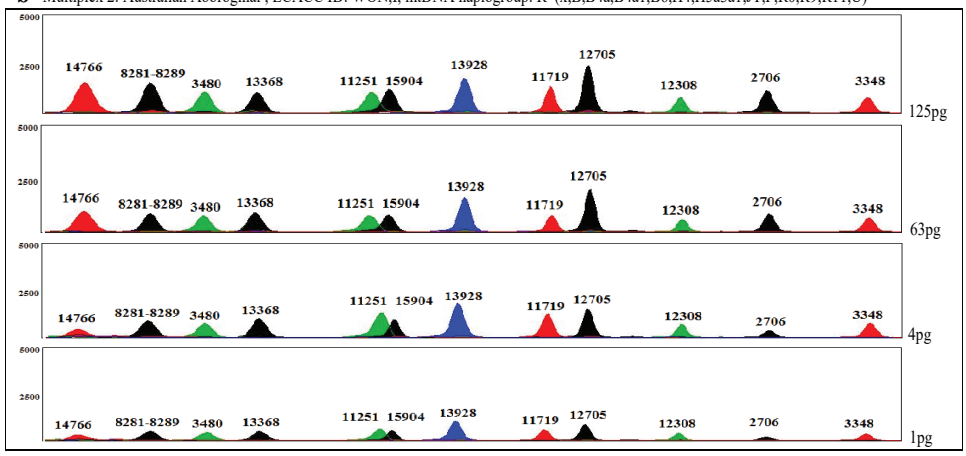
In order to determine the limit of detection of the five assays, they were applied to DNA amounts (genomic DNA) ranging from 1000 pg to 1 pg. Five samples from five different populations – 'French' (ECACC ID: HO104); 'Australian Aboriginal' (ECACC ID: WON,I); 'Italian' (ECACC ID: J0528239); 'Black African' (ECACC ID: MP0008) and 'South American Indian' (ECACC ID: GRC-138) were analyzed and the results of the sensitivity study on samples with DNA inputs ranging from 125 pg to 1 pg are summarized in Figure 2. Most notably, it was possible to achieve high sensitivity in all samples across all the five assays even at 1 pg input DNA. In the previous publication describing the Multiplexes 1–3 [1], dropouts were observed at 1 pg, but the increased amount of AmpliTaq polymerase as currently applied with the modified protocol has enhanced the sensitivity and therefore the robustness of the assays. Even at a very low template amount of 1 pg DNA input a full profile was obtained with a high intensity of above 100 RFUs for all peaks in Multiplexes 2 and 6. However, the marker 7146 (Multiplex 1); markers 6285, 290–291 and 12007 (Multiplex 3); markers 11251, 8404, 10214 and 14025 (Multiplex 4) were detected below 85 RFUs, but still above the usually applied detection threshold of 50 RFUs. Notably, our five assays demonstrated a sensitivity that is similar to two previously published publications [10,22] and higher than the GenPlex SNP typing system [23] and SNPfor ID 52-plex [24] assay which failed to produce 100% profiles from 500 pg; higher than the commercially available Powerplex1 ESI System that produced full profiles only from 62.5 pg [25]; and higher than the commercially available STR multiplex kit – AmpFlSTR1 NGM Select that showed stochastic dropout of alleles below 125 pg [26]. Our multiplexes were also undoubtedly higher than many other mtDNA multiplexes [3,9,11,27].

The use of 23 s injection time (previously 10 s) and the high abundance of mtDNA compared to nuclear DNA would provide an explanation for the high sensitivity achieved here. Nonetheless, our results indicated that the five mt-SNP multiplexes studied here are efficient and advantageous for extremely difficult samples. It is to be noted that both positive and negative controls were included in the reactions, the negative controls were clean, and no other artefacts were observed which could interfere with allele calling.

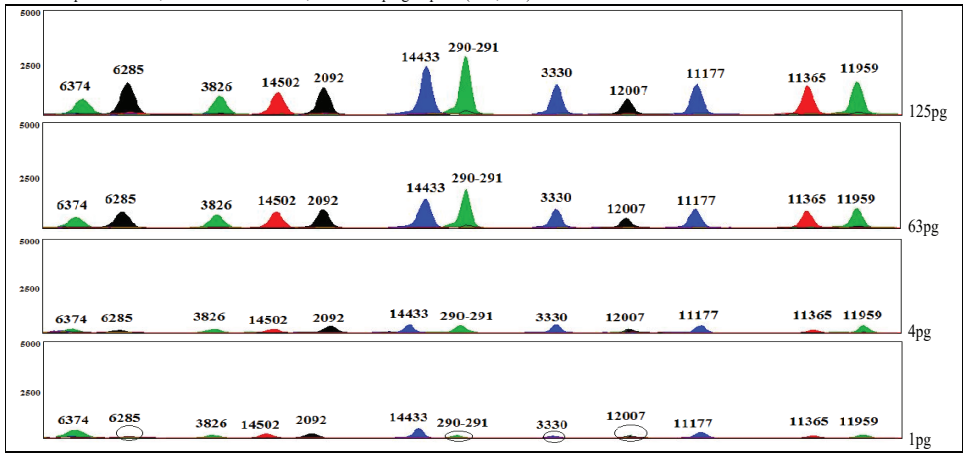
a Multiplex 1: French, ECACC ID: HO104, mtDNA haplogroup: U*(xU6,U8b)



b Multiplex 2: Australian Aboroginal , ECACC ID: WON,I, mtDNA haplogroup: R*(x,B,B4a,B4a1,B6,H4,H5a3a1,JT,P,R0,R9,R11,U)



c Multiplex 3: Italian, ECACC ID: J0528239, mtDNA haplogroup: U*(xU6,U8b)



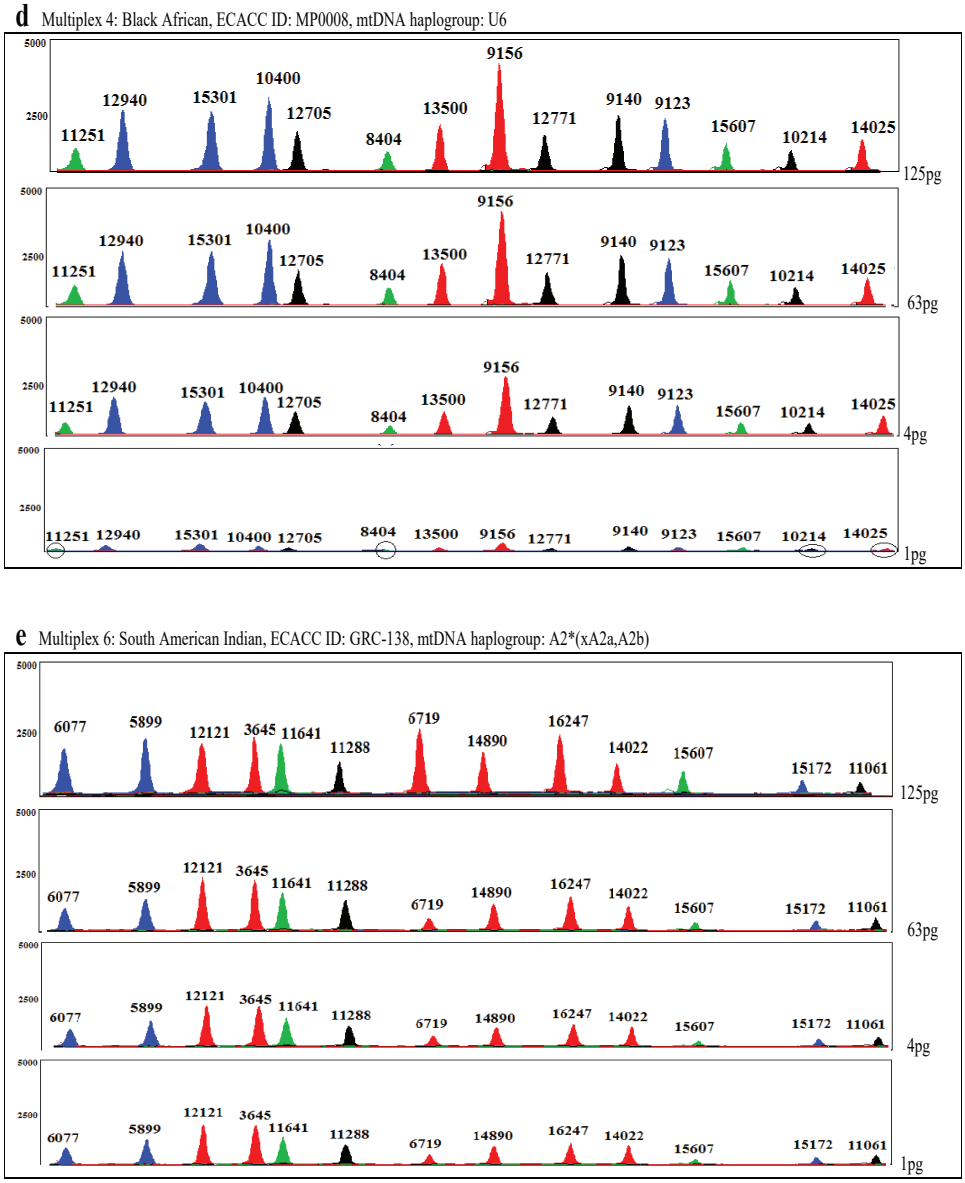


Figure 2*:* Sensitivity plots for all five mt-SNP multiplex assays performed on samples belonging to different haplogroups with DNA inputs ranging from 125 pg to 1 pg (scaled to 5000 RFUs). Grey circles indicate markers which were detected below 100 RFUs (but still above the detection threshold of 50 RFUs). Geographic sample designation and IDNr. as used by the ECACC Ethnic Diversity DNA Panel (Sigma).

An important caveat to be kept in mind for all the assays is that the primer quality drastically affects the sensitivity of the assays. We experienced that repeated rounds of freezing and thawing of the primers resulted in primer degradation and failure to produce optimal results with full profiles even at 250 pg DNA input. This shall be considered when applying the assays, especially to challenging samples in future case work. For the final experiments herein described we therefore always used freshly ordered primers to avoid primer degradation that would affect our results. Another solution to avoid the repeated rounds of freezing and thawing could be aliquoting the freshly ordered primers into smaller volumes upon arrival.

Precision studies

The sizing precision and accuracy of an instrument is very important to avoid incorrect calling of alleles, which is crucial when dealing with unknown DNA samples. To address this issue in our validation study, three samples were injected five times and the average standard deviation of the peak positions for each marker across the five multiplexes was calculated. The maximum deviation observed for each marker (in nt) between samples was 0.32 (Multiplex 1); 0.42 (Multiplex 2); 0.37 (Multiplex 3); 0.25 (Multiplex 4); 0.17 (Multiplex 6) And the minimum was 0.04 (Multiplex 1); 0.01 (Multiplex 2); 0.04 (Multiplex 3); 0.03 (Multiplex 4); 0.02 (Multiplex 6). The average standard deviation is less than 0.5 across all multiplexes. However, it is to be noted that the precision is affected by a number of factors, of which a few are temperature fluctuations, capillary length, voltage, and polymer [28] not tested here.

Mixture studies

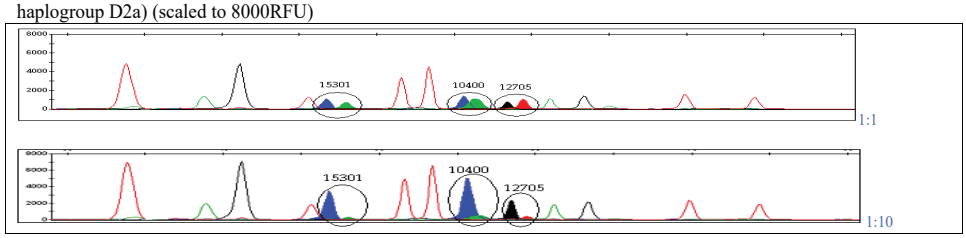
Samples from two individuals of different biogeographic origin (Multiplexes 1 and 2 – 'Italian', BM15 and 'Oriental', HOSONUM; Multiplexes 3 and 4 – 'South American Indian', GRC 212 and 'Oriental', HOSONUM; Multiplex 6 – 'Australian Aboriginals', NON,L and BUR,E) were mixed in the ratios 1:1, 1:5 and 1:10. The results from all five assays with samples mixed in the ratios 1:1 and 1:10 are illustrated in Figure 3. Obviously, a mixture can only be detected if a mt-SNP in the different contributors has a different allele (i.e. if the contributors belong to a different haplogroup at the resolution level offered by the respective multiplex). Given that some of the mtDNA haplogroups detectable with our assays are quite common in certain geographic areas (e.g. haplogroup H is typically observed in Europe

at 40–50%) there is a considerable chance that two contributors coming from the same geographic region belong to the same haplogroup and thus have the same allele at the haplogroup defining mt-SNP included in our assays, which would leave the mixture undetectable from the mtDNA perspective. However, when the components of the mixture represent different haplogroups, our results demonstrate that those components can be resolved with this assay at mixture ratios of 1:10 [14].

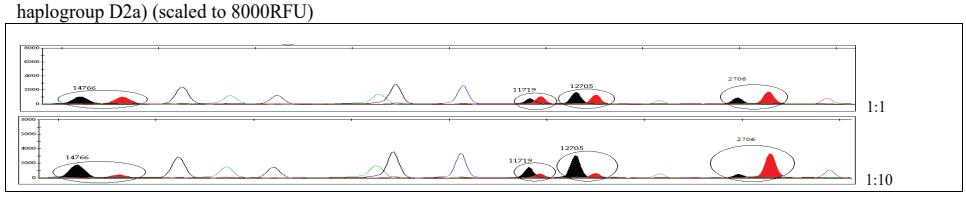
Reproducibility studies

To determine the reproducibility of the assays, a set of 48 samples from individuals with different mtDNA haplogroups, from different populations and of different gDNA concentrations were genotyped and assessed by three independent labs. The two external labs were the Forensic Laboratory for DNA Research (FLDO) at the Leiden University Medical Centre (LUMC) and the R&D group of the Department of Human Biological Stains at the Netherlands Forensic Institute (NFI). Both labs had no previous experience with these particular assays, but did have general experience in the use of the SNaPshot technology for SNP genotyping. Protocols and the phylogenetic trees to determine the haplogroups were provided to both participants. Such a blind concordance test examines the efficiency and practical applicability of the assays when performed by individuals with no prior experience of the five multiplexes. The results from the concordance testing are summarized in Supplementary Table 1. The majority of the results were in concordance; however in one instance one of the laboratoriescould not correctly predict the mixture samples. Though the genotypes were in concordance, they faced difficulty in assigning the appropriate haplogroup. Also, it is to be noted that one of the laboratories faced difficulties with one marker each in Multiplexes 1 (marker 3552), and 3 (marker 6374). These markers in the hands of this lab often showed allelic dropouts (marker 3552 – 85.4% and marker 6374 – 79.1%) and a plausible explanation could be primer degradation as described earlier. However, for this set of test samples the two affected markers were not critical for haplogroup inference, hence when tested across all the five assays, the haplogroups of the samples and thereby their biogeographic origins were still determined correctly despite the genotyping problems that occurred. Another issue that the laboratory faced was bin shifts, especially with the marker 11251. It is known that even slight temperature variations can cause mobility shifts and this issue could be resolved by adjusting the bins [29].

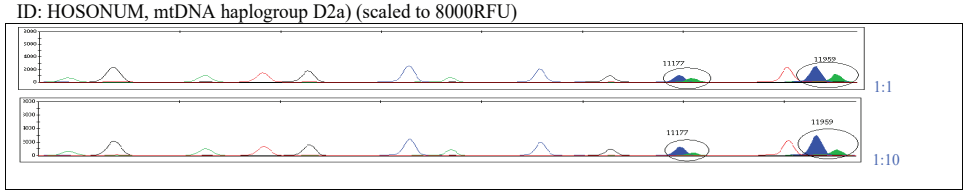
a Multiplex1 Mixture of Italian (ECACC ID:BM15, mtDNA haplogroup H) and Oriental (ECACC ID: HOSONUM, haplogroup D2a) (scaled to 8000RFU)



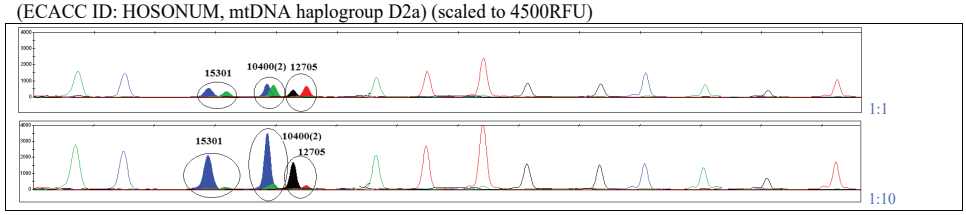
b Multiplex 2: Mixture of Italian (ECACC ID:BM15, mtDNA haplogroup H) and Oriental (ECACC ID: HOSONUM, haplogroup D2a) (scaled to 8000RFU)



c Multiplex 3:Mixture of South American Indian (ECACC ID: GRC 212, mtDNA haplogroup B2)and Oriental (ECACC ID: HOSONUM, mtDNA haplogroup D2a) (scaled to 8000RFU)



d Multiplex 4: Mixture of South American Indian (ECACC ID: GRC 212, mtDNA haplogroup B2)and Oriental (ECACC ID: HOSONUM, mtDNA haplogroup D2a) (scaled to 4500RFU)



e Multiplex 6:Mixture of two Australian Aboriginals (ECACC ID: NON,L mtDNA haplogroup P5 and ECACC ID: BUR,E mtDNA haplogroup M42a) scaled to 7000RFU)

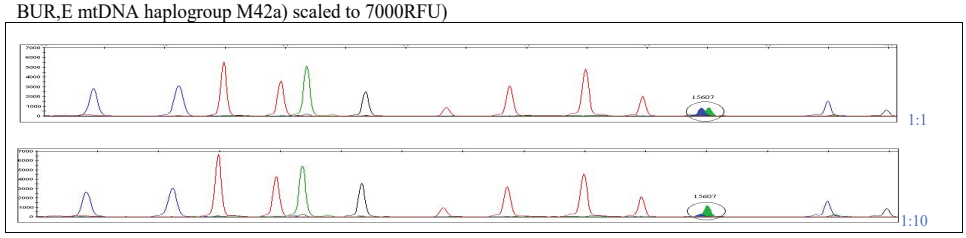


Figure 3: Mixture study for all five mt-SNP multiplex assays of 1:1 and 1:10 mixtures from DNA samples of individuals with different geographic origin and mtDNA haplogroups. Circles indicate those mtDNA SNPs that clearly indicate the presence of a mixture. Geographic sample designation and ID-Nr. as used by the ECACC Ethnic Diversity DNA panel (Sigma).

Specificity testing

Non-human DNA samples were tested across all five multiplexes. A chimpanzee profile is depicted in Figure 4; peaks were present at certain loci as expected due to the close evolutionary relationship between chimpanzees and humans. All the other, evolutionary more distant species tested such as dog, cat, rat, mouse, pig, bovine and chicken showed no amplification, even at 50 ng input DNA, across majority of the loci. In Multiplex 4 a weak peak for position 8404 was observed in the dog sample but at very low relative fluorescence units (54 RFUs). Nevertheless, the specificity study provides enough evidence for the application of these assays in routine casework.

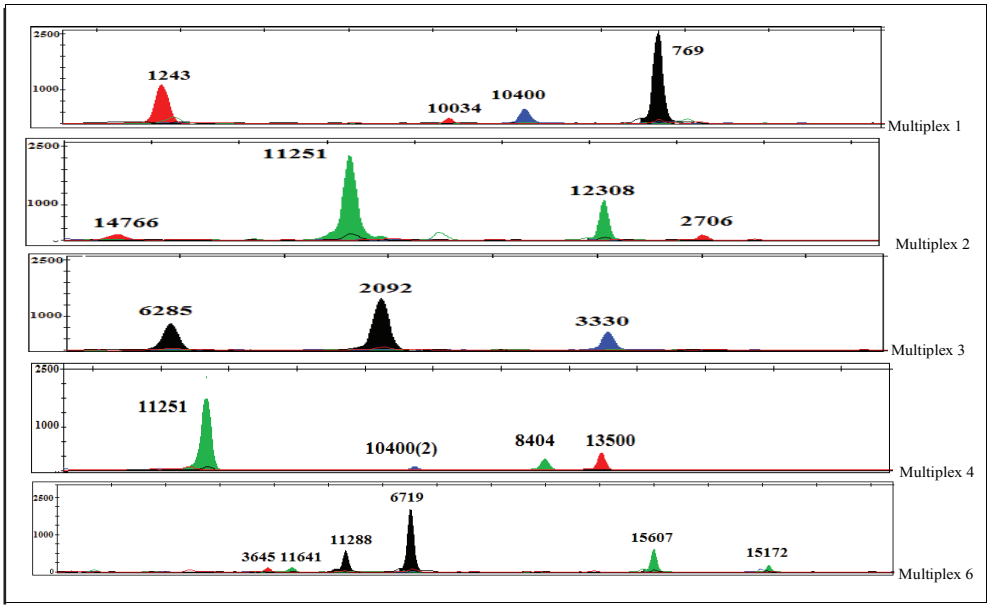


Figure 4: Specificity study performed for all five mt-SNP multiplex assays on 1 ng of chimpanzee DNA (scaled to 2500 RFUs). Loci in each of the multiplex assay where peaks were observed are indicated in bold black.

Stability/degradation testing

Two individuals with different haplogroups, a 'Japanese' (ECACC ID: KT12) with haplogroup M800D and a 'South American Indian' (ECACC ID: GRC-138) with haplogroup A2*(xA2a,A2b), were subjected to ultraviolet exposure at time intervals of 0 seconds, 30 seconds, 60 seconds, 90 seconds, 5 minutes, 15 minutes, 30 minutes, and 60 minutes, to test the effect of DNA degradation on the performance of the multiplexes. UV light is known to fragment DNA [30] and we therefore used this experiment to assess the impact of DNA degradation on the five mt-SNP multiplex assays. Full profiles were observed up to 15 minutes of continuous exposure to UV light in all the multiplexes. Multiplexes 1, 4 and 6 showed full profiles even up to 30 minutes of UV exposure, whereas Multiplexes 2 and 3 experienced dropouts at 30 minutes at position 8281–8289 and 6374 respectively. All five multiplexes showed dropouts after 60 minutes of UV exposure (see Table 2 for the full results). Our results thus indicate that beyond 30 minutes of exposure to UV light, DNA damage is extreme for the amplicon size range employed, and dropout is expected as the DNA is becoming fully degraded. However, the robustness of the assays is clearly evident from the full profiles still obtained from the DNA samples with up to 30 minutes of UV degradation.

Time	Multiplex 1		Multiplex 2		Muliplex 3		Multiplex 4		Multiplex 5	
	Individual 1	Individual 2	Individual 1	Individual 2	Individual 1	Individual 2	Individual 1	Individual 2	Individual 1	Individual 2
0sec	✓		✓		✓		✓		✓	
15min	✓		✓		✓		✓		✓	
30min	✓		X		X		✓		✓	
60min	X		X		almost no profile		X		X	

Table 2: Effect of UV light exposure on the five mt-SNP multiplex assays for two samples: a Japanese (ECACC ID: KT12 with haplogroup M800D and a South American Indian (ECACC ID:GRC-138) with haplogroup A2*(xA2a,A2b) at different exposure time intervals. Full profiles are indicated by H and X indicate dropout of one or more alleles. Geographic sample designation and ID-Nr. as used by the ECCAC sample panel provider.

Simulated case samples

Twenty-five mock casework samples prepared from blood, semen, saliva, hair, trace DNA, high and low concentrations of haematin inhibited samples and mixtures, were blind-tested to assess the ability of the assays to generate reproducible results from different forensically relevant substrates. As seen in Supplementary Table 2, it was possible to achieve full profiles for the majority of the samples. However, it was not possible to obtain full profiles with high and low concentrations of haematin inhibitor. Addition of bovine serum albumin (BSA) could be used to overcome this problem. It should also be noted that in case of a mixture, as explained earlier, the discerning haplogroups can be resolved only if the contributors belonged to different haplogroups. For example, as seen in the Supplementary Table 2, sample 22 is a mixture of vaginal DNA and semen but the mixture could not be detected by the assays. It can be concluded that the two contributors belonged to the same haplogroup and hence it did not generate two different haplogroups. However, the five assays were effective in producing results across different substrates and highly accurate for a number of casework type scenarios.

Population studies

The performance efficiency and reliability of the multiplexes for predicting the haplogroups and inferring the bio-geographical origin was tested here on the ECACC Ethnic Diversity DNA panel comprising individuals of diverse known biogeographic origin. The final haplogroup results with their corresponding geographic origin for the 92 ECACC individuals as inferred by the combined results from the five multiplex assays are shown in Table 3. The haplogroup results obtained from each of the five assays specifically are presented in Supplementary Table 3. All the haplogroups determined for the ECACC samples were in agreement with their biogeographic origin as provided by the panel distributor (although it must be noted that it was difficult to evaluate the ancestry of the 'Oriental' population, since 'Oriental' does not correspond to a well-defined geographic area). The availability of a commercially available set of samples for which we hereby provide mtDNA haplogroup information obtained with the five multiplex assays will allow other researchers to use them as positive control samples in their future studies. Note, however, that not all haplogroups detectable with our assays were observed in the current sample set from the ECACC panel. Of further relevance for population studies and validation is that previously,

the Multiplexes 1–3 were theoretically tested using the available complete mtDNA sequences using 75 samples from Centre d'Etude du Polymorphisme Humain-Human Genome Diversity Panel (CEPH-HGDP) [1]; and the Multiplexes 4–6 were previously empirically tested using 161 samples from Near Oceania, Remote Oceania and Australia [2].

In order to further stress on the effectiveness of the mtDNA tool for bio-geographic ancestry prediction, we inferred, in silico, the mtDNA haplogroups of the 1000 Genomes Project samples for which the whole mtDNA sequence data are available [31,32]. From the full mtDNA sequences, the alleles of those SNP sites that are included in the five multiplex assays were considered and from the resulting genotypes, the haplogroups were inferred. This information, together with the previously reported haplogroup frequencies in Oceanian individuals [2], can be found in Supplementary Table 4. This table merely serves to illustrate the continental discrimination capacity of our assays and should not be regarded as a representative worldwide reference database from which reliable frequency estimates can be obtained.

Sample ID	Haplogroup Origin (known)	Haplogroup detected	Haplogroup Origin (detected)
591	Japanese Male	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
AKIBA	Japanese Female	R9	E Eurasia
EK-TOK	Japanese Male	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
HID	Japanese Male	M7a	E Eurasia
HS67	Japanese Male	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
KOZ	Japanese Female	B*(xB4a,B2)	E Eurasia/America
KT12	Japanese Male	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
KT14	Japanese Male	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
KT17	Japanese Female	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
KT2	Japanese Female	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
KT3	Japanese Male	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
SA	Japanese Male	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
SM	Japanese Male	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
TAB089	Japanese Male	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
YI	Japanese Male	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
BM14	Italian Male	J	W Eurasia
BM15	Italian Male	H	W Eurasia/N Africa
BM16	Italian Female	T	W Eurasia
BM21	German/Italian Male	H	W Eurasia/N Africa
BM9	Italian Female	H	W Eurasia/N Africa
BRIP	Italian Male	H	W Eurasia/N Africa
CALOGERO	Italian Male	T	W Eurasia
J05282239	Italian Male	U*(xU6,U8b)	W Eurasia/S Asia
SPO010	Italian Female	HV*(xHV0a,H)	W Eurasia
WT100BIS	Italian Male	U*(xU6,U8b)	W Eurasia/S Asia
WT47	Italian Male	U*(xU6,U8b)	W Eurasia/S Asia
HAY,KJ	Australian Aborigine Male	S	Australia
IHL,AD031	Australian Aborigine Male	S	Australia
IHL,AD036	Australian Aborigine Male	M42a	Australia
NON,L	Australian Aborigine Female	P5	Australia
WON,C	Australian Aborigine Male	S	Australia
WON,I	Australian Aborigine Male	R*(x,B,B4a,B4a1,B6,H4,H5a3a1,JT,P,R0,R9,R11,U)	E Eurasia/S Asia/Oceania
WON,M	Australian Aborigine Male	S	Australia

Sample ID	Haplogroup Origin (known)	Haplogroup detected	Haplogroup Origin (detected)
WOO,EM	Australian Aborigine Female	M42*(xM42a)	Australia/Southern Asia
BUR,E	Australian Aborigine Female	M42a	Australia
CHI-007	Thai Male	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
CHI-034	Thai Female	R9	E Eurasia
DCH002	Thai Female	R9	E Eurasia
DCH006	Thai Female	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
DCH007	Thai Male	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
DCH008	Thai Male	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
DCH009	Thai Male	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
DCH010	Thai Male	R9	E Eurasia
DCH011	Thai Female	R9	E Eurasia
DCH012	Thai Male	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
PETCH	Oriental Male	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
HARA	Oriental Unknown	M7a	Eastern Eurasia
HAU,ML	Oriental Male	R9	E Eurasia
HOKKAIDO	Oriental Male	R9	E Eurasia
HOSONUM	Oriental Male	D2a	Beringia
KAWASAKI	Oriental Male	M7a	Eastern Eurasia
KIME	Oriental Male	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
KUROIWA	Oriental Male	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
NP369	Oriental Male	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
WATANABE	Oriental Male	L2 / L3*(xM,N) / L4 / L6	Africa
WHONP192	Oriental Male	N*(xN1,W,A,X,R)	E Eurasia/S Asia/Oceania
WHONP199	Oriental Female	R9	E Eurasia
MP0008	Black African Male	U6	Africa
MP0010	Black African Female	L3*(xM,N)	Africa
MP0014	Black African Female	L2 / L4 / L6	Africa
MP0021	Black African Male	L0*(xL0k,L0d1a,L0d3) / L1 (with independent mutation at 12007)	Africa
FPAF	Ashkenazi Jewish Male	H	W Eurasia/N Africa
IDF	Ashkenazi Jewish Male	U8b	W Eurasia
LWAGS	Ashkenazi Jewish Male	J	W Eurasia
MZ070782	Ashkenazi Jewish Male	H	W Eurasia/N Africa
YAR	Ashkenazi Jewish Male	W	W Eurasia

Sample ID	Haplogroup Origin (known)	Haplogroup detected	Haplogroup Origin (detected)
BON-CT	French Male	H	W Eurasia/N Africa
BONA	French Caucasian Male	H	W Eurasia/N Africa
CF996	French Male	J	W Eurasia
DUCAF	French Male	H	W Eurasia/N Africa
FJO	French Caucasian Male	J	W Eurasia
FORE	French Caucasian Female	H	W Eurasia/N Africa
HO104	French Male	U*(xU6,U8b)	W Eurasia/S Asia
HO301	French Female	I	W Eurasia
15310-LN	French Caucasian Female	H	W Eurasia/N Africa
MARMarg	French Caucasian Female	H	W Eurasia/N Africa
PF04015	French Female	W	W Eurasia
RML	South American Indian Male	C1*(xC1a,C1e)	America
AMALA	South American Indian Female	C1*(xC1a,C1e)	America
SPL	South American Indian Male	C1*(xC1a,C1e)	America
LZL	South American Indian Female	C1*(xC1a,C1e)	America
KRC-110	South American Indian Female	B2	America
GRC-138	South American Indian Male	A2*(xA2a,A2b)	America
GRC-187	South American Indian Male	D1	America
GRC-212	South American Indian Male	B2	America
C0935	Caucasian Male	U*(xU6,U8b)	W Eurasia/S Asia
C0915	Caucasian Male	T	W Eurasia
C0919	Caucasian Female	H	W Eurasia/N Africa
C0125	Caucasian Female	U8b	W Eurasia
C0163	Caucasian Female	U*(xU6,U8b)	W Eurasia/S Asia
C0175	Caucasian Female	H	W Eurasia/N Africa
C0176	Caucasian Female	U*(xU6,U8b)	W Eurasia/S Asia

Table 3: The ECACC Ethnic Diversity DNA panel with sample ID and sample origin as reported by sample provider, mtDNA haplogroups as detected using the 5 mt-SNP multiplex assays, and the likely geographic origin of the detected mtDNA haplogroups.

Conclusions

The five SNaPshot mtDNA multiplex assays comprising a total of 62 ancestry-informative mt-SNPs that together delineate 70 worldwide matrilineal haplogroups have been successfully validated according to the SWGDAM guidelines. The tests performed showed that all five multiplex assays satisfy all the conditions in terms of sensitivity, precision, mixtures, reproducibility, species specificity, stability, simulated case samples and population studies. This developmental validation thus demonstrates the ease, robustness, and reliability of the five mtDNA multiplexes for matrilineal bio-geographic ancestry prediction in forensic (and other) applications. This hierarchical genotyping tool is highly sensitive and it is possible to obtain full profiles with a minimum amount of 1 pg DNA input, it exhibits minimum cross reactivity with non-human DNA, it is successful in producing accurate results from degradation samples and simulated casework samples of different types as well as from population samples, and it produces concordant results between different laboratories with various pre-existing experiences. We would like to emphasize that with mtDNA only the matrilineal ancestry information of an individual can be revealed and that local demographic histories have to be kept in mind when utilizing mtDNA evidence. In order to procure a more comprehensive picture of an individual's overall ancestry, it is recommended to combine mtDNA evidence with evidence from paternally inherited Y chromosomal (in the case of males) and biparentally inherited autosomal DNA markers [1]. Overall, this validated genotyping tool can now be used to establish the matrilineal bio-geographic ancestry of an unknown individual at the continental level for instance in forensic application. As such, this validated tool can provide investigative leads in those criminal cases where conventional STR profiling fails or does not produce a match with a known suspect, or in those missing person identification cases where conventional STR profiling fails or does not produce a match with ante mortem samples nor highlights true biological relatives.

Acknowledgements

This work was funded in part by the Netherlands Forensic Institute (NFI), and received additional support by a grant from the Netherlands Genomic Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands (FGCN).

References

[1] van Oven M, Vermeulen M, Kayser M, Multiplex genotyping system for efficient inference of matrilineal genetic ancestry with continental resolution, Investig. Genet. 2 (March) (2011) 6.

[2] Ballantyne KN, van Oven M, Ralf A, Stoneking M, Mitchell RJ, van Oorschot RA, Kayser M, MtDNA SNP multiplexes for efficient inference of matrilineal genetic ancestry within Oceania, Forensic Sci. Int. Genet. 6 (July (4)) (2012) 425–436.

[3] Schlebusch CM, Naidoo T, Soodyall H, SNaPshot minisequencing to resolve mitochondrial macro-haplogroups found in Africa, Electrophoresis 30 (November (21)) (2009) 3657–3664.

[4] Phillips C, Salas A, Sanchez JJ, Fondevila M, Gomez-Tato A, Alvarez-Dios J, Calaza M, de Cal MC, Ballard D, Lareu MV, Carracedo A, SNPforID Consortium. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, Forensic Sci. Int. Genet. 1 (December (3–4)) (2007) 273–280.

[5] Halder I, Shriver M, Thomas M, Fernandez JR, Frudakis T, A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications, Hum. Mutat. 29 (May (5)) (2008) 648–658.

[6] Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, De La Vega FM, Seldin MF, Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America, Hum. Mutat. 30 (January (1)) (2009) 69–78.

[7] Lao O, Vallone PM, Coble MD, Diegoli TM, van Oven M, van der Gaag KJ, Pijpe J, de Knijff P, Kayser M, Evaluating self-declared ancestry of U.S. Americans with autosomal, Y-chromosomal and mitochondrial DNA, Hum. Mutat. 31 (December (12)) (2010) E1875–E1893.

[8] Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ, Harvesting the fruit of the human mtDNA tree, Trends Genet. 22 (June (6)) (2006) 339–345.

[9] Grignani P, Turchi C, Achilli A, Peloso G, Alu M , Ricci U, Robino C, Pelotti S, Carnevali E, Boschi I, Tagliabracci A, Previdere C , Multiplex mtDNA coding region SNP assays for molecular dissection of haplogroups U/K and J/T, Forensic Sci. Int. Genet. 4 (December (1)) (2009) 21–25.

[10] Brandstatter A, Parsons TJ, Parson W, Rapid screening of mtDNA coding region SNPs for the identification of west European Caucasian haplogroups, Int. J. Legal Med. 117 (October (5)) (2003) 291–298.

[11] Alvarez-Iglesias V, Jaime JC, Carracedo A, Salas A, Coding region mitochondrial DNA SNPs: targeting East Asian and Native American haplogroups, Forensic Sci. Int. Genet. 1 (March (1)) (2007) 44–55.

[12] Scientific Worzzzking Group on DNA Analysis Methods. SWGDAM Validation Guidelines, 2012 Available at: http://swgdam.org/SWGDAM_Validation_Guidelines_APPROVED_Dec_2012.pdf (accessed December 2012).

[13] Børsting C, Rockenbauer E, Morling N, Validation of a single nucleotide polymorphism (SNP) typing assay with 49 SNPs for forensic genetic testing in a laboratory accredited according to the ISO 17025 standard, Forensic Sci. Int. Genet. 4 (December (1)) (2009) 34–42.

[14] Walsh S, Lindenbergh A, Zuniga SB, Sijen T, de Knijff P, Kayser M, Ballantyne KN, Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence, Forensic Sci. Int. Genet. 5 (November (5)) (2011) 464–471.

[15] Vallone PM, Butler JM, AutoDimer: a screening tool for primer-dimer and hairpin structures, Biotechniques 37 (August (2)) (2004) 226–231.

[16] van Oven M, Kayser M, Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation, Hum. Mutat. 30 (February (2)) (2009) E386–E394.

[17] Scholes C, Siddle K, Ducourneau A, Crivellaro F, Jarve M, Rootsi S, Bellatti M, Tabbada K, Mormina M, Reidla M, Villems R, Kivisild T, Lahr MM, Migliano AB, Genetic diversity and evidence for population admixture in Batak Negritos from Palawan, Am. J. Phys. Anthropol. 146 (September (1)) (2011) 62–72.

[18] Duggan AT, Stoneking M, A highly unstable recent mutation in human mtDNA, Am. J. Hum. Genet. 92 (February (2)) (2013) 279–284.

[19] Hagelberg E, Quevedo S, Turbon D, Clegg JB, DNA from ancient Easter Islanders, Nature 369 (May (6475)) (1994) 25–26.

[20] Lum JK, Rickards O, Ching C, Cann RL, Polynesian mitochondrial DNAs reveal three deep maternal lineage clusters, Hum. Biol. 66 (August (4)) (1994) 567–590.

[21] Trejaut JA, Kivisild T, Loo JH, Lee CL, He CL, Hsu CJ, Lee ZY, Lin M, Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations, PLoS Biol. 3 (August (8)) (2005) e247.

[22] Vallone PM, Just RS, Coble MD, Butler JM, Parsons TJ, A multiplex allele specific primer extension assay for forensically informative SNPs distributed throughout the mitochondrial genome, Int. J. Legal Med. 118 (June (3)) (2004) 147–157.

[23] Musgrave-Brown E, Ballard D, Alvarez MF, Fang R, Harrison C, Phillips C, Prasad Y, Rey BS, Thacker C, Wiluhn J, Carracedo A, Schneider PM, Court Syndercombe D, Forensic validation of the Genplex SNP typing system-Results of an interlaboratory study, Forensic Sci. Int. Genet. Suppl. Ser. 1 (August (1)) (2008) 389–393.

[24] Sanchez JJ, Phillips C, Børsting C, Balogh K, Bogus M, Fondevila M, Harrison CD, Musgrave-Brown E, Salas A, Syndercombe-Court D, Schneider PM, Carracedo A, Morling N, A multiplex assay with 52 single nucleotide polymorphisms for human identification, Electrophoresis 27 (May (9)) (2006) 1713–1724.

[25] Tucker VC, Hopwood AJ, Sprecher CJ, McLaren RS, Rabbach DR, Ensenberger MG, Thompson JM, Storts DR, Developmental validation of the PowerPlex(1) ESI 16 and PowerPlex(1) ESI 17 Systems: STR multiplexes for the new European standard, Forensic Sci. Int. Genet. 5 (November (5)) (2011) 436–448.

[26] Green RL, Lagace´ RE , Oldroyd NJ, Hennessy LK, Mulero JJ, Developmental validation of the AmpFlSTR1 NGM SElectTM PCR Amplification Kit: A next generation STR multiplex with the SE33 locus, Forensic Sci. Int. Genet. 7 (January (1)) (2013) 41–51.

[27] Quintans B, Alvarez-Iglesias V, Salas A, Phillips C, Lareu MV, Carracedo A, Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing, Forensic Sci. Int. 140 (March (2–3)) (2004) 251–257.

[28] Guichoux E, Lagache L, Wagner S, Chaumeil P, Leger P, Lepais O, Lepoittevin C, Malausa T, Revardel E, Salin F, Petit RJ, Current trends in microsatellite genotyping, Mol. Ecol. Resour. 11 (July (4)) (2011) 591–611.

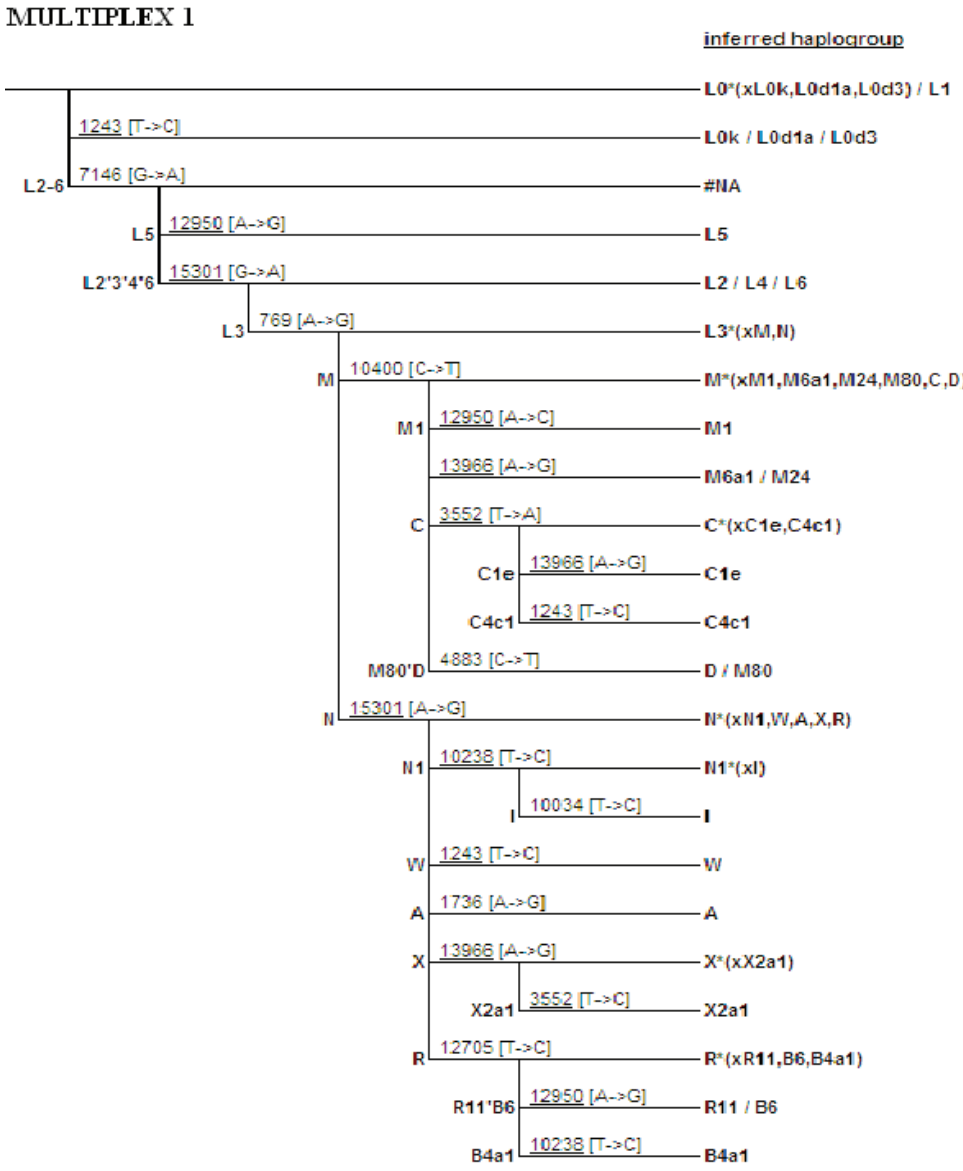
[29] McCord B, Troubleshooting capillary electrophoresis systems, Profiles DNA 6 (September (2)) (2003) 10–12.

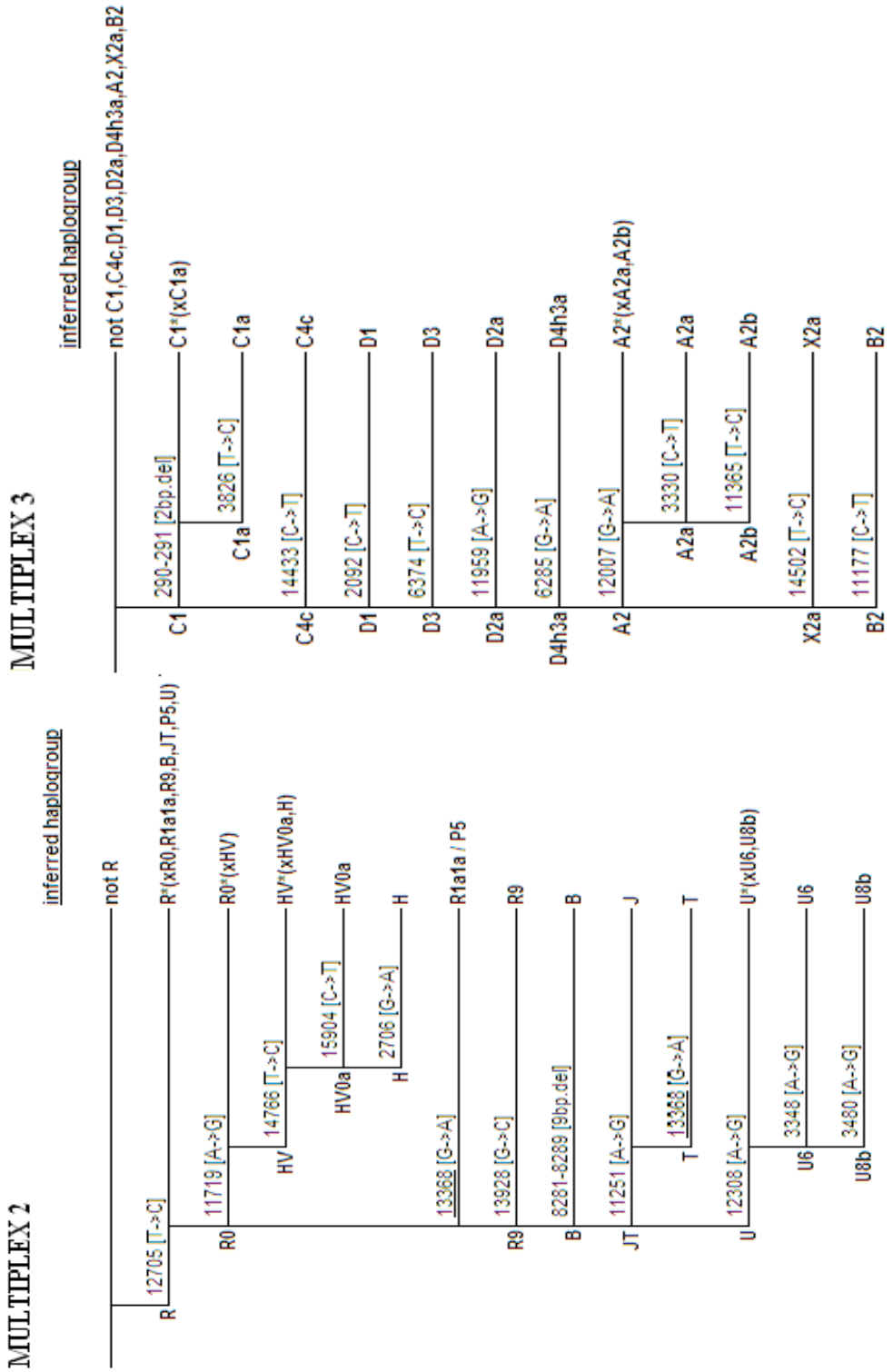
[30] Sinha RP, Hader DP, UV-induced D.N.A. damage and repair: a review, Photochem. Photobiol. Sci. 1 (April (4)) (2002) 225–236.

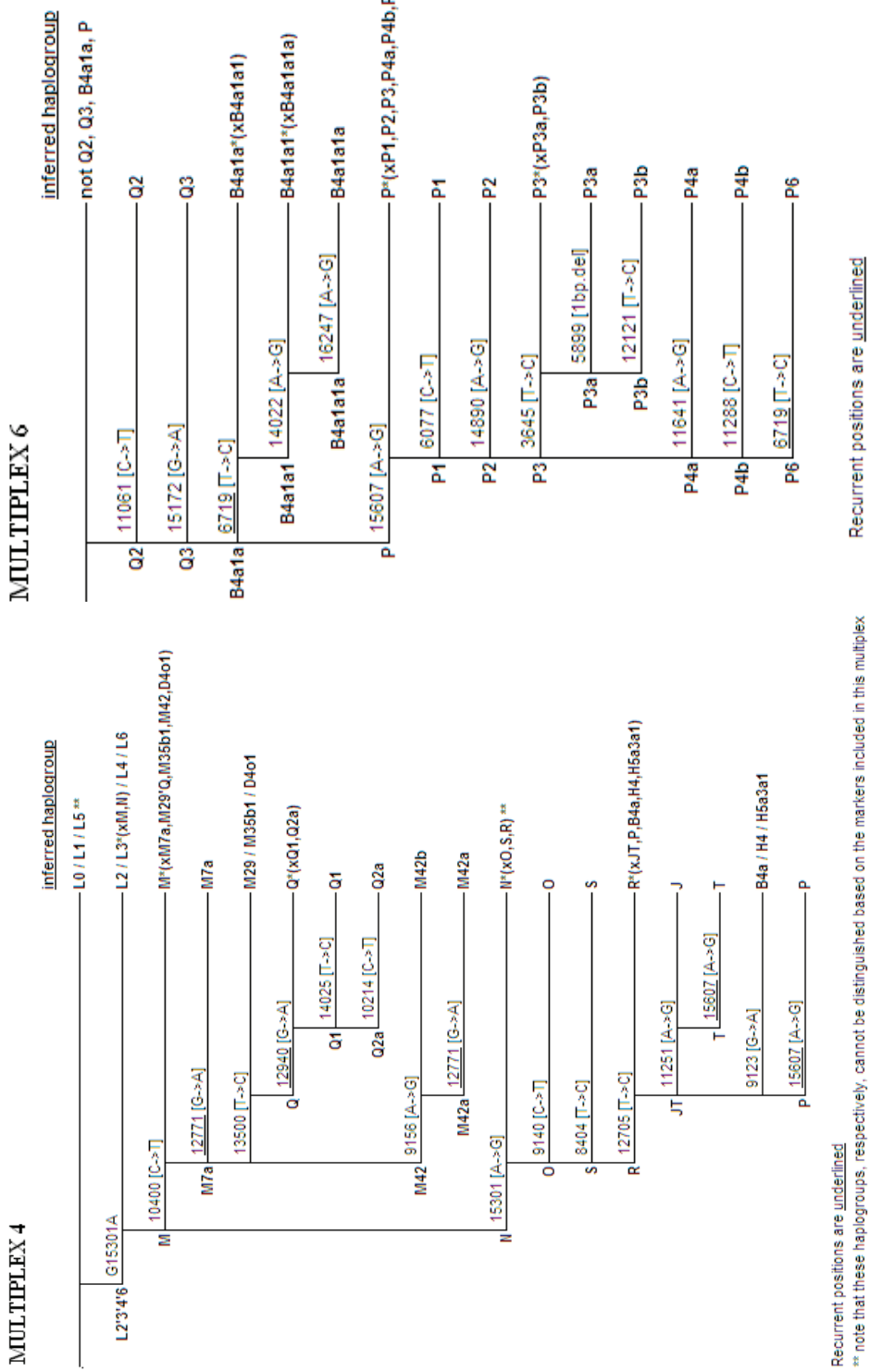
[31] Zheng HX, Yan S, Qin ZD, Wang Y, Tan JZ, Li H, Jin L, Major population expansion of East Asians began before neolithic time: evidence of mtDNA genomes, PLoS ONE 6 (10) (2011) e25835.

[32] Zheng HX, Yan S, Qin ZD, Jin L, MtDNA analysis of global populations support that major population expansions began before Neolithic Time, Sci. Rep. 2 (2012) 745.

Supplementary materials




MULTIPLEX 2
inferred haplogroup
not R
R
12705 [T->C]
R*(xR0,R1a1a,R9,B,JT,P5,U)
R0
11719 [A->G]
R0*(xHV)
HV
14766 [T->C]
HV*(xHV0a,H)
HV0a
15904 [C->T]
HV0a
H
2706 [G->A]
H
R1a1a / P5
13368 [G->A]
R9
13928 [G->C]
R9
B
8281-8289 [9bp.del]
B
JT
11251 [A->G]
J
T
13388 [G->A]
T
U
12308 [A->G]
U*(xU6,U8b)
U6
3348 [A->G]
U6
U8b
3480 [A->G]
U8b
MULTIPLEX 3
inferred haplogroup
not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2
C1
290-291 [2bp.del]
C1*(xC1a)
C1a
3826 [T->C]
C1a
C4c
14433 [C->T]
C4c
D1
2092 [C->T]
D1
D3
6374 [T->C]
D3
D2a
11959 [A->G]
D2a
D4h3a
6285 [G->A]
D4h3a
A2
12007 [G->A]
A2*(xA2a,A2b)
A2a
3330 [C->T]
A2a
A2b
11365 [T->C]
A2b
X2a
14502 [T->C]
X2a
B2
11177 [C->T]
B2



Supplementary Figure 1: Mt-SNP phylogenies and haplogroups of the five multiplex assays.

Sample	Concentration (pg.)	Laboratory 1	Laboratory 2
Blank	-	-	-
Sample 1	980	✓	✓
Sample 2	289	✓	✓
Sample 3 (mix)	590	✓	X
Sample 4	900	✓	✓
Sample 5	673	✓	✓
Sample 6	51	✓	✓
Sample 7	834	✓	✓
Sample 8	76	✓	✓
Sample 9 (mix)	643	✓	X
Sample 10	873	✓	✓
Blank	-	✓	✓
Sample 11	616	✓	✓
Chimp_78ng	78	✓	✓
Sample 12 (mix)	728	✓	X
Sample 13	108	✓	✓
Sample 2	289	✓	✓
Sampe 14	53	✓	✓
Sample 15	289	✓	✓
Sample 1	980	✓	✓
Sample 16	409	✓	✓
Blank	-	✓	✓
Blank	-	✓	✓
Sample 9 (mix)	643	✓	X
Sample 73	485	✓	✓
Sample 4	900	✓	✓
Sample 3 (mix)	590	✓	X
Blank	-	✓	✓
Sample 13	108	✓	✓
Sample 78	568	✓	✓
Sample 18	48	✓	✓
Sample 18	48	✓	✓
Sample 19	116	✓	✓
Blank	-	✓	✓
Sample 17	59	✓	✓
Sample 20	98	✓	✓
Sample 5	673	✓	✓
Sample 21	339	✓	✓
Sample 10	873	✓	✓
Sample 22	417	✓	✓
Blank	-	✓	✓
Sample 6	51	✓	✓
Sample 8	76	✓	✓
Sample 14	53	✓	✓
Sample 12 (mix)	728	✓	✓
Sample 17	59	✓	✓
Blank	-	✓	✓
Blank	-	✓	✓

Supplementary Table 1: Results of concordance testing from two independent laboratories. ✓ indicates the expected results across all the five assays and X indicates incorrect genotyping across one or more assays.

	Simulated Casework Sample	DNA concentration in sample (pg)	Multiplex 1	Multiplex 2	Multiplex 3	Multiplex 4	Multiplex 5	Final Interpretation
1	Fresh Blood	1000	✓	✓	✓	x- no profile	✓	H
2	Fresh Blood	1000	✓	✓	✓	✓	✓	M80
3	Dried Blood	225	✓	✓	✓	✓	✓	J
4	Semen	1000	✓	✓	✓	✓	✓	H
5	Saliva	1000	✓	✓	✓	✓	✓	H
6	Vaginal Swab	550	✓	✓	✓	✓	✓	J
7	Saliva	1000	✓	✓	✓	✓	✓	U*(xU6,U8b)
8	Mixture of semen and blood	630	✓	✓ mixture	✓	✓	✓	H/ U*(xU6,U8b)
9	Saliva	1000	✓	✓	✓	✓	✓	H
10	Mucus	739	x- dropout@snp3552	✓	✓	✓	✓	U8b
11	Dried saliva	950	✓	✓	✓	✓	✓	H
12	Mixture of saliva and blood	550	✓	✓ mixture	✓	✓	✓	H/ U*(xU6,U8b)
13	Hair	858	x- no profile	✓	✓	✓	✓	H
14	Trace DNA	94	✓	✓	✓	✓	✓	U*(xU6,U8b)
15	Trace DNA	127	✓	✓	✓	✓	✓	H
16	Mixture of mucus and saliva	426	✓	✓ mixture	✓	✓ mixture	✓	H/ U*(xU6,U8b)
17	Touched item	238	✓	✓	✓	✓	✓	H
18	Frozen blood (Haematin 2%)	98	x- dropout @snp3552;very low peak height @769(54RFU)	x- partial profile	x- partial profile	x- partial profile	x- dropout @11061	Inhibition present
19	Frozen blood (Haematin 10%)	57	x- partial profile	x- partial profile	x- no profile	x- no profile	x- no profile	Inhibition present
20	Touched item	119	✓	✓	✓	✓	✓	H
21	Saliva	1000	✓	✓	✓	✓	✓	M*(xM1,M6a1,M24,M80,C,D)
22	Mixture of vaginal DNA and semen	768	✓	✓	✓	✓	✓	H
23	Saliva	460	✓	✓	✓	✓	✓	H
24	Vaginal DNA	45	x- no profile	x- dropout @ 12308	x- dropout @ 6374;290-291	✓	✓	H
25	Trace DNA	78	x- no profile	✓	✓	✓	✓	H

Supplementary Table 2: Simulated casework samples prepared in house and tested where only the concentrations were known, the sample type and the simulated casework scenario were not provided.

ECCAC ID	Reported ethnicity	Haplogroup as detected with the respective multiplex: Multiplex 1	Multiplex 2	Multiplex 3	Multiplex 4	Multiplex 6	Final haplogroup	Haplogroup-based inferred matrilineal origin
RML	South American Indian Male	C*(xC1e,C4c1)	not R	C1*(xC1a)	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	C1*(xC1a,C1e)	America
EK-TOK	Japanese Male	M*(xM1,M6a1,M24,M80,C,D)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
BM14	Italian Male	R*(xR11,B6,B4a1)	J	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	J	not Q2, Q3, B4a1a, P	J	W Eurasia
LWAGS	Ashkenazi Jewish Male	R*(xR11,B6,B4a1)	J	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	J	not Q2, Q3, B4a1a, P	J	W Eurasia
CF996	French Male	R*(xR11,B6,B4a1)	J	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	J	not Q2, Q3, B4a1a, P	J	W Eurasia
FJO	French Caucasian Male	R*(xR11,B6,B4a1)	J	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	J	not Q2, Q3, B4a1a, P	J	W Eurasia
AMALA	South American Indian Female	C*(xC1e,C4c1)	not R	C1*(xC1a)	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	C1*(xC1a,C1e)	America
MP0014	Black African Female	L2 / L4 / L6	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	L2 / L3*(xM,N) / L4 / L6	not Q2, Q3, B4a1a, P	L2 / L4 / L6	Africa
MP0010	Black African Female	L3*(xM,N)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	L2 / L3*(xM,N) / L4 / L6	not Q2, Q3, B4a1a, P	L3*(xM,N)	Africa
WATANABE	Oriental Male	M*(xM1,M6a1,M24,M80,C,D)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M7a	not Q2, Q3, B4a1a, P	L2 / L3*(xM,N) / L4 / L6	Africa
SPL	South American Indian Male	C*(xC1e,C4c1)	not R	C1*(xC1a)	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	C1*(xC1a,C1e)	America
LZL	South American Indian Female	C*(xC1e,C4c1)	not R	C1*(xC1a)	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	C1*(xC1a,C1e)	America
GRC-187	South American Indian Male	D / M80	not R	D1	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	D1	America
HOSONUM	Oriental Male	M*(xM1,M6a1,M24,M80,C,D)	not R	D2a	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	D2a	Beringia
591	Japanese Male	D / M80	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
KT2	Japanese Female	D / M80	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
KT3	Japanese Male	D / M80	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
SA	Japanese Male	D / M80	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
SM	Japanese Male	D / M80	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
CHI-007	Thai Male	D / M80	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	inconclusive	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
HS67	Japanese Male	D / M80	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
NP369	Oriental Male	D / M80	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
KT12	Japanese Male	D / M80	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
KT14	Japanese Male	D / M80	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
KT17	Japanese Female	D / M80	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	D*(xD1,D3,D2a,D4h3a,D4o1) / M80	E Eurasia
TAB089	Japanese Male	M*(xM1,M6a1,M24,M80,C,D)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
YI	Japanese Male	M*(xM1,M6a1,M24,M80,C,D)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
DCH006	Thai Female	M*(xM1,M6a1,M24,M80,C,D)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
DCH007	Thai Male	M*(xM1,M6a1,M24,M80,C,D)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
DCH008	Thai Male	M*(xM1,M6a1,M24,M80,C,D)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
DCH009	Thai Male	M*(xM1,M6a1,M24,M80,C,D)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
DCH012	Thai Male	M*(xM1,M6a1,M24,M80,C,D)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
PETCH	Oriental Male	M*(xM1,M6a1,M24,M80,C,D)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
KIME	Oriental Male	M*(xM1,M6a1,M24,M80,C,D)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
KUROIWA	Oriental Male	M*(xM1,M6a1,M24,M80,C,D)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M*(xM7a,M29'Q,M35b1,M42,D4o1)	not Q2, Q3, B4a1a, P	M*(xM1,M6a1,M24,M80,C,D,M7a,M29'Q,M35b1,M42,D4o1)	E Eurasia/S Asia/Oceania
WOO,EM	Australian Aborigine Female	M*(xM1,M6a1,M24,M80,C,D)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M42*(xM42a)	not Q2, Q3, B4a1a, P	M42*(xM42a)	Australia/Southern Asia
IHL,AD036	Australian Aborigine Male	M*(xM1,M6a1,M24,M80,C,D)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M42a	not Q2, Q3, B4a1a, P	M42a	Australia
BUR,E	Australian Aborigine Female	M*(xM1,M6a1,M24,M80,C,D)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M42a	not Q2, Q3, B4a1a, P	M42a	Australia
HID	Japanese Male	M*(xM1,M6a1,M24,M80,C,D)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M7a	not Q2, Q3, B4a1a, P	M7a	E Eurasia
HARA	Oriental Unknown	M*(xM1,M6a1,M24,M80,C,D)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M7a	not Q2, Q3, B4a1a, P	M7a	Eastern Eurasia
KAWASAKI	Oriental Male	M*(xM1,M6a1,M24,M80,C,D)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	M7a	not Q2, Q3, B4a1a, P	M7a	Eastern Eurasia
GRC-138	South American Indian Male	A	not R	A2*(xA2a,A2b)	N*(xO,S,R) / L0 / L1 / L5	not Q2, Q3, B4a1a, P	A2*(xA2a,A2b)	America
MP0021	Black African Male	L0*(xL0k,L0d1a,L0d3) / L1	not R	A2*(xA2a,A2b)	N*(xO,S,R) / L0 / L1 / L5	not Q2, Q3, B4a1a, P	L0*(xL0k,L0d1a,L0d3) / L1(with independent mutation at 12007	Africa

HO301	French Female	I	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	N*(xO,S,R) / L0 / L1/ L5	not Q2, Q3, B4a1a, P	I	W Eurasia
WHONP192	Oriental Male	N*(xN1,W,A,X,R)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	N*(xO,S,R) / L0 / L1/ L5	not Q2, Q3, B4a1a, P	N*(xN1,W,A,X,R)	E Eurasia/S Asia/Oceania
YAR	Ashkenazi Jewish Male	W	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	N*(xO,S,R) / L0 / L1/ L5	not Q2, Q3, B4a1a, P	W	W Eurasia
PF04015	French Female	W	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	N*(xO,S,R) / L0 / L1/ L5	not Q2, Q3, B4a1a, P	W	W Eurasia
NON,L	Australian Aborigine Female	R*(xR11,B6,B4a1)	R1a1a / P5	inconclusive	P	P*(xP1,P2,P3,P4a,P4b,P6) / T	P5	Australia
WT47	Italian Male	R*(xR11,B6,B4a1)	U*(xU6,U8b)	inconclusive	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	U*(xU6,U8b)	W Eurasia/S Asia
KRC-110	South American Indian Female	R*(xR11,B6,B4a1)	B	B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	B2	America
GRC-212	South American Indian Male	R*(xR11,B6,B4a1)	B	B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	B2	America
HAU,ML	Oriental Male	R*(xR11,B6,B4a1)	R9	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	R9	Eurasia
KOZ	Japanese Female	R*(xR11,B6,B4a1)	B	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	B*(xB4a,B2)	E Eurasia/America
BM15	Italian Male	R*(xR11,B6,B4a1)	H	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	H	W Eurasia/N Africa
BM21	German/Italian Male	R*(xR11,B6,B4a1)	H	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	H	W Eurasia/N Africa
BM9	Italian Female	R*(xR11,B6,B4a1)	H	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	H	W Eurasia/N Africa
BRIP	Italian Male	R*(xR11,B6,B4a1)	H	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	H	W Eurasia/N Africa
FPAF	Ashkenazi Jewish Male	R*(xR11,B6,B4a1)	H	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	H	W Eurasia/N Africa
MZ070782	Ashkenazi Jewish Male	R*(xR11,B6,B4a1)	H	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	H	W Eurasia/N Africa
BON-CT	French Male	R*(xR11,B6,B4a1)	H	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	H	W Eurasia/N Africa
BONA	French Caucasian Male	R*(xR11,B6,B4a1)	H	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	H	W Eurasia/N Africa
DUCAF	French Male	R*(xR11,B6,B4a1)	H	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	H	W Eurasia/N Africa
FORE	French Caucasian Female	R*(xR11,B6,B4a1)	H	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	H	W Eurasia/N Africa
15310-LN	French Caucasian Female	R*(xR11,B6,B4a1)	H	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	H	W Eurasia/N Africa
MARMarg	French Caucasian Female	R*(xR11,B6,B4a1)	H	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	H	W Eurasia/N Africa
C0919	Caucasian Female	R*(xR11,B6,B4a1)	H	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	H	W Eurasia/N Africa
C0175	Caucasian Female	R*(xR11,B6,B4a1)	H	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	H	W Eurasia/N Africa
SPO010	Italian Female	R*(xR11,B6,B4a1)	HV*(xHV0a,H)	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	HV*(xHV0a,H)	W Eurasia/N Africa
WON,I	Australian Aborigine Male	R*(xR11,B6,B4a1)	R*(xR0,R9,B,JT,U)	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	R*(x,B,B4a,B4a1,B6,H4,H5a3a1,JT,P,R0,R9,R11,U)	E Eurasia/S Asia/Oceania
AKIBA	Japanese Female	R*(xR11,B6,B4a1)	R9	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	R9	E Eurasia
CHI-034	Thai Female	R*(xR11,B6,B4a1)	R9	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	R9	E Eurasia
DCH002	Thai Female	R*(xR11,B6,B4a1)	R9	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	R9	E Eurasia
DCH010	Thai Male	R*(xR11,B6,B4a1)	R9	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	R9	E Eurasia
DCH011	Thai Female	R*(xR11,B6,B4a1)	R9	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	R9	E Eurasia
HOKKAIDO	Oriental Male	R*(xR11,B6,B4a1)	R9	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	R9	E Eurasia
WHONP199	Oriental Female	R*(xR11,B6,B4a1)	R9	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	R9	E Eurasia
J0528239	Italian Male	R*(xR11,B6,B4a1)	U*(xU6,U8b)	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	U*(xU6,U8b)	W Eurasia/S Asia
WT100BIS	Italian Male	R*(xR11,B6,B4a1)	U*(xU6,U8b)	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	U*(xU6,U8b)	W Eurasia/S Asia
HO104	French Male	R*(xR11,B6,B4a1)	U*(xU6,U8b)	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	U*(xU6,U8b)	W Eurasia/S Asia
C0935	Caucasian Male	R*(xR11,B6,B4a1)	U*(xU6,U8b)	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	U*(xU6,U8b)	W Eurasia/S Asia
C0163	Caucasian Female	R*(xR11,B6,B4a1)	U*(xU6,U8b)	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	U*(xU6,U8b)	W Eurasia/S Asia
C0176	Caucasian Female	R*(xR11,B6,B4a1)	U*(xU6,U8b)	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	U*(xU6,U8b)	W Eurasia/S Asia
MP0008	Black African Male	R*(xR11,B6,B4a1)	U6	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	U6	Africa
IDF	Ashkenazi Jewish Male	R*(xR11,B6,B4a1)	U8b	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	U8b	W Eurasia
C0125	Caucasian Female	R*(xR11,B6,B4a1)	U8b	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	R*(xJT,P,B4a,H4,H5a3a1)	not Q2, Q3, B4a1a, P	U8b	W Eurasia
HAY,KJ	Australian Aborigine Male	N*(xN1,W,A,X,R)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	S	not Q2, Q3, B4a1a, P	S	Australia
IHL,AD031	Australian Aborigine Male	N*(xN1,W,A,X,R)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	S	not Q2, Q3, B4a1a, P	S	Australia
WON,C	Australian Aborigine Male	N*(xN1,W,A,X,R)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	S	not Q2, Q3, B4a1a, P	S	Australia
WON,M	Australian Aborigine Male	N*(xN1,W,A,X,R)	not R	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	S	not Q2, Q3, B4a1a, P	S	Australia
BM16	Italian Female	R*(xR11,B6,B4a1)	T	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	T	P*(xP1,P2,P3,P4a,P4b,P6) / T	T	W Eurasia
CALOGERO	Italian Male	R*(xR11,B6,B4a1)	T	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	T	P*(xP1,P2,P3,P4a,P4b,P6) / T	T	W Eurasia
C0915	Caucasian Male	R*(xR11,B6,B4a1)	T	not C1,C4c,D1,D3,D2a,D4h3a,A2,X2a,B2	T	not Q2, Q3, B4a1a, P	T	W Eurasia

Supplementary Table 3: Haplogroups as detected by the respective multiplex for all the 92 individuals in the ECCAC DNA Panel.

	African ancestry*			European ancestry*						East Asian ancestry*				Mixed (largely Native American) ancestry*		Oceanian ancestry**			
	ASW US African Americans	YRI Yorubans (Nigeria)	LWK Luhya (Kenya)	CEU US European Americans	FIN Finnish	GBR British	IBS Spanish	TSI Toscani (Italy)	CHD US Chinese Americans	CHB Han Chinese (Beijing)	CHS Southern Han Chinese	JPT Japanese	MXL US Mexican Americans	CLM Colombians	PUR Puerto Ricans	Australian Aboriginals	Papua New Guinea Highlands	Papua New Guinea Coast	Fiji
MtDNA multiplex-based haplogroup classification	n=61	n=136	n=116	n=103	n=97	n=94	n=14	n=105	n=73	n=121	n=55	n=118	n=67	n=62	n=55	n=93	n=30	n=21	n=17
L0k / L0d1a / L0d3																			
L0*(xL0k,L0d1a,L0d3) / L1	20	27	27					1						2	5				
L5			11																
L2 / L4 / L6	14	48	26												1				
L3*(xM,N)	22	60	52											3	3				
M*(xM1,M6a1,M7a,M24,M29,M35b1,M42,M80,C,D,Q)	1								21	30	13	25				2	1	1	
M1														1					
M6a1 / M24																			
M7a												8							
M29 / M35b1																			
Q*(xQ1,Q2,Q3)																			
Q1																	6	1	
Q2*(xQ2a)																			
Q2a																		1	
Q3																	2		
M42a																2			
M42*(xM42a)																4			
C*(xC1,C4c)									5	4	1								
C1*(xC1a,C1e)	1												10	2	13				
C1a																			
C1e																			
C4c*(xC4c1)																			
C4c1																			
D*(xD1,D3,D2a,D4h3a,D4o1) / M80								1	13	26	9	42							
D1	1												9	1					
D3																			
D2a																			
D4h3a														1					
D4o1																			
N*(xN1,A,O,R,S,W,X)									1	8	7	10				3			
N1*(xI)																			
I				1	2	3													

W				5	2	2		4					1						
A*(xA2)									7	9	4	8							
A2*(xA2a,A2b)	1												25	24	20				
A2a																			
A2b																			
X*(xX2a)					1	5		1											
X2a*(xX2a1)																			
X2a1																			
O																6			
S																29			
R*(xR0,R1a1a,R9,R11,B,J,P,T,U)																2			
R0*(xHV)																			
HV*(xHV0a,H)				4	1		1	4		1									
HV0a (incl V)				3	4		1	2					1			1			
H*(xH4,H5a3a1)				49	34	39	7	49					6	1	2	1			
H4 / H5a3a1				2	1	2		1											
R1a1a																			
J				8	8	10	1	8							3				
T				10	3	9	2	12											
R9 (incl F)									11	23	7	9							
R11									1	2	3								
B*(xB4a,B2,B6)									11	12	10	13							
B4a*(xB4a1)									1	4	1	1							
B4a1*(B4a1a)										1		2							
B4a1a*(xB4a1a1)																			
B4a1a1*(xB4a1a1a)																		3	1
B4a1a1a																		7	13
B2									1				13	27	4				
B6																			
P*(xP1,P2,P3,P4a,P4b,P5,P6)																28			
P1																	17	3	1
P2																	4		
P3*(xP3a,P3b)																1			
P3a																1			
P3b																4			
P4a																			
P4b																3			
P5																			
P6																5			
U*(xU6,U8b)				18	35	19	1	12	1				1		3				
U6	1	1											1						
U8b (incl K)				3	6	5	1	10		1					1	1			

* 1000Genomes data from Zheng et al. 2011 [31] and Zheng et al. 2012 [32]

** data from Ballantyne et al. 2011 [2]

Supplementary Table 4: mtDNA haplogroups and their geographic distribution observed in the 1000 Genome Project data [31,32] and in the 161 Oceanian individuals [2].

Chapter 5.2

High-quality mtDNA control region sequences from 680 individuals sampled across the Netherlands to establish a national forensic mtDNA reference database

Forensic Science International: Genetics 21 (2016) 158 – 167

Abstract

The use of mitochondrial DNA (mtDNA) for maternal lineage identification often marks the last resort when investigating forensic and missing-person cases involving highly degraded biological materials. As with all comparative DNA testing, a match between evidence and reference sample requires a statistical interpretation, for which high-quality mtDNA population frequency data are crucial. Here, we determined, under high quality standards, the complete mtDNA control-region sequences of 680 individuals from across the Netherlands sampled at 54 sites, covering the entire country with 10 geographic sub-regions. The complete mtDNA control region (nucleotide positions 16024–16569 and 1–576) was amplified with two PCR primers and sequenced with ten different sequencing primers using the EMPOP protocol. Haplotype diversity of the entire sample set was very high at 99.63% and, accordingly, the random-match probability was 0.37%. No population substructure within the Netherlands was detected with our dataset. Phylogenetic analyses were performed to determine mtDNA haplogroups. Inclusion of these high-quality data in the EMPOP database (accession number: EMP00666) will improve its overall data content and geographic coverage in the interest of all EMPOP users worldwide. Moreover, this dataset will serve as (the start of) a national reference database for mtDNA applications in forensic and missing person casework in the Netherlands.

Introduction

The eminence of the circular, small-size and highly polymorphic mitochondrial DNA (mtDNA) in the field of forensics is evident, especially in situations where nuclear DNA analysis is technically impossible. The riveting features of mtDNA, i.e., its high cellular copy number, stability and haploidity have made it a pivotal tool in crime-scene investigations, phylogenetics, population genetics and kinship analyses, especially from ancient and/or degraded biological samples which are devoid of intact nuclear DNA. The overall higher effective mutation rate of mtDNA compared to nuclear DNA provides it with higher sequence variability, which has further bolstered its forensic value. The high sequence variability is especially present in three hypervariable segments (HVS) within the non-coding control region or displacement loop (D-loop) [1–3]. With the augmenting significance of mtDNA in forensics, reliable mtDNA databases are crucial for population frequency estimations of mtDNA sequences observed in forensic casework, given an exact match between a crime-scene sample and an individual [4,5]. Knowledge about the rarity of an mtDNA profile is always valuable, especially when the mtDNA profile from an evidence sample and the reference sample cannot be excluded as originating from the same individual.

The collaborative effort of several DNA laboratories under the lead of the DNA laboratory of the Institute of Legal Medicine (GMI) in Innsbruck, Austria, established an easily accessible high-quality mtDNA population database: the European DNA Profiling Group (EDNAP) MtDNA population database, in short EMPOP (http://www.empop.org), for frequency estimations in routine forensic and missing person casework [6], which has grown substantially since its start. However, typically only relatively small sample sets per country are included in EMPOP, limiting the reliability of the obtained frequency estimates and thus their applications in forensic and missing-person cases. In the present study, the complete mtDNA control-region was sequenced under EMPOP's high-quality standards, following the ISFG recommendations [7], in a total of 680 randomly selected Dutch individuals sampled at 54 sites across the Netherlands (Figure1) within ten geographic sub-regions (Table 1).

The generated mtDNA population data are incorporated into the EMPOP database under accession number EMP00666 for worldwide use, and in addition to their usability within EMPOP, will particularly serve as (start of a) national reference database for mtDNA applications in forensic and missing person casework in the Netherlands.

Region	N
Leiden	57
cluster SW: Bergen op Zoom, Hulst, Middelburg, Oostburg, Zierikzee.	84
cluster SE: Heerlen, Maastricht, Roermond, Sittard, Weert.	59
cluster CS: Druten, Oss, Veghel, Waalwijk, Zaltbommel.	94
cluster CNE: Genemuiden, Hoogeveen, Kampen, Meppel, Steenwijk.	79
cluster NW: Heemskerk, Heerhugowaard, Hoorn, Purmerend, Schagen.	54
cluster C: Barneveld, Hilversum, Maarsenbroek, Mijdrecht, Woerden.	76
cluster N: Dokkum, Drachten, Harlingen, Kollum, Leeuwarden, islands (Hollum, Midsland, Nes, Ameland)	81
cluster NE: Assen, Borger, Ter Apel, Veendam, Winschoten + Beilen, Delfzijl, Eelde, Emmen, Hoogezand.	49
cluster CE: Denekamp, Haaksbergen, Losser, Markelo, Tubbergen.	47
Total	**680**

Table 1: The 10 sampling regions and the total number of samples sequenced (N).

Figure 1. Map of Netherlands depicting the 54 sampling sites (filled red circles) within 10 geographic regions (circled black).

Materials and methods

Samples

Sample collection was previously described by Lao et al. [8] as the 680 samples used in this study are a subset of the 999 samples used by Lao et al. [8]. In short, all the samples were from healthy, male blood-donor volunteers at Sanquin, who have given their informed consent. To enable a more-or-less even coverage of the entire Netherlands, samples of known unrelated donors from specific towns and cities of Netherlands were collected. All the samples were received anonymously and the only information that was available was the place of residence.

DNA extraction, amplification and sequencing

DNA was extracted from blood samples, obtained from 680 randomly selected individuals sampled at 54 sites across the Netherlands, using the QIAamp DNA Mini kit (Qiagen, Hagen, Germany) according to the manufacturer's guidelines. The entire mtDNA control region was amplified and sequenced according to EMPOP recommendations [6,7]. PCR amplification of the mtDNA control region was performed in a total reaction volume of 25 µL using the primers F15851 and R639 (Table 2) in a MicroAmp Optical 96-well reaction plate (Thermo Fisher Scientific Inc., Foster City, CA, USA). Each PCR reaction comprised 1X Advantage1 HD buffer (Clontech, California, USA), 2.5 mg/mL bovine serum albumin (BioLabs, New England), 10 mM of dNTPs (Roche, Mannheim, Germany), 10 mM of each primer (F15851 and R639), 5 U/mL Advantage1 HD polymerase and 10 ng of DNA template. All the amplifications were performed on a GeneAmp PCR System 9700 thermocycler (Thermo Fisher Scientific) with the following thermocycling conditions: 95ºC for 2 minutes; 35 cycles of 95ºC for 15 seconds, 56ºC for 30 seconds; 72ºC for 90 seconds, and a final extension step at 72ºC for 10 minutes. The PCR products were purified by adding 2 µL of ExoSAP-IT (USB Corporation, OH, USA) to 5 µL of PCR product and incubating at 37ºC for 45 minutes, followed by enzyme deactivation at 80ºC for 15 minutes. The entire mtDNA control region was sequenced with ten internal primers—six forward primers and four reverse primers (Table 2) to obtain coverage from both strands. The cycle sequencing reaction was carried out in a final volume of 10 µL, containing 2 µL BigDye® Terminator Ready Reaction Mix v1.1 (Thermo Fisher Scientific),

2 µL BigDye Terminator Sequencing Buffer v1.1 (Thermo Fisher Scientific), 0.5 µL of 10 mM primer and 2 µL of the purified PCR product. Cycle sequencing was performed on a GeneAmp PCR System 9700 thermocycler (Thermo Fisher Scientific) with the following thermocycling conditions: 96ºC for 1 min; 25 cycles of 96ºC for 10 seconds, 50ºC for 5 seconds; and finally 60ºC for 4 minutes. The sequencing reaction products were further purified using Sephadex G-50 superfine (GE Healthcare, Uppsala, Sweden) and Multiscreen 96-well filter plates (Millipore, Ireland) according to the manufacturer's protocol. The purified products were electrophoretically separated and detected on a 3130xl Genetic Analyzer (Thermo Fisher Scientific) with POP-7 polymer on a 36-cm capillary array.

Amplification Primers	
F15851	5' ATCTCCCTAATTGAAAACAAAATACTCAAA 3'
R639	5' GGGTGATGTGAGCCCGTCTA 3'
Sequencing Primers	
F16268	5' CACTAGGATACCAACAAACC 3'
F15971	5' TTAACTCCACCATTAGCACC 3'
F15851	5' ATCTCCCTAATTGAAAACAAAATACTCAAA 3'
F314	5' CCGCTTCTGGCCACAGCACT 3'
F29	5' CTCACGGGAGCTCTCCATGC 3'
F15	5' CACCCTATTAACCACTCACG 3'
R16	5' TGATAGACCTGTGATCCATCGTGA 3'
R159	5' AAATAATAGGATGAGGCAGGAATC 3'
R484	5' TGAGATTAGTAGTATGGGAG 3'
R639	5' GGGTGATGTGAGCCCGTCTA 3'
Additional Sequencing Primers	
F16406	5' TCCTCCGTGAAATCAATA 3'
F361	5' ACAAAGAACCCTAACACCAGC 3'
R460	5' AGGGGAAAATAATGTGTTAGT 3'
F74	5' TGCACGCGATAGCATTG 3'
R619	5' AACATTTTCAGTGTATTGCTTTGA 3'

Table 2: Primer sequences used for the amplification and sequencing of the mtDNA control region.

Data analysis and haplogroup assignment

The sequence reads obtained were aligned to the revised Cambridge Reference Sequence (rCRS) [9] and evaluated using the DNA sequence analysis software Sequencher v5.1 (Gene Codes Corporation, Ann Arbor, MI, USA). The sequence quality was manually evaluated, the nomenclature used was

as per the recommended guidelines and variants scored accordingly [10,11]. The haplogroups were assigned using the software HaploGrep [12] and for some samples manually, using PhyloTree Build 16 [13].

Random-match probability and pairwise comparisons

Within the Dutch sample set, the random match probability, which is defined as the probability that an individual chosen at random from an appropriate population will have the same profile as the evidence sample, was calculated as the sum of the squares of the observed haplotype frequencies with a correction for sampling effects [14]. Molecular diversity indices, haplotype frequencies, analysis of molecular variance (AMOVA) and pairwise F_{ST} values were calculated using theArlequin software v3.5.1.2 [15]. In addition, the Dutch data was compared with thirteen other worldwide populations for which entire control-region data could be obtained from the literature: 273 control-region sequences from Austria [5], 200 from Macedonia [16], 129 from Angola [17], 191 from Ghana [18], 249 from Dubai [19], 173 from Hungary [20], 509 from Morocco [21], 182 from Iraq [22], 190 from Somalia [23], 277 from Egypt [24], 187 from Vietnam [25], 377 from Hong Kong [26] and 381 from Kuwait [27]. Pairwise F_{ST} values and AMOVA were computed with Arlequin and multidimensional scaling (MDS) analysis was performed using IBM SPSS Statistics 21.0 to evaluate the genetic distances between populations.

Results and discussion

The numerous appreciative features of mtDNA have made it an indispensable tool for forensic identification and evolutionary studies. However, in comparison to autosomal markers, larger mtDNA datasets are needed to evaluate the extent of variation due to the lack of recombination of mtDNA [28] likewise to Ychromosome markers [29]. As an attempt towards starting a national reference database for mtDNA applications in forensic and missing person casework in the Netherlands, we established a high-quality mtDNA control region dataset of 680 individuals, sampled from 10 geographic regions across the Netherlands (Figure 1), using the Sanger sequencing technique following the EMPOP protocol and make the data available via EMPOP.

For almost all the samples (92.9%), the ten forward and reverse sequencing primers, as listed in Table 2, generated full doublestrand coverage with overlapping sequences across

the entire mtDNA control region. However, for 48 samples (7.1%) ambiguous nucleotide sequences were observed due to the homopolymeric C stretch flanking nucleotide position 16189 and in such cases additional primers F16406, F361, F74, R460 and R619 were used accordingly and then overlapping sequences across the control region were obtained.

Heteroplasmic positions

A total of 49 point heteroplasmies was observed in 45 of the 680 samples (6.61%), of which three were transversions and forty-six were transitions (Supplementary Table 2). Both sequence or point and length heteroplasmy require individual consideration due to the differences in the frequency of occurrence and cause. Most commonly observed sequence or point heteroplasmy have been positions 16093 and 16129 in HVI and 73, 152 and 189 in HVII [4,30,31]. Length heteroplasmy is commonly observed in three regions: in HVI, it begins at position 16184 and when there is a transition at 16189 (T to C), it creates a long homopolymeric stretch making it difficult to determine the precise number of cytosine residues present. The second length heteroplasmy is observed in HVII between positions 303–315, and with a transition at position 310, it becomes difficult to determine the exact number of cytosine residues. The third region of length heteroplasmy is observed in HVIII between positions 576–573 [30]. In most cases, length heteroplasmy is not considered for interpretations.

The following heteroplasmies were observed more than once in this entire dataset: 204Y (5 times), 16093Y (4 times), 16129R (3 times), 146Y (2 times), 152Y (2 times), 16092Y (2 times), 16189Y (2 times), 16218Y (2 times) 16311Y (2 times) and 16362Y (2 times). Some of the heteroplasmic positions observed in this study, such as at 16093, 16129 and 152, are among the most commonly observed heteroplasmic positions [4,30,31]. Additionally, the most frequently noted length heteroplasmy at positions 16184 (HVI); between positions 303–315 (HVII) and between positions 568–574 (HVIII) [30]—have also been observed in this study.

When analysing mtDNA in forensic case work, both the reference and evidence sequences are compared to see if there is a complete match or not. Several publications have provided uniform interpretation guidelines for assessing heteroplasmy in forensic applications [4,30,32]. In brief, when a heteroplasmy is observed, very careful analysis and comparison between the evidence sample and multiple reference samples (also from

different tissues of the same individual) is done. When both the evidence sample and the reference samples originating from the same tissue as the evidence sample comes from have identical heteroplasmic position, and are also identical in all other positions analysed being heteroplasmic or not, then a match is reported. A confidence interval is calculated depending on the size of the database and number of times a particular sequence is observed in the database. As Sanger sequencing has limitations in detecting heteroplasmic alleles <15% [33,34], the use of high-coverage massively parallel sequencing (MPS) overcoming such limitations [34] is expected to be the method of choice for future mtDNA usage in forensic genetics [35].

Haplotypes and haplogroup assignment

The observed variants (relative to the rCRS) from the sequencing of the control region, together with their respective haplogroups determined by HaploGrep and manual inference, are reported in Supplementary Table 1. The haplogroups were initially assigned using the software HaploGrep [12], which uses the most recent version of PhyloTree (Build 16) [13]. For every sample, HaploGrep provides a detailed explanation of how and why a haplogroup was given the highest rank [12]. To explain it in brief, a phylogenetic weight is calculated for every mutation. For each mutation a so called "fluctuation rate" is calculated—depending on how often a mutation occurs in PhyloTree. If a mutation only occurs for 1 haplogroup, this has a higher impact than such mutations occurring more often. If the scores of 2 or more haplogroups are identical, all of them appear as best hits—if a common shared haplogroup exists, this one is the resulting one: HG Quality gives an information if all haplogroup defining mutations were found in the sample, and the sample quality gives an information on how many remaining mutations could not be associated with the haplogroup. The overall score is the mean of those 2 values: haplogroup quality and sample quality. Additionally, a tree structure is provided for every derived haplogroup, showing the current position in the PhyloTree. Taking into account all these factors, we infer a likely haplogroup for a given sample.

However, due to HaploGrep's suboptimal performance with partial mtDNA sequences (i.e., the lack of coding sequence information which would help in assigning the correct haplogroup), some questionable haplogroup predictions were obtained; e.g., all the sequences with the rCRS variants 16519C, 263G, 315.1C were reported as haplogroup

H2a2a. Furthermore, sample Don 1022 with rCRS variants 16140C, 16519C, 263G, 315.1C was predicted as B5 by HaploGrep, but given the absence of variants 73G and 16189C (which are expected with haplogroup B5) we preferred to assign it as HV*. Therefore, the haplogroup assignments were re-evaluated by manual inference using PhyloTree Build 16 [13], to correct and improve the HaploGrep predictions.

A total of 509 different haplotypes were identified, the most common of which was shared among 17 individuals from different sampling sites (4 samples from cluster 1; 6 samples from cluster 4, 2 samples from cluster 3 and 1 sample each from clusters 5, 6, 7, 8 and 9). Using Build 16 of PhyloTree [13], the samples were assigned to 57 different haplogroups/paragroups; for 14 samples, however, the haplogroup could not be confidently determined from the control-region data currently available. The most common haplogroup was HV* (22.9%), followed by haplogroups J1c (7.1%) and K1a (5.7%), all of which were almost equally distributed across the country (Table 3a). At a broader haplogroup level, the most common was HV (45.3%), followed by U (25.6%) and T (11.6%) (Table 3b and Figure 2). In addition, Supplementary Figure 1 illustrates the frequency distribution of each of the haplogroups in the 10 geographic sampling regions. It is noticeable that the mtDNA haplogroup composition of our Dutch population sample set was overall similar to that of other European populations at a broader level of resolution [36,37]. The frequency distribution of the major mtDNA haplogroups in the Dutch population as obtained by us is very similar to that observed in the Genome of the Netherlands (GoNL) Project [38]. However, it is to be noted that five individuals with haplogroups of clearly non-European origin (such as 1 individual each belonging to haplogroups B5, L2, L3, M3 and M7) were present in our Dutch dataset too. These samples, although most likely reflecting relatively recent immigration events from outside Europe, were not eliminated and were considered for all analyses, as they represent people currently living in the Netherlands. Although the sampling design applied [8] avoided the large cities with relatively bigger expected immigrant populations, genealogical background of the donors was not considered when collecting the samples, which explains these findings.

(3a)

Haplogroup	Number	Percent	Haplogroup	Number	Percent
HV*	156	22.9	H3v	2	0.3
J1c	48	7.1	H3z	2	0.3
K1a	39	5.7	H74	2	0.3
T2b	32	4.7	I1	2	0.3
U5a1	30	4.4	I2	2	0.3
K2a	27	4.0	J	2	0.3
H5	26	3.8	J2b	2	0.3
T1a	24	3.5	K1a4a1a2b	2	0.3
H1a	23	3.4	K1b	2	0.3
HV0	23	3.4	T2c	2	0.3
U5b	22	3.2	U	2	0.3
U4	18	2.6	B5a	1	0.1
H1c	17	2.5	H1ap1	1	0.1
J2a	16	2.4	H1b1f	1	0.1
H11a	15	2.2	H1ba	1	0.1
undetermined	14	2.1	H24	1	0.1
W	13	1.9	H39a	1	0.1
H6	9	1.3	H5'36*	1	0.1
H2a1	8	1.2	H7a1	1	0.1
U5a2	8	1.2	H82	1	0.1
I3	7	1.0	HV9a	1	0.1
T2*	7	1.0	K*	1	0.1
T2e	7	1.0	L2c2	1	0.1
H27	5	0.7	L3f1b	1	0.1
J1b	5	0.7	M3	1	0.1
U3a	5	0.7	M7c1	1	0.1
K1c	4	0.6	N1b1	1	0.1
K1d	4	0.6	T	1	0.1
U2e	4	0.6	T2b4b	1	0.1
U8a	4	0.6	T2f1	1	0.1
H1b	3	0.4	T2i	1	0.1
H2a2a1	3	0.4	U5b1e	1	0.1
H2a2b	3	0.4	U7	1	0.1
T2a1b	3	0.4	X2c	1	0.1
H36	2	0.3	X2c1c	1	0.1

(3b)

Broad Haplogroup	Number	Percent
HV	308	45.3
U	174	25.6
T	79	11.6
J	73	10.7
undetermined	14	2.1
W	13	1.9
N1	12	1.8
X	2	0.3
B5	1	0.1
L2c	1	0.1
L3f	1	0.1
M3	1	0.1
M7c	1	0.1

Table 3: (a) Haplogroup frequencies and (b) broad haplogroup frequencies of 680 samples from Netherlands based on the full mtDNA control-region sequences (16,024–16,569, 1–576); haplogroup nomenclature according to PhyloTree build 16 [12].

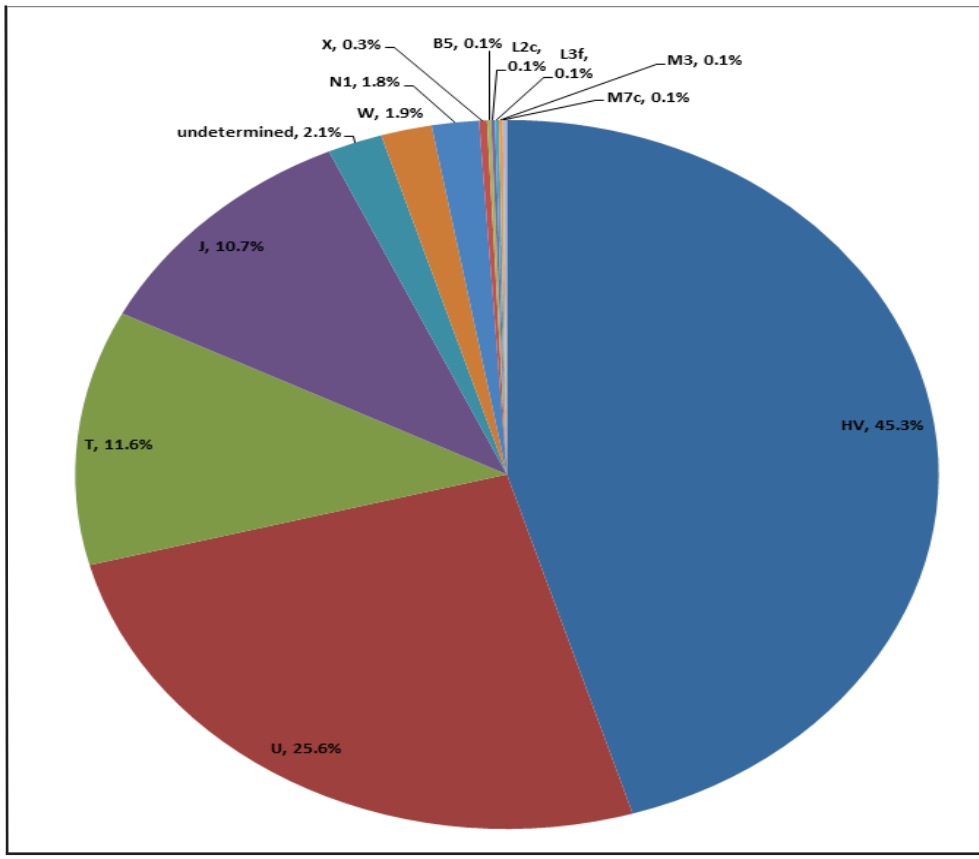


Figure 2: Frequency distribution of the major mtDNA haplogroups in the Dutch population.

Random-match probability, pairwise comparisons, and population substructure

The probability that two individuals selected at random from this Dutch dataset have an identical mtDNA control-region profile was calculated as 0.37%. Accordingly, the Dutch population has a high power of discrimination (99.63%), which in practice will be likely higher given the sampling strategy to avoid large cities applied here, which is a favorable circumstance in forensic genetic investigations. The diversity parameters calculated using the Arlequin software v3.5.1.2 [15] is given in Table 4. The comparison of pairwise F_{ST} distances and their p-values (Table 5a and b) revealed that there is no statistically significant difference in mtDNA distributions between the 10 different geographic regions of the Netherlands sampled. The two-dimensional MDS analysis was performed based on the pairwise differences (Figure 3). Analysis of Molecular Variance (AMOVA) was performed to test for significant variation in the mtDNA distributions among the various regions. Of the variance observed among the 10 groups, 99.814% is attributable to differences within populations and 0.19% represents differences among populations (Table 6a) indicating absence of population substructure based on these mtDNA data in line with the other analyses.

	Parameters				
	Number of Samples	Number of Haplotypes	Gene Diversity	Nucleotide Diversity (average over loci)	Average Pairwise differences
Leiden	57	47	0.9825 +/- 0.0113	0.008330 +/- 0.004306	9.371554 +/- 4.367347
SW	84	73	0.9965 +/- 0.0024	0.007573 +/- 0.003919	8.512048 +/- 3.975383
SE	59	55	0.9978 +/- 0.0035	0.008169 +/- 0.004227	9.174167 +/- 4.279541
CS	94	77	0.9978 +/- 0.0035	0.009062 +/- 0.004625	10.204072 +/- 4.701072
CNE	79	70	0.9947 +/- 0.0033	0.009132 +/- 0.004668	10.292113 +/- 4.747066
NW	54	47	0.9953 +/- 0.0047	0.008243 +/- 0.004269	9.264850 +/- 4.324621
C	76	66	0.9957 +/- 0.0031	0.009082 +/- 0.004647	10.217193 +/- 4.716687
N	81	67	0.9941 +/- 0.0032	0.007755 +/- 0.004008	8.716975 +/- 4.065419
NE	49	48	0.9949 +/- 0.0075	0.007780 +/- 0.004054	8.736395 +/- 4.101669
CE	47	42	0.9943 +/- 0.0090	0.007954 +/- 0.004142	8.924144 +/- 4.186458

Table 4: Molecular diversity indices estimated from the analysis of the entire mtDNA control region in a population sample from the Netherlands.

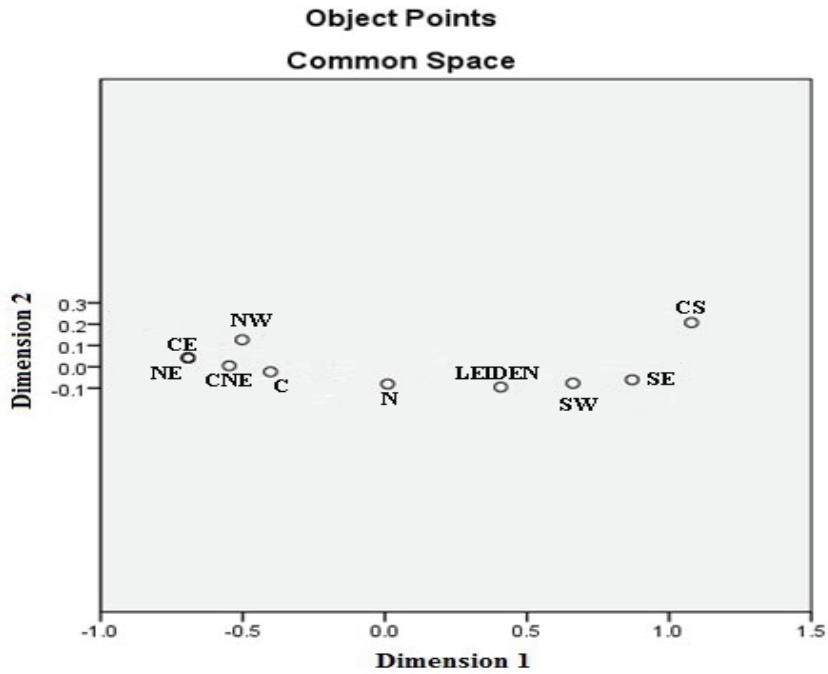


Figure 3: Multi-dimensional scaling plot illustrating the inter-population substructure between the 10 geographic sampling regions from this study.

Due to the highly variable Dutch geographic landscape e.g., sea level fluctuations, extensive peat areas changing over time, the human demographic history of the Netherlands population is complex as the geographic areas accessible to human occupation were dramatically changing over time. Lao et al. [8] and subsequently the Genome of the Netherlands (GoNL) consortium [38] using large numbers of autosomal SNPs found genetic evidence for a subtle population substructure within the Netherlands. In particular, Lao et al. [8] elucidated a genomic gradient, allowing differentiation of a north-eastern, centralwestern, central-northern and a southern ancestry component. Such a substructure is not reflected in our mtDNA data. One reason for such a difference could be relatively high migration rate of females as compared to males leading to higher mtDNA gene flow, which prevents substantial geographic mtDNA differentiation to develop over time [38]. Another reason may be in the genetic resolution applied. For instance, the Dutch population substructure from autosomal data [8,38] was only detectable by using large numbers of SNPs. The genetic resolution applied with mtDNA studies based on parts of the mt genome such as the hypervariable

region studies here is low and future studies based on whole mt genome will shed more light on this matter. However, what matters for the use of the mtDNA information presented here for the Netherlands is that because no population substructure is detected with these mtDNA markers, population substructure effects do not need to be taken into account when interpreting forensic mtDNA evidence in the Netherlands as long as it comes from Dutch European sample donors.

The pairwise F_{ST} distances among fourteen worldwide populations, including the Dutch population sample generated here, are given in Table 5c and d. The F_{ST} values revealed that there are significant differences in the mtDNA distributions between the Netherlands and the thirteen other populations compared here. The average pairwise difference for the Netherlands is higher than that of Austria and that of Hungary (Table 5e). An MDS plot (Figure 4) was constructed to visualize the pairwise F_{ST} distances among the fourteen populations. From the MDS plot it is evident that the populations that are closest to the Netherlands geographically (Austria and Macedonia) also clustered with the Netherlands based on their mtDNA diversity, as expected. When compared to autosomal genome data, theoretically one could expect mtDNA – based F_{ST} values to be somewhat larger than autosomally – based F_{ST} values because of the lower Ne of mtDNA compared to autosomes. It has been reported that the mean F_{ST} value across all genome-wide autosomal SNPs used and the mean combined F_{ST} value between pairs of subpopulations was 0.003 and 0.00038 [8] These results suggest very small overall genetic differentiation among the 54 Dutch subpopulations sampled across the entire country when it comes to the mtDNA control region

Further, AMOVA (Table 6b) revealed that, of the variance observed among the 14 populations, 89.43% is attributable to differences within populations, and 10.57% represents differences among populations, illustrating the fact that, on a global level, mtDNA variation is significantly geographically structured.

Our study provides considerably more data for the Netherlands than most previous mtDNA studies did for other countries [5,16–27]. Nevertheless, and because of the high diversity of mtDNA CR sequences, we would advocate for a further increase in data content of the Dutch mtDNA reference database for forensic usage, for which our currently presented data serve as a (reasonably good) start. The further expansion shall also include major cities, which were deliberately left out from the present study to avoid detecting effects of more recent genetic admixture when setting the base line of Dutch European mtDNA

diversity as was the goal of this study. Furthermore, the major population groups living in the Netherlands besides Dutch Europeans will need to be included in a growing reference database to become more representative of the Netherlands.

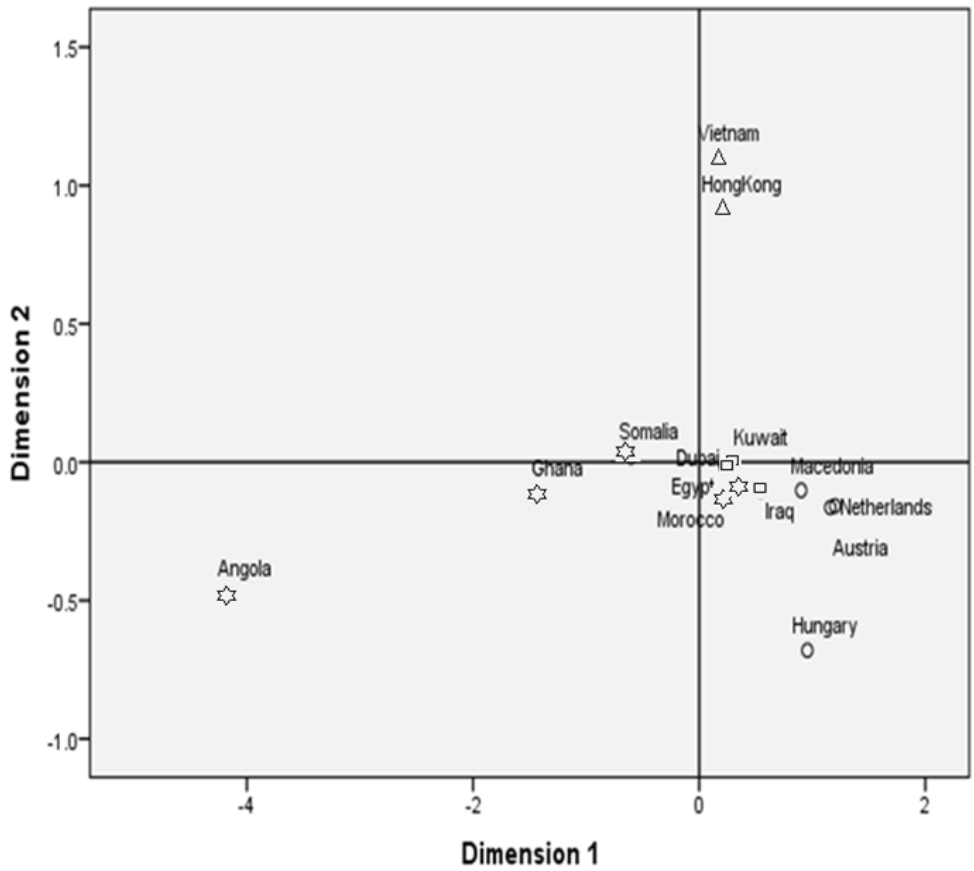


Figure 4: Multi-dimensional scaling plot illustrating the genetic distance between the Netherlands from this study and published mtDNA control region (16,024–16,569, 1–576) sequences of populations from Austria, Macedonia, Angola, Ghana, Dubai, Hungary, Morocco, Iraq, Somalia, Vietnam, Hong Kong, Kuwait and Egypt based on population pairwise differences represents O European Countries, ▫ for Middle-Eastern countries, * for African countries and Δ represents Asian countries.

(a) Population Pairwise FSTs

	Leiden	SW	SE	CS	CNE	NW	C	N	NE	CE
Leiden	0									
SW	-0.00269	0								
SE	0.00076	0.00126	0							
CS	-0.00471	0.00581	0.00423	0						
CNE	0.00083	0.00372	-0.00064	0.00294	0					
NW	-0.00173	-0.0017	0.00136	0.00761	-0.00016	0				
C	-0.00297	0.00354	0.00393	0.00027	-0.00179	0.00186	0			
N	0.00662	0.00252	0.00482	0.0062	0.00328	0.00654	0.00618	0		
NE	-0.00215	-0.00171	-0.00181	0.00514	-0.00647	-0.00472	0.00251	0.00107	0	
CE	0.00291	0.00378	0.01087	0.01095	-0.00066	-0.00387	0.00025	0.00303	-0.00803	0

(b) FST P Values

	Leiden	SW	SE	CS	CNE	NW	C	N	NE	CE
Leiden	*									
SW	0.71680 +-0.0123	*								
SE	0.40723 +-0.0163	0.28809 +-0.0166	*							
CS	0.89941 +-0.0085	0.04492 +-0.0061	0.15332 +-0.0106	*						
CNE	0.38770 +-0.0149	0.13477 +-0.0089	0.48438 +-0.0143	0.16113 +-0.0108	*					
NW	0.57129 +-0.0167	0.57617 +-0.0140	0.38379 +-0.0106	0.06738 +-0.0091	0.42773 +-0.0151	*				
C	0.75586 +-0.0120	0.14453 +-0.0107	0.18555 +-0.0121	0.42676 +-0.0190	0.63184 +-0.0167	0.29199 +-0.0159	*			
N	0.07910 +-0.0086	0.21777 +-0.0175	0.14746 +-0.0096	0.04688 +-0.0067	0.18555 +-0.0126	0.11035 +-0.0084	0.07324 +-0.0089	*		
NE	0.59961 +-0.0161	0.61035 +-0.0153	0.55957 +-0.0104	0.11035 +-0.0098	0.93750 +-0.0075	0.75488 +-0.0075	0.65430 +-0.0144	0.37402 +-0.0129	*	
CE	0.26074 +-0.0155	0.21680 +-0.0139	0.05859 +-0.0060	0.01758 +-0.0042	0.46777 +-0.0185	0.46777 +-0.0185	0.36523 +-0.0140	0.23535 +-0.0125	0.93457 +-0.0075	*

c) Population Pairwise FSTs

	Netherlands	Austria	Macedonia	Angola	Ghana	Dubai	Hungary	Morocco	Iraq	Somalia	Vietnam	Hong Kong	Kuwait	Egypt
Netherlands	0													
Austria	0.0021	0												
Macedonia	0.00574	0.00342	0											
Angola	0.51454	0.50715	0.47345	0										
Ghana	0.25196	0.23247	0.20252	0.30283	0									
Dubai	0.04119	0.03931	0.02841	0.37912	0.12195	0								
Hungary	0.03934	0.04633	0.04382	0.49434	0.2291	0.05365	0							
Morocco	0.04069	0.03497	0.02415	0.40211	0.1161	0.01835	0.05845	0						
Iraq	0.01221	0.01029	0.00741	0.43321	0.16519	0.01158	0.04385	0.01828	0					
Somalia	0.14716	0.13396	0.10734	0.30834	0.05861	0.04785	0.12803	0.05489	0.08164	0				
Vietnam	0.13711	0.12807	0.11144	0.41707	0.16952	0.06908	0.1496	0.08075	0.0862	0.08164	0			
Hong Kong	0.11713	0.10794	0.09497	0.41707	0.17167	0.05711	0.12656	0.07091	0.07376	0.07723	0.00821	0		
Kuwait	0.0306	0.02751	0.02048	0.40528	0.14896	0.00783	0.05647	0.02166	0.00507	0.07029	0.07832	0.06371	0	
Egypt	0.03292	0.02972	0.01889	0.39819	0.13055	0.00981	0.05017	0.00959	0.00699	0.04665	0.06441	0.05308	0.01186	0

(d) FST P Values

	Netherlands	Austria	Macedonia	Angola	Ghana	Dubai	Hungary	Morocco	Iraq	Somalia	Vietnam	Hong Kong	Kuwait	Egypt
Netherlands	*													
Austria	0.01270+-0.0031	*												
Macedonia	0.00098+-0.0010	0.00684+-0.0023	*											
Angola	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	*										
Ghana	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	*									
Dubai	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	*								
Hungary	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	*							
Morocco	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	*						
Iraq	0.00000+-0.0000	0.00000+-0.0000	0.00098+-0.0010	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	*					
Somalia	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	*				
Vietnam	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	*			
Hong Kong	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00195+-0.0014	*		
Kuwait	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00098+-0.0010	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	*	
Egypt	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	*

e)

	Netherlands	Austria	Macedonia	Angola	Ghana	Dubai	Hungary	Morocco	Iraq	Somalia	Vietnam	Hong Kong	Kuwait	Egypt
Average Pairwise difference	9.421710 +/- 4.330641	9.102402 +/- 4.201510	10.061307 +/- 4.617804	12.588178 +/- 5.714862	15.522142 +/- 6.960867	14.526655 +/- 6.527644	8.692365 +/- 4.032658	12.454674 +/- 5.630466	12.567361 +/- 5.695231	15.939181 +/- 7.139967	13.093497 +/- 5.920218	13.398781 +/- 6.037707	14.014201 +/- 6.300905	13.306964 +/- 6.003149

Table 5: Pairwise differences between geographic sampling regions within Netherlands (5a and b) and among fourteen worldwide populations (5c, d and e) as calculated using Arlequin v3.5.1.2 [14].

Source of Variation	Degrees of Freedom	Sum of Squares	Variance Components	Percentage of variation
Among Populations	9	47.815	0.00905	0.19
Within Populations	670	3149.857	4.70128	99.81
Total	679	3197.672	4.71033	

Source of Variation	Degrees of Freedom	Sum of Squares	Variance Components	Percentage of variation
Among Populations	13	2702.209	0.72127	10.57
Within Populations	3985	24317.365	6.10222	89.43
Total	3998	27019.574	6.82349	

Table 6: (a) AMOVA Results for the Netherlands (b) AMOVA Results for the 14 countries

Conclusions

High-quality control-region mtDNA sequences have been determined from 680 individuals sampled across the entire Netherlands and are made available via the EMPOP database to all users. The high genetic diversity, and thus low random-match probability, obtained indicates that this data can be used as a valuable tool in routine forensic casework in the Netherlands. The availability of such data will aid the assessment of the rarity of mtDNA haplotype frequencies as can be necessary in forensic or missing person cases. Such information can further corroborate the use of mtDNA as an effective forensic tool, especially when an autosomal STR profile is unattainable from limited or degraded sample material. Future emphasis should be placed to further increase the dataset aiming as (the start of) a reference database for mtDNA applications in forensic and missing person cases in the Netherlands.

Acknowledgments

We thank all volunteers to provide their samples used in this study, as well as, Mayra Eduardoff, Liane Fendt, Simone Nagl and Alexander Röck for their valuable support to this project. This work was funded in part by previous grants from the Netherlands Forensic Institute (NFI) and the Netherlands Genomic Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands (FGCN), and received additional support by the Erasmus MC University Medical Center Rotterdam. Some of the work leading to the presented results has received funding from the European Union Seventh Framework Programme (FP7/2007- 2013) (EUROFORGEN-NoE) and was in part supported by the Austrian Science Fund (FWF) [P22880-B12].

References

[1] Butler JM, Levin BC, Forensic applications of mitochondrial DNA, Trends Biotechnol. 16 (4) (1998) 158–162.

[2] Budowle B, Wilson MR, DiZinno JA, Stauffer C, Fasano MA, Holland MM, Monson KL, Mitochondrial DNA regions HVI and HVII population data, Forensic Sci. Int. 103 (1) (1999) 23–35.

[3] Sullivan KM, Hopgood R, Gill P, Identification of human remains by amplification and automated sequencing of mitochondrial DNA, Int. J. Legal Med. 105 (2) (1992) 83–86.

[4] Holland MM, Parsons TJ, Mitochondrial DNA sequences analysis—validation and use for forensic casework, Forensic Sci. Rev. 11 (1) (1999) 21–49.

[5] Brandstätter A, Niederstätter H, Pavlic M, Grubwieser P, Parson W, Generating population data for the EMPOP database—an overview of the mtDNA sequencing and data evaluation processes considering 273 Austrian control region sequences as example, Forensic Sci. Int. 166 (2-3) (2007) 164–175.

[6] Parson W, Dür A, EMPOP—a forensic mtDNA database, Forensic Sci. Int. Genet. 1 (2) (2007) 88–92.

[7] Parson W, Gusmão L, Hares DR, Irwin JA, Mayr WR, Morling N, Pokorak E, Prinz M, Salas A, Schneider PM, Parsons TJ, DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing, Forensic Sci. Int. Genet. 13 (2014) 134–142.

[8] Lao O, Altena E, Becker C, Brauer S, Kraaijenbrink T, van Oven M, Nürnberg P, de Knijff P, Kayser M, Clinal distribution of human genomic diversity across the Netherlands despite archaeological evidence for genetic discontinuities in Dutch population history, Invest. Genet. 4 (1) (2013) 9.

[9] Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N, Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA, Nat. Genet. 23 (2) (1999) 147.

[10] Parson W, Bandelt HJ, Extended guidelines for mtDNA typing of population data in forensic science, Forensic Sci. Int. Genet. 1 (1) (2007) 13–19.

[11] Bandelt HJ, Parson W, Consistent treatment of length variants in the human mtDNA control region: a reappraisal, Int. J. Legal Med. 122 (1) (2008) 11–21.

[12] Kloss-Brandstätter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G, Kronenberg F, HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups, Hum. Mutat. 32 (1) (2011) 25–32.

[13] van Oven M, Kayser M, Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation, Hum. Mutat. 30 (2) (2009) E386–94.

[14] Stoneking M, Hedgecock D, Higuchi RG, Vigilant L, Erlich HA, Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes, Am. J. Hum. Genet. 48 (2) (1991) 370–382.

[15] Excoffier L, Lischer HE, Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows, Mol. Ecol. Resour. 10 (3) (2010) 564–567.

[16] Zimmermann B, Brandstätter A, Duftner N, Niederwieser D, Spiroski M, Arsov T, Parson W, Mitochondrial DNA control region population data from Macedonia, Forensic Sci. Int. Genet. 1 (3-4) (2007) e4–9.

[17] Fendt L, Huber G, Röck AW, Zimmermann B, Bodner M, Delport R, Schmidt K, Parson W, Mitochondrial DNA control region data from indigenous Angolan Khoe–San lineages, Forensic Sci. Int. Genet. 6 (5) (2012) 662–663.

[18] Fendt L, Röck A, Zimmermann B, Bodner M, Thye T, Tschentscher F, Owusu-Dabo E, Göbel TM, Schneider PM, Parson W, MtDNA diversity of Ghana: a forensic and phylogeographic view, Forensic Sci. Int. Genet. 6 (2) (2012) 244–249.

[19] Alshamali F, Brandstätter A, Zimmermann B, Parson W, Mitochondrial DNA control region variation in Dubai, United Arab Emirates, Forensic Sci. Int. Genet. 2 (1) (2008) e9–10.

[20] Brandstätter A, Egyed B, Zimmermann B, Tordai A, Padar Z, Parson W, Mitochondrial DNA control region variation in Ashkenazi Jews from Hungary, Forensic Sci. Int. Genet. 2 (1) (2008) e4–e6.

[21] Aboukhalid R, Sturk-Andreaggi K, Bouabdellah M, Squalli D, Irwin JA, Amzazi S, Mitochondrial DNA control region variation from samples of the Moroccan population, Int. J. Legal Med. 127 (4) (2013) 757–759.

[22] Al-Zahery N, Saunier J, Ellingson K, Parson W, Parsons TJ, Irwin JA, Characterization of mitochondrial DNA control region lineages in Iraq, Int. J. Legal Med. 127 (2) (2013) 373–375.

[23] Mikkelsen M, Fendt L, Röck AW, Zimmermann B, Rockenbauer E, Hansen AJ, Parson W, Morling N, Forensic and phylogeographic characterisation of mtDNA lineages from Somalia, Int. J. Legal Med. 126 (4) (2012) 573–579.

[24] Saunier JL, Irwin JA, Strouss KM, Ragab H, Sturk KA, Parsons TJ, Mitochondrial control region sequences from an Egyptian population sample, Forensic Sci. Int. Genet. 3 (3) (2009) e97–e103.

[25] Irwin JA, Saunier JL, Strouss KM, Diegoli TM, Sturk KA, O'Callaghan JE, Paintner CD, Hohoff C, Brinkmann B, Parsons TJ, Mitochondrial control region sequences from a Vietnamese population sample, Int. J. Legal Med. 122 (3) (2008) 257–259.

[26] Irwin JA, Saunier JL, Beh P, Strouss KM, Paintner CD, Parsons TJ, Mitochondrial DNA control region variation in a population sample from Hong Kong, China, Forensic Sci. Int. Genet. 3 (4) (2009) e119–25.

[27] Scheible M, Alenizi M, Sturk-Andreaggi K, Coble MD, Ismael S, Irwin JA, Mitochondrial DNA control region variation in a Kuwaiti population sample, Forensic Sci. Int. Genet. 5 (4) (2011) e112–3.

[28] Budowle B, Allard MW, Wilson MR, Chakraborty R, Forensics and mitochondrial DNA: applications, debates, and foundations, Annu. Rev. Genomics Hum. Genet. 4 (2003) 119–141.

[29] Wilson Sayres MA, Lohmueller KE, Nielsen R, Natural selection reduced diversity on human y chromosomes, PLoS Genet. 10 (1) (2014) e1004064.

[30] Tully G, Lareu M, European DNA Profiling (EDNAP) group: controversies over heteroplasmy, Electrophoresis 22 (1) (2001) 180–182.

[31] Melton T, Mitochondrial DNA heteroplasmy, Forensic Sci. Rev. 16 (1) (2004) 1–20.

[32] Carracedo A, Bär W, Lincoln P, Mayr W, Morling N, Olaisen B, Schneider P, Budowle B, Brinkmann B, Gill P, Holland M, Tully G, Wilson M, DNA commission of the international society for forensic genetics: guidelines for mitochondrial DNA typing, Forensic Sci. Int. 110 (2) (2000) 79–85.

[33] Rohlin A, Wernersson J, Engwall Y, Wiklund L, Björk J, Nordling M, Parallel sequencing used in detection of mosaic mutations: comparison with four diagnostic DNA screening techniques, Hum. Mutat. 30 (6) (2009) 1012–1020.

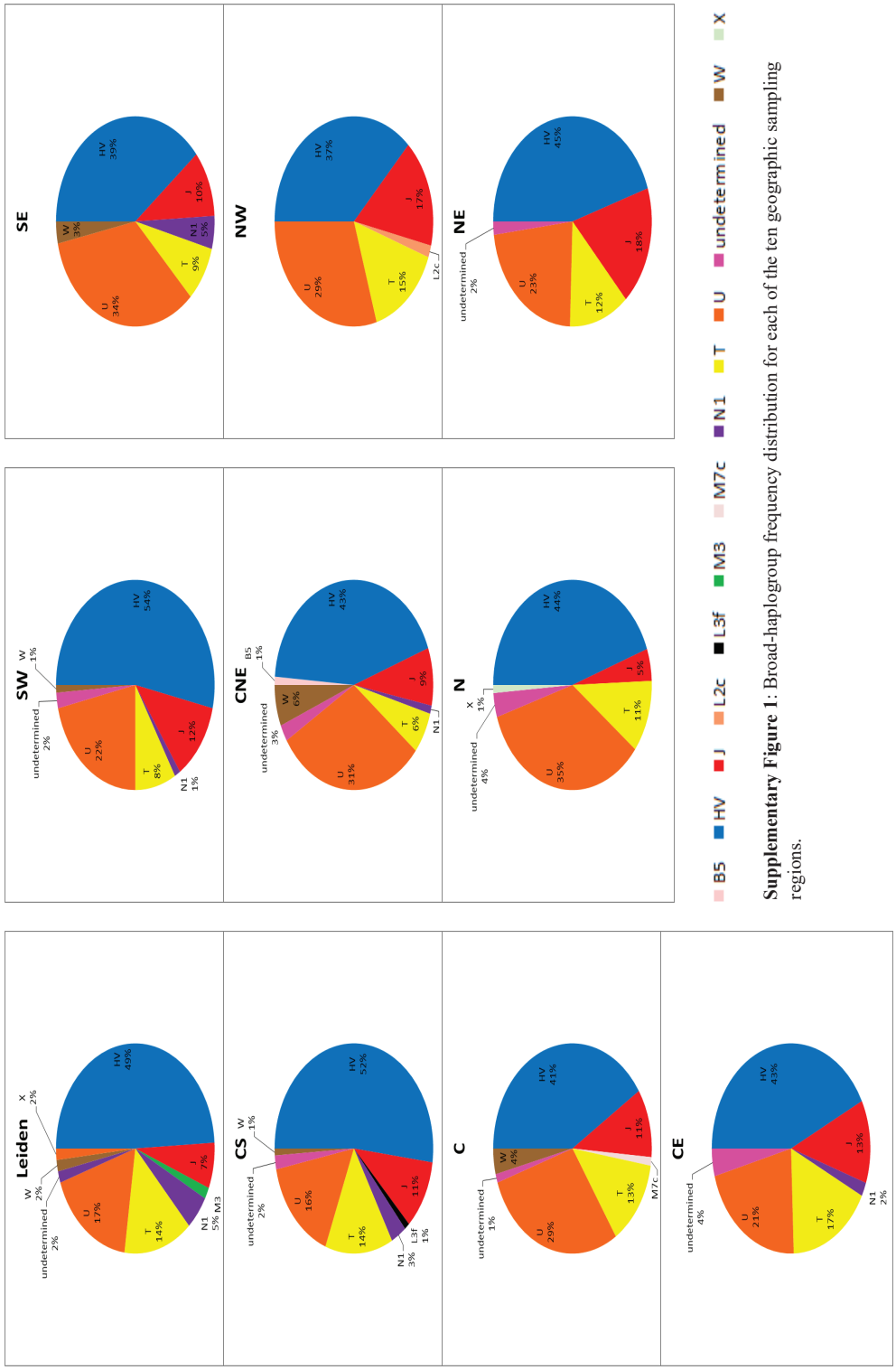
[34] Just RS, Irwin JA, Parson W, Mitochondrial DNA heteroplasmy in the emerging field of massively parallel sequencing, Forensic Sci. Int. Genet. 18 (2015) 131–139.

[35] Chaitanya L, Ralf A, van Oven M, Kupiec T, Chang J, Lagacé R, Kayser M, Simultaneous whole mitochondrial genome sequencing with short overlapping amplicons suitable for degraded DNA Using the Ion Torrent Personal Genome Machine, Hum. Mutat. 36 (12) (2015) 1236–1247.

[36] Finnilä S, Lehtonen MS, Majamaa K, Phylogenetic network for European mtDNA, Am. J. Hum. Genet. 68 (6) (2001) 1475–1484.

[37] Karachanak S, Carossa V, Nesheva D, Olivieri A, Pala M, Hooshiar Kashani B, Grugni V, Battaglia V, Achilli A, Yordanov Y, Galabov AS, Semino O, Toncheva D, Torroni A, Bulgarians vs the other European populations: a mitochondrial DNA perspective, Int. J. Legal Med. 126 (4) (2012) 497–503.

[38] Genome of the Netherlands Consortium, Whole-genome sequence variation, population structure and demographic history of the Dutch population, Nat. Genet. 46 (8) (2014) 818–825.



Supplementary Figure 1: Broad-haplogroup frequency distribution for each of the ten geographic sampling regions.

Sample ID	Region	Geographic cluster	Control-region variants (listed as differences to the rCRS)	Haplogroup assignment (PhyloTree Build 16)	Broad haplogroup
Don0001	Leiden	1	16129A,16223T,16311C,16391A,16491C,16494T,16519C,73G,152C,199C,204C,207A,239C,250C,263G,309.1C,309.2C,315.1C,573.1C,573.2C,573.3C,573.4C,573.5C	I3	N1
Don0002	Leiden	1	16239G,16311C,195C,263G,309.1C,315.1C	H11a	HV
Don0003	Leiden	1	16519C,263G,309.1C,315.1C	HV*	HV
Don0004	Leiden	1	16086C,16126C,16294T,16296T,16304C,16519C,73G,195C,263G,315.1C	T2b	T
Don0005	Leiden	1	16356C,16519C,263G,293C,309.1C,315.1C,493G	H3z	HV
Don0006	Leiden	1	16086C,16129A,16223T,16391A,16519C,73G,152C,199C,204C,207A,239C,250C,263G,309.1C,315.1C,524.1A,524.2C,524.3A,524.4C,573.1C,573.2C,573.3C,573.4C	I3	N1
Don0007	Leiden	1	16069T,16126C,16278T,16362C,73G,185A,228A,263G,295T,315.1C,462T,489C	J1c	J
Don0008	Leiden	1	16184T,16519C,263G,315.1C,523d,524d	HV*	HV
Don0009	Leiden	1	16409C,16519C,73G,195C,263G,315.1C	undetermined	undetermined
Don0010	Leiden	1	16126C,16294T,16296T,16304C,16519C,73G,263G,315.1C	T2b	T
Don0011	Leiden	1	16192T,16234T,16256T,16270T,16311C,16399G,73G,263G,309.1C,315.1C	U5a1	U
Don0012	Leiden	1	16092C,16301T,16311C,195C,263G,309.1C,315.1C	HV*	HV
Don0013	Leiden	1	16519C,263G,315.1C	HV*	HV
Don0014	Leiden	1	16146G,16519C,131C,263G,315.1C	HV*	HV
Don0015	Leiden	1	16146G,16162G,16342C,73G,263G,282C,315.1C	U8a	U
Don0016	Leiden	1	16428A,16519C,263G,315.1C	HV*	HV
Don0017	Leiden	1	16224C,16311C,16519C,73G,146C,152C,263G,309.1C,315.1C	K2a	U
Don0018	Leiden	1	16223T,16292T,16362C,16519C,73G,189G,194T,195C,204C,207A,263G,309.1C,315.1C	W	W
Don0019	Leiden	1	16519C,152C,263G,309.1C,309.2C,315.1C,524.1A,524.2C,	HV*	HV
Don0020	Leiden	1	16069T,16126C,16278T,16362C,73G,185A,228A,263G,295T,315.1C,462T,489C	J1c	J
Don0021	Leiden	1	16192T,16239T,16311C,73G,150T,263G,315.1C	U5b	U
Don0022	Leiden	1	16519C,263G,315.1C	HV*	HV
Don0023	Leiden	1	16126C,16153A,16294T,16296T,16519C,73G,150T,173C,263G,309.1C,315.1C	T2e	T
Don0024	Leiden	1	16256T,16270T,16362C,16399G,73G,152C,195C,263G,309.1C,309.2C,315.1C,524.1A,524.2C	U5a1	U
Don0025	Leiden	1	16220C,16324C,16362C,16519C,263G,309.1C,309.2C,315.1C	HV*	HV
Don0026	Leiden	1	16070G,16304C,152C,263G,309.1C,315.1C,456T	H36	HV
Don0029	Leiden	1	16126C,16163G,16186T,16189C,16294T,16519C,73G,152C,195C,263G,309.1C,315.1C,468T	T1a	T
Don0030	Leiden	1	16298C,16311C,16519C,72C,263G,309.1C,315.1C	HV0	HV
Don0031	Leiden	1	16519C,263G,315.1C,	HV*	HV
Don0032	Leiden	1	16519C,263G,309.1C,315.1C	HV*	HV
Don0033	Leiden	1	195C,263G,315.1C,523d,524d	HV*	HV
Don0034	Leiden	1	16126C,16223T,16465T,16519C,73G,146C,203A,204C,263G,309.1C,315.1C,482C,489C	M3	M3
Don0035	Leiden	1	16126C,16163G,16186T,16189C,16294T,16519C,73G,152C,195C,263G,309.1C,315.1C	T1a	T
Don0036	Leiden	1	16093C,16126C,16189Y,16294T,16296T,16304C,16519C,73G,199C,263G,315.1C	T2b	T
Don0037	Leiden	1	16519C,263G,309.1C,315.1C	HV*	HV
Don0038	Leiden	1	16193T,16219G,16274A,18862C,16482G,204C,239C,263G,315.1C	H6	HV
Don0039	Leiden	1	16298C,72C,263G,309.1C,315.1C	HV0	HV
Don0040	Leiden	1	16129A,16223T,16391A,16519C,73G,152C,199C,204C,207A,239C,250C,263G,315.1C,524.1A,524.2C,524.3A,524.4C,574C	I3	N1
Don0041	Leiden	1	16129A,16264T,16311C,16519C,263G,315.1C	HV*	HV
Don0042	Leiden	1	16519C,263G,309.1C,315.1C	HV*	HV
Don0045	Leiden	1	16069T,16126C,16145A,16231C,16261T,73G,150T,152C,195C,198Y,215G,263G,295T,310.1T,315.1C,319C,489C,513A	J2a	J
Don0046	Leiden	1	16519C,263G,315.1C	HV*	HV
Don0048	Leiden	1	16126C,16294T,16296T,16304C,16519C,3C,73G,263G,309.1C,315.1C	T2b	T
Don0049	Leiden	1	16093C,16224C,16311C,16362C,16400T,16519C,65C,73G,263G,315.1C,497T	K1a	U
Don0050	Leiden	1	16075C,16189C,16260T,16519C,263G,315.1C	H1b	HV
Don0051	Leiden	1	16224C,16245T,16311C,16519C,73G,263G,309.1C,315.1C,497T,524.1A,524.2C,573.1C,573.2C,573.3C,573.4C	K1a	U
Don0052	Leiden	1	16519C,152C,153G,263G,309.1C,315.1C	HV*	HV

Don0054	Leiden	1	16069T,16126C,73G,185A,188G,228A,263G,295T,309.1C,315.1C,462T,489C	J1c	J	
Don0055	Leiden	1	16093C,16224C,16311C,16519C,73G,114T,152C,234G,263G,315.1C,497T,524.1A,524.2C	K1a	U	
Don0056	Leiden	1	16209C,16342C,73G,227T,263G,282C,315.1C	U8a	U	
Don0057	Leiden	1	16126C,16172C,16294T,16296T,16304C,16519C,3C,73G,263G,315.1C	T2b	T	
Don0058	Leiden	1	16224C,16311C,16320T,16519C,73G,146C,152C,263G,315.1C,498d	K1c	U	
Don0059	Leiden	1	195C,263G,315.1C,523d,524d	HV*	HV	
Don0060	Leiden	1	16335G,16519C,152C,263G,315.1C	HV*	HV	
Don0063	Leiden	1	16519C,152C,263G,309.1C,315.1C,524.1A,524.2C	HV*	HV	
Don0065	Leiden	1	16304C,263G,315.1C,338T,456T	H5	HV	
Don0067	Leiden	1	16183C,16189C,16193.1C,16223T,16230G,16255A,16278T,16300G,16519C,73G,195C,225A,227G,263G,315.1C	X2c	X	
Don0121	Bergen	2	16311C,16519C,93G,263G,315.1C	HV*	HV	
Don0122	Bergen	2	16093C,16224C,16311C,16519C,73G,146C,152C,263G,315.1C,498d	K1c	U	
Don0123	Bergen	2	16224C,16311C,16519C,73G,146C,195C,263G,309.1C,315.1C,524.1A,524.2C,524.3A,524.4C	K1b	U	
Don0124	Bergen	2	16221T,16311C,16519C,263G,309.1C,315.1C	HV*	HV	
Don0125	Bergen	2	16255A,16354T,194T,263G,309.1C,309.2C,315.1C	H2a1	HV	
Don0126	Bergen	2	16069T,16126C,16209C,16519C,73G,185A,188G,228A,263G,295T,309.1C,315.1C,462T,489C	J1c	J	
Don0127	Bergen	2	16129A,16519C,93G,263G,315.1C	HV*	HV	
Don0128	Bergen	2	16311C,16519C,152C,263G,315.1C	HV*	HV	
Don0129	Bergen	2	16172C,16304C,16311C,263G,315.1C,4446,456T,523d,524d	H5	HV	
Don0151	Zierikzee	2	16311C,16519C,263G,315.1C	HV*	HV	
Don0152	Zierikzee	2	16093C,16129A,16316G,16519C,185A,263G,315.1C	H27	HV	
Don0153	Zierikzee	2	16519C,263G,309.1C,309.2C,315.1C,523d,524d	HV*	HV	
Don0154	Zierikzee	2	16293G,16519C,263G,315.1C	HV*	HV	
Don0155	Zierikzee	2	16162G,16519C,73G,263G,315.1C	H1a	HV	
Don0157	Zierikzee	2	16227G,16519C,263G,309.1C,315.1C	HV*	HV	
Don0158	Zierikzee	2	16255A,16354T,194T,263G,309.1C,309.2C,315.1C	H2a1	HV	
Don0159	Zierikzee	2	16356C,16519C,73G,195C,263G,310C,499A,524.1A,524.2C,524.3A,524.4C	U4	U	
Don0160	Zierikzee	2	16224C,16311C,16320T,16519C,73G,146C,152C,263G,315.1C,498d	K1c	U	
Don0181	Hulst	2	16215G,16519C,152C,263G,315.1C	HV*	HV	
Don0182	Hulst	2	16519C,152C,263G,315.1C	HV*	HV	
Don0183	Hulst	2	16235G,16291T,16293G,16400T,263G,309.1C,309.2C,315.1C	H2a2b	HV	
Don0184	Hulst	2	16126G,16294T,16296T,16304C,16519C,73G,263G,315.1C	T2b	T	
Don0185	Hulst	2	16223T,16292T,16295T,16324C,16519C,73G,189G,195C,204C,207A,263G,315.1C	W	W	
Don0186	Hulst	2	16069T,16126C,16145A,16172C,16222T,16254G,73G,185A,242T,263G,295T,315.1C,462T,489C	J1b	J	
Don0187	Hulst	2	16224C,16311C,16519C,73G,146C,152C,263G,309.1C,315.1C	K2a	U	
Don0188	Hulst	2	16519C,73G,263G,315.1C,573.1C	undetermined	undetermined	
Don0190	Hulst	2	16224C,16311C,16519C,73G,195C,263G,315.1C,497T,524.1A,524.2C,524.3A,524.4C	K1a	U	
Don0192	Middelburg	2	16093C,16224C,16311C,16519C,73G,114Y,152C,263G,315.1C,497T,524.1A,524.2C	K1a	U	
Don0193	Middelburg	2	16216G,16261T,16298C,16519C,72C,263G,309.1C,315.1C	HV0	HV	
Don0194	Middelburg	2	16126C,16294T,16296T,16304C,16519C,73G,263G,309.1C,315.1C	T2b	T	
Don0195	Middelburg	2	16224C,16242T,16270T,16519C,73G,150T,263G,309.1C,315.1C,513A,517T	U5b	U	
Don0196	Middelburg	2	16209C,16519C,263G,315.1C	HV*	HV	
Don0197	Middelburg	2	16192T,16209C,16239T,16311C,73G,150T,263G,315.1C	U5b	U	
Don0198	Middelburg	2	16311C,16354T,263G,315.1C,575T	H2a1	HV	
Don0199	Middelburg	2	16519C,195C,263G,309.1C,315.1C,523d,524d	HV*	HV	
Don0200	Middelburg	2	16140C,16239G,16311C,195C,263G,315.1C	H11a	HV	
Don0201	Middelburg	2	16519C,263G,309.1C,309.2C,315.1C	HV*	HV	
Don0222	Zierikzee	2	16069T,16126C,16325C,73G,185A,228A,263G,295T,315.1C,462T,489C,574C	J1c	J	
Don0223	Zierikzee	2	16092C,16235G,16239T,16519C,146C,263G,309.1C,315.1C	HV*	HV	
Don0224	Zierikzee	2	16311C,73G,150T,263G,315.1C	U5b	U	

Don0225	Zierikzee	2	16092C,16239T,16519C,263G,315.1C,560T	HV*	HV
Don0226	Zierikzee	2	16069T,16126C,16193T,16278T,73G,150T,152C,263G,295T,315.1C,489C	J2b	J
Don0227	Zierikzee	2	16263C,16519C,263G,309.1C,315.1C,477C	H1c	HV
Don0228	Zierikzee	2	16239T,16304C,146C,195C,263G,309.1C,315.1C,456T	H5	HV
Don0229	Zierikzee	2	16069T,16126C,16311C,16519C,73G,150T,195C,263G,295T,309.1C,315.1C,489C	J2a	J
Don0230	Zierikzee	2	16216G,16261T,16298C,16519C,72C,263G,309.1C,315.1C	HV0	HV
Don0231	Zierikzee	2	16215G,16519C,152C,263G,315.1C	HV*	HV
Don0271	Oostburg	2	16519C,143A,263G,315.1C	HV*	HV
Don0272	Oostburg	2	16086C,16129A,16223T,16391A,16519C,73G,152C,199C,204C,207A,239C,250C,263G,309.1C,315.1C,524.1A,524.2C,524.3C,524.4C,573.1C,573.2C,573.3C	I3	N1
Don0273	Oostburg	2	16129A,16224C,16311C,16519C,73G,146C,195C,263G,315.1C	K2a	U
Don0275	Oostburg	2	16311C,16519C,263G,309.1C,315.1C	HV*	HV
Don0276	Oostburg	2	16169T,16299G,16519C,263G,309.1C,315.1C	H39a	HV
Don0278	Oostburg	2	16126C,16172C,16294T,16304C,16354T,16519C,73G,263G,309.1C,315.1C	T2b	T
Don0279	Oostburg	2	16519C,263G,309.1C,315.1C	HV*	HV
Don0280	Middelburg	2	16519C,263G,309.1C,315.1C	HV*	HV
Don0281	Middelburg	2	16293G,16519C,131C,263G,309.1C,315.1C	HV*	HV
Don0282	Middelburg	2	16224C,16311C,16519C,73G,146C,152C,263G,315.1C	K2a	U
Don0283	Middelburg	2	16129A,16519C,263G,309.1C,315.1C	HV*	HV
Don0284	Middelburg	2	16519C,131C,263G,315.1C	HV*	HV
Don0286	Middelburg	2	16126C,16163G,16186T,16189C,16294T,16519C,73G,152C,195C,263G,309.1C,315.1C	T1a	T
Don0287	Middelburg	2	16069T,16126C,73G,185A,263G,295T,315.1C,408A,462T,489C	J1c	J
Don0288	Middelburg	2	16069T,16093C,16126C,16145A,16172C,16222T,16261T,73G,242T,263G,295T,315.1C,462T,489C	J1b	J
Don0299	Bergen	2	16126C,16241G,16256T,16294T,16304C,16519C,73G,263G,309.1C,315.1C	T2b	T
Don0302	Bergen	2	16230G,16261T,16519C,228C,263G,309.1C,315.1C	HV*	HV
Don0303	Bergen	2	16213A,16362C,16482G,16497G,152C,239C,263G,315.1C	H6	HV
Don0305	Bergen	2	16126C,16294T,16296T,16304C,16519C,3C,73G,263G,315.1C	T2b	T
Don0306	Bergen	2	16519C,73G,263G,315.1C	undetermined	undetermined
Don0308	Bergen	2	16263C,16519C,64T,263G,315.1C,477C	H1c	HV
Don0309	Bergen	2	16192T,16270T,16304C,16526A,73G,150T,263G,309.1C,315.1C	U5b	U
Don0310	Hulst	2	16126C,16163G,16186T,16189C,16294T,16519C,73G,152C,195C,263G,309.1C,315.1C	T1a	T
Don0311	Hulst	2	16094C,16134T,16146G,16356C,16519C,73G,152C,195C,263G,315.1C,499A	U4	U
Don0313	Hulst	2	16256T,16270T,16362C,16399G,73G,204C,263G,185A,309.1C,315.1C,524.1A,524.2C	U5a1	U
Don0314	Hulst	2	16162G,16258C,16519C,73G,263G,309.1C,315.1C	H1a	HV
Don0315	Hulst	2	16356C,16519C,73G,146C,195C,263G,309.1C,315.1C,499A	U4	U
Don0316	Hulst	2	16192T,16270T,16304C,16311Y,16526A,73G,150T,263G,309.1C,315.1C	U5b	U
Don0327	Oostburg	2	16124C,16354T,16519C,263G,309.1C,309.2C,315.1C,575T	H2a1	HV
Don0328	Oostburg	2	16298C,72C,263G,309.1C,309.2C,315.1C	HV0	HV
Don0329	Oostburg	2	16069T,16126C,16145A,16231C,16261T,73G,150T,152C,195C,263G,295T,315.1C,319C,489C,513A	J2a	J
Don0330	Oostburg	2	16069T,16126C,16319A,73G,185A,228A,234G,263G,295T,309.1C,315.1C,462T,489C,524.1A,524.2C,524.3A,524.4C	J1c	J
Don0331	Oostburg	2	16129A,16304C,152C,263G,309.1C,315.1C,456T	H5	HV
Don0332	Oostburg	2	16519C,263G,309.1C,315.1C	HV*	HV
Don0333	Oostburg	2	16069T,16126C,16145A,16231C,16261T,73G,150T,152C,195C,215G,263G,295T,309.1C,310.1T,315.1C,319C,489C,513A	J2a	J
Don0334	Oostburg	2	16298C,16527T,263G,315.1C	HV0	HV
Don0335	Oostburg	2	16129A,16224C,16311C,16519C,73G,146C,195C,263G,315.1C	K2a	U
Don0385	Sittard	3	16184T,16304C,263G,309.1C,315.1C,456T,523d,524d	H5	HV
Don0387	Sittard	3	16519C,263G,315.1C	HV*	HV
Don0389	Sittard	3	16287T,16343G,16390A,16519C,16527Y,73G,150T,263G,315.1C	U3a	U
Don0390	Sittard	3	16192T,16256T,16270T,16291T,16294T,16399G,73G,263G,309.1C,315.1C	U5a1	U
Don0391	Sittard	3	16192T,16304C,263G,315.1C,456T	H5	HV
Don0392	Sittard	3	16093C,16163G,16179T,16224C,16311C,16519C,73G,195C,263G,309.1C,315.1C,497T,524.1A,524.2C,524.3A,524.4C	K1a	U

Don0393	Sittard	3	16311C,16356C,16519C,73G,195C,263G,310C,499A,524.1A,524.2C,524.3A,524.4C	U4	U
Don0394	Sittard	3	16290T,16294T,16304C,16320T,93G,152C,263G,315.1C,456T	H5	HV
Don0395	Sittard	3	16192T,16304C,263G,315.1C,456T,	H5	HV
Don0396	Sittard	3	16356C,16519C,73G,146C,152C,182T,195C,263G,309.1C,315.1C,499A,524.1A,524.2C,524.3A,524.4C	U4	U
Don0397	Sittard	3	16298C,72C,195C,263G,309.1C,315.1C	HV0	HV
Don0399	Sittard	3	16287T,16304C,263G,309.1C,315.1C,456T,524d,524d,573.1C,573.2C,573.3C,573.4C	H5	HV
Don0400	Sittard	3	16066G,16192T,16304C,152C,263G,315.1C,456T	H5	HV
Don0402	Sittard	3	16134T,16356C,16519C,73G,152C,195C,215G,263G,315.1C,499A,524.1A,524.2C	U4	U
Don0403	Sittard	3	16093C,16224C,16311C,16519C,73G,152C,263G,309.1C,315.1C,497T,524.1A,524.2C	K1a	U
Don0405	Oss	4	16224C,16311C,16519C,73G,195C,263G,315.1C,497T,524.1A,524.2C	K1a	U
Don0406	Oss	4	16223T,16292T,16519C,73G,189G,195C,204C,207A,263G,309.1C,315.1C	W	W
Don0407	Roermond	3	16519C,263G,309.1C,309.2C,315.1C,477C	H1c	HV
Don0408	Oss	4	16192T,16209C,16256T,16270T,16362C,16399G,73G,263G,315.1C	U5a1	U
Don0409	Oss	4	16519C,152C,263G,315.1C,513A	HV*	HV
Don0410	Oss	4	16256T,16270T,16399G,16519C,73G,263G,309.1C,309.2C,315.1C	U5a1	U
Don0411	Oss	4	93G,315.1C,523d,524d	H2a2a1	HV
Don0412	Oss	4	16192T,16223T,16292T,16325C,16519C,73G,148A,189G,194T,195C,204C,207A,263G,309.1C,315.1C	W	W
Don0413	Oss	4	16126C,16294T,16296T,16324C,16519C,73G,263G,309.1C,309.2C,315.1C	T2a1b	T
Don0414	Oss	4	16162G,16209C,16519C,73G,263G,315.1C	H1a	HV
Don0415	Oss	4	16162G,16209C,16519C,73G,315.1C	H1a	HV
Don0416	Oss	4	16126C,16294T,16296T,16304C,16519C,73G,263G,315.1C	T2b	T
Don0417	Oss	4	16256T,16270T,16399G,73G,263G,309.1C,315.1C,573.1C,573.2C,573.3C	U5a1	U
Don0418	Oss	4	16086C,16129A,16223T,16243C,16311C,16391A,16519C,73G,152C,199C,204C,207A,239C,250C,263G,315.1C,456T,524.1A,524.2C,524.3A,524.4C,524.5C,524.6C,524.7C,524.8C,573.1C,573.2C,573.3C,573.4C,573.5C	I3	N1
Don0419	Oss	4	16224C,16311C,16519C,73G,146C,152C,204Y,263G,315.1C,324T	K2a	U
Don0420	Oss	4	16069T,16126C,16166G,16519C,73G,146C,185A,188G,228A,263G,295T,315.1C,462T,489C	J1c	J
Don0421	Oss	4	16192T,16256T,16270T,16291T,16399G,73G,263G,315.1C	U5a1	U
Don0422	Oss	4	16069T,16126C,16145A,16231C,16261T,73G,150T,152C,195C,215G,263G,295T,310.1T,315.1C,319C,489C,513A,523d,524d	J2a	J
Don0423	Oss	4	16126C,16298C,16346C,72C,152C,200G,263G,309.1C,315.1C,523d,524d	HV0	HV
Don0424	Oss	4	16069T,16126C,16362C,73G,185A,263G,295T,315.1C,462T,489C	J1c	J
Don0465	Zaltbommel	4	16224C,16245T,16311C,16519C,16524G,73G,263G,309.1C,315.1C,497T,524.1A,524.2C,524.3A,524.4C,573.1C,573.2C,573.3C	K1a	U
Don0466	Zaltbommel	4	16069T,16126C,73G,185A,228A,263G,295T,315.1C,462T,489C	J1c	J
Don0467	Zaltbommel	4	16354T,263G,309.1C,315.1C,575T	H2a1	HV
Don0468	Zaltbommel	4	16192T,16238T,16256T,16270T,16399G,73G,263G,309.1C,315.1C	U5a1	U
Don0469	Zaltbommel	4	16126C,16294T,16304C,16519C,73G,263G,315.1C	T2b	T
Don0470	Zaltbommel	4	16519C,263G,309.1C,315.1C	HV*	HV
Don0471	Zaltbommel	4	16291T,16519C,263G,315.1C,499A	HV*	HV
Don0472	Zaltbommel	4	16134T,16356C,16519C,73G,152C,195C,263G,309.1C,315.1C,499A,524.1A,524.2C	U4	U
Don0473	Zaltbommel	4	16069T,16126C,16519C,73G,185A,188G,228A,263G,295T,315.1C,462T,489C	J1c	J
Don0474	Zaltbommel	4	16519C,263G,315.1C	HV*	HV
Don0476	Zaltbommel	4	16519C,263G,315.1C	HV*	HV
Don0477	Zaltbommel	4	16298C,16311C,72C,263G,309.1C,315.1C	HV0	HV
Don0478	Zaltbommel	4	16519C,263G,315.1C	HV*	HV
Don0479	Zaltbommel	4	16129A,16223T,16311C,16391A,16491C,16494T,16519C,73G,152C,199C,204C,207A,239C,250C,263G,309.1C,315.1C,315.1C,524.1A,524.2C,573.1C,573.2C,573.3C,573.4C,573.5C	I3	N1
Don0480	Zaltbommel	4	16519C,263G,309.1C,315.1C	HV*	HV
Don0481	Zaltbommel	4	16179T,16356C,16362C,16519C,73G,150T,195C,263G,315.1C,499A,524.1A,524.2C,524.3A,524.4C	U4	U
Don0482	Zaltbommel	4	16145A,16176G,16209C,16223T,16390A,16519C,73G,152C,263G,309.1C,315.1C	N1b1	N1
Don0483	Zaltbommel	4	16092C,16209C,16256T,16270T,16399G,73G,153G,263G,315.1C	U5a1	U
Don0484	Zaltbommel	4	16126C,16294T,16296T,16304C,16519C,73G,263G,315.1C	T2b	T
Don0510	Maastricht	3	16069T,16126C,73G,185A,228A,263G,295T,315.1C,462T,489C	J1c	J
Don0511	Maastricht	3	16192T,16256T,16270T,16362C,16399G,73G,263G,315.1C	U5a1	U

Don0512	Sittard	3	16192T,16239T,16311C,73G,150T,263G,315.1C	U5b	U
Don0519	Maastricht	3	16126C,16163G,16186T,16189C,16294T,16519C,73G,152C,195C,263G,309.1C,315.1C	T1a	T
Don0521	Maastricht	3	16519C,263G,315.1C	HV*	HV
Don0522	Maastricht	3	16104T,16126C,16294T,16304C,16519C,73G,152C,263G,309.1C,315.1C	T2b	T
Don0523	Maastricht	3	16209C,16223T,16311C,16519C,73G,150T,152C,189G,200G,263G,309.1C,315.1C,523d,524d	L3f1b	L3f
Don0527	Heerlen	3	16069T,16126C,16166G,16519C,73G,146C,185A,188G,195C,228A,263G,295T,315.1C,462T,489C	J1c	J
Don0530	Heerlen	3	16093C,16263C,16519C,263G,315.1C,477C	H1c	HV
Don0531	Heerlen	3	16126C,16294T,16296T,16324G,16519C,73G,263G,309.1C,315.1C	T2a1b	T
Don0543	Heerlen	3	16192T,16256T,16270T,16526A,73G,263G,315.1C	U5a2	U
Don0544	Heerlen	3	16126C,16153A,16294T,16519C,411,73G,150T,263G,315.1C	T2e	T
Don0545	Weert	3	16069T,16126C,16222T,16366T,73G,228A,263G,295T,315.1C,462T,489C	J1c	J
Don0546	Weert	3	16231C,16256T,16270T,16399G,73G,152C,263G,315.1C	U5a1	U
Don0547	Weert	3	16051G,16069T,16126C,16147T,73G,185A,228A,263G,295T,315.1C,462T,489C	J1c	J
Don0548	Weert	3	16239T,16519C,263G,315.1C	HV*	HV
Don0549	Weert	3	16189C,16239T,16519C,263G,315.1C	HV*	HV
Don0558	Weert	3	16309G,16318T,16519C,73G,152C,263G309.1C,315.1C,523d,524d	U7	U
Don0559	Weert	3	16354T,263G,315.1C	H2a1	HV
Don0560	Weert	3	16239T,16362C,16519C,263G,309.1C,315.1C	HV*	HV
Don0561	Weert	3	16519C,152C,204C,263G,315.1C	HV*	HV
Don0562	Weert	3	16213A,16362C,16482G,152C,239C,263G,309.1C,309.2C,315.1C	H6	HV
Don0563	Weert	3	16129R,16174T,16519C,263G,309.1C,315.1C,477C	H1c	HV
Don0564	Weert	3	16114T,16519C,263G,315.1C	HV*	HV
Don0586	Veghel	4	16519C,73G,263G,315.1C	undetermined	undetermined
Don0587	Veghel	4	16126C,16163G,16186T,16189C,16294T,16519C,73G,146C,152C,195C,263G,309.1C,315.1C,469T	T1a	T
Don0588	Veghel	4	16129A,16519C,263G,315.1C	HV*	HV
Don0589	Veghel	4	16126C,16192T,16294T,16519C,73G,263G,315.1C	T2b	T
Don0590	Veghel	4	16093C,16362C,16519C,152C,263G,309.1C,315.1C	HV*	HV
Don0591	Veghel	4	16093C,16263C,16519C,152C,263G,315.1C,477C	H1c	HV
Don0592	Veghel	4	16519C,263G,309.1C,315.1C	HV*	HV
Don0593	Veghel	4	16239G,16311C,195C,207A,263G,309.1C,315.1C	H11a	HV
Don0594	Veghel	4	16126C,16298C,16346C,72C,200G,263G,309.1C,315.1C,523d,524d	HV0	HV
Don0595	Veghel	4	16223T,16292T,16519C,73G,189G,194T,195C,199C,204C,207A,263G,309.1C,315.1C	W	W
Don0596	Veghel	4	16126C,16163G,16186T,16189C,16294T,16519C,73G,152C,195C,263G,315.1C	T1a	T
Don0597	Veghel	4	16193T,16219G,16319A,16362C,16482G,195C,203A,204C,239C,263G,315.1C	H6	HV
Don0598	Veghel	4	16126C,16163G,16186T,16189C,16294T,16519C,73G,152C,195C,263G,309.1C,309.2C,315.1C,469T	T1a	T
Don0599	Veghel	4	16129A,16519C,263G,309.1C,315.1C	HV*	HV
Don0600	Veghel	4	16519C,263G,315.1C	HV*	HV
Don0601	Veghel	4	16193T,16219G,16319A,16362C,16482G,195C,239C,263G,315.1C	H6	HV
Don0603	Veghel	4	16519C,263G,315.1C	HV*	HV
Don0604	Waalwijk	4	16069T,16126C,16527T,73G,228A,263G,295T,309.1C,315.1C,462T,489C	J1c	J
Don0605	Waalwijk	4	16174T,16189C,16192T,16270T,16311C,73G,150T,263G,309.1C,315.1C,516T	U5b	U
Don0606	Waalwijk	4	16224C,16311C,16519C,73G,195C,263G,309.1C,315.1C,497T,524.1A,524.2C	K1a	U
Don0607	Waalwijk	4	16192T,16256T,16270T,16311C,16320T,16399G,73G,263G,315.1C	U5a1	U
Don0608	Waalwijk	4	16129A,16189C,16519C,152C,263G,315.1C	HV*	HV
Don0609	Waalwijk	4	16129A,16223T,16259Y,16391A,16519C,73G,152C,199C,204C,207A,250C,263G,309.1C,315.1C,573.1C,573.2C,573.3C,573.4C	I2	N1
Don0610	Waalwijk	4	16263C,16519C,263G,315.1C,477C	H1c	HV
Don0611	Waalwijk	4	16304C,143A,204C,263G,315.1C,316A,456T	H5	HV
Don0612	Waalwijk	4	16051G,16092C,16129C,16182C,16183C,16189C,16362C,16519C,73G,152C,217C,263G,309.1C,309.2C,309.3C,315.1C,508G	U2e	U
Don0613	Waalwijk	4	16193T,16219G,16319A,16362C,16482G,195C,204C,239C,263G,315.1C	H6	HV
Don0614	Waalwijk	4	16086C,16129A,16223T,16319A,16391A,16519C,73G,152C,199C,204C,207A,239C,246C,250C,263G,309.1C,315.1C,524.1A,524.2C,524.3A,524.4C,573.1C,573.2C,573.3C	I3	N1

Don0615	Waalwijk	4	16086Y,16224C,16245T,16311C,16519C,73G,263G,309.1C,315.1C,497T,524.1A,524.2C,573.1C,573.2C,573.3C	K1a	U
Don0616	Waalwijk	4	16239T,16519C,263G,315.1C,524.1A,524.2C	HV*	HV
Don0617	Waalwijk	4	16519C,263G,315.1C	HV*	HV
Don0618	Waalwijk	4	16263C,16519C,263G,315.1C,477C	H1c	HV
Don0619	Waalwijk	4	16129A,16183C,16189C,16193.1C,16519C,263G,309.1C,309.2C,315.1C,408A	H3v	HV
Don0620	Waalwijk	4	16092C,16189C,16519C,146C,153G,204C,263G,309.1C,309.2C,315.1C	HV*	HV
Don0621	Waalwijk	4	16224C,16311C,16519C,73G,195C,263G,315.1C,417A,497T,524.1G,524.2C	K1a	U
Don0622	Waalwijk	4	16069T,16126C,16145A,16231C,16261T,73G,150T,152C,195C,215G,263G,295T,310.1T,315.1C,319C,489C,513A	J2a	J
Don0623	Waalwijk	4	16126C,16182C,16183C,16189C,16294T,16296T,16298C,16519C,73G,195C,263G,315.1C	T2f1	T
Don0624	Druten	4	16519C,263G,309.1C,315.1C	HV*	HV
Don0625	Druten	4	16129A,16172C,16223T,16311C,16391A,16519C,73G,188G,199C,203A,204C,250C,263G,315.1C,455.1T,573.1C,573.2C,573.3C	I1	N1
Don0626	Druten	4	16069T,16126C,16209C,16519C,73G,185A,188G,228A,263G,295T,309.1C,315.1C,462T,489C	J1c	J
Don0627	Druten	4	16304C,143A,204C,263G,309.1C,315.1C,316A,456T	H5	HV
Don0628	Druten	4	16298C,72C,195C,263G,309.1C,315.1C	HV0	HV
Don0629	Druten	4	16519C,73G,195C,263G,315.1C	undetermined	undetermined
Don0630	Druten	4	16069T,16126C,16304C,16382C,73G,185A,263G,295T,315.1C,462T,489C	J1c	J
Don0631	Druten	4	16298C,72C,195C,263G,309.1C,315.1C	HV0	HV
Don0632	Druten	4	16051G,16129C,16183C,16189C,16362C,16519C,73G,152C,217C,263G,315.1C,340T,508G,524.1A,524.2C,524.3A,524.4C	U2e	U
Don0633	Druten	4	16126C,16183G,16186T,16189C,16274A,16294T,16519C,73G,199C,263G,309.1C,309.2C,315.1C,512G	T1a	T
Don0634	Druten	4	16185T,16263C,16519C,263G,315.1C,477C	H1c	HV
Don0635	Druten	4	16162G,16209C,16519C,73G,146Y,263G,315.1C	H1a	HV
Don0636	Druten	4	16343G,16390A,16519C,73G,150T,263G,309.1C,315.1C	U3a	U
Don0637	Druten	4	16192T,16256T,16270T,16362C,16399G,16428A,73G,263G,309.1C,315.1C	U5a1	U
Don0638	Druten	4	16192T,16256T,16270T,16362C,16399G,73G,263G,309.1C,315.1C	U5a1	U
Don0639	Druten	4	16129A,16519C,263G,309.1C,315.1C	HV*	HV
Don0640	Druten	4	16183C,16188C,16193.1C,16519C,263G,315.1C,540T	HV*	HV
Don0642	Druten	4	16126C,16153A,16182C,16183C,16189C,16265T,16294T,16296T,16519C,73G,150T,204V,263G,309.1C,309.2C,315.1C	T2e	T
Don0643	Druten	4	16519C,263G,309.1C,309.2C,315.1C	HV*	HV
Don0665	Roermond	3	16311C,93G,152C,195C,263G,315.1C	HV*	HV
Don0666	Roermond	3	16126C,16294T,16304C,16519C,73G,152C,263G,309.1C,315.1C	T2b	T
Don0667	Roermond	3	16184T,16256T,16270T,16399G,73G,263G,315.1C	U5a1	U
Don0668	Roermond	3	16519C,146C,263G,309.1C,315.1C	HV*	HV
Don0669	Roermond	3	16051G,16093C,16162G,16278T,16519C,73G,263G,309.1C,315.1C	H1a	HV
Don0670	Roermond	3	16126C,16292T,16294T,16296T,16324C,16519C,73G,263G,309.1C,315.1C	T2a1b	T
Don0671	Roermond	3	16264T,16519C,146C,150T,263G,309.1C,315.1C	HV*	HV
Don0672	Roermond	3	16069T,16126C,16519C,73G,185A,188G,228A,263G,295T,315.1C,462T,489C	J1c	J
Don0673	Roermond	3	16180G,16298C,72C,263G,309.1C,309.2C,315.1C	HV0	HV
Don0674	Roermond	3	16298C,72C,195C,198T,263G,309.1C,309.2C,315.1C	HV0	HV
Don0675	Roermond	3	263G,309.1C,315.1C,456T,523d,524d	H5'36*	HV
Don0676	Roermond	3	16129A,16316G,16519C,263G,315.1C	H27	HV
Don0677	Roermond	3	16069T,16126C,16207G,16311C,16519C,73G,185A,188G,228A,263G,295T,315.1C,462T,489C,523d,524d	J1c	J
Don0678	Roermond	3	16126C,16183G,16186T,16189C,16294T,16519C,73G,152C,195C,263G,315.1C	T1a	T
Don0680	Roermond	3	16145A,16311C,16519C,263G,309.1C,315.1C	HV*	HV
Don0681	Roermond	3	16069T,16126C,16166G,16519C,73G,146C,185A,188G,228A,263G,295T,315.1C,462T,489C	J1c	J
Don0683	Roermond	3	16298C,72C,263G,309.1C,315.1C	HV0	HV
Don0684	Roermond	3	16257T,263G,309.1C,309.2C,315.1C	HV*	HV
Don0685	Roermond	3	16134T,16356C,16519C,73G,152C,195C,215G,263G,315.1C,499A,524.1A,524.2C	U4	U
Don0687	Meppel	5	16261T,16519C,263G,309.1C,315.1C	HV*	HV
Don0688	Meppel	5	16265C,16356C,16400Y,16519C,73G,146C,152C,195C,263G,309.1C,315.1C,499A	U4	U
Don0691	Losser	10	16311C,263G,309.1C,315.1C	HV*	HV

Don0692	Losser	10	no difference to rCRS	H2a2a1	HV
Don0704	Losser	10	16519C,73G,263G,315.1C	undetermined	undetermined
Don0705	Losser	10	16224C,16311C,16519C,73G,195C,263G,315.1C,497T	K1a	U
Don0707	Losser	10	16223T,16256T,16292T,16519C,73G,119C,189G,195C,204C,207A,214G,227G,263G,315.1C,	W	W
Don0708	Losser	10	16235R,16291T,16293G,263G,315.1C	H2a2b	HV
Don0709	Losser	10	16224C,16311C,16519C,73G,146C,152C,263G,309.1C,315.1C	K2a	U
Don0715	Haaksbergen	10	16093C,16224C,16311C,16519C,73G,152C,263G,315.1C,497T,524.1A,524.2C	K1a4c	U
Don0716	Haaksbergen	10	16069T,16126C,16265C,16319A,73G,185A,228A,263G,295T,309.1C,315.1C,462T,489C,524.1A,524.2C	J1c8a2	J
Don0717	Haaksbergen	10	16183C,16189C,16261T,16519C,146C,152C,263G,315.1C	H1ap1	HV
Don0718	Haaksbergen	10	16261T,16519C,263G,309.1C,315.1C	H7a1	HV
Don0719	Haaksbergen	10	16063C,16069T,16126C,73G,228A,263G,295T,315.1C,462T,489C	J1c3f	J
Don0720	Haaksbergen	10	16126C,16163G,16186T,16189C,16294T,16344T,16519C,73G,195C,263G,315.1C	T1a1l	T
Don0721	Haaksbergen	10	16134T,16356C,16519C,73G,152C,195C,263G,309.1C,315.1C,499A,524.1A,524.2C	U4a1	U
Don0722	Haaksbergen	10	16076T,16189C,16270T,16465T,73G,150T,152C,263G,315.1C,523d,524d	U5b1e	U
Don0723	Haaksbergen	10	16224C,16311C,16519C,73G,146C,263G,296T,309.1C,315.1C	K2a	U
Don0724	Haaksbergen	10	16215G,16224C,16311C,16368C,73G,263G,309.1C,315.1C,385G,497T	K1a	U
Don0725	Haaksbergen	10	16224C,16311C,16519C,73G,1147,152C,263G,315.1C,497T	K1a	U
Don0729	Haaksbergen	10	16519C,152C,263G,315.1C	HV*	HV
Don0731	Haaksbergen	10	16093C,16224C,16311C,16519C,73G,152C,263G,315.1C,497T,524.1A,524.2C	K1a	U
Don0732	Haaksbergen	10	16256T,16270T,16399G,73G,263G,309.1C,315.1C	U5a1	U
Don0748	Mijdrecht	7	16189C,16356C,16519C,199C,263G,315.1C	H1b	HV
Don0749	Mijdrecht	7	16224C,16311C,16519C,73G,263G,309.1C,315.1C,497T	K1a	U
Don0759	Mijdrecht	7	16304C,16342C,263G,315.1C,456T,523d,524d	H5	HV
Don0760	Mijdrecht	7	16192T,16223T,16292T,16325C,16519C,73G,189G,194T,195C,204C,207A,263G,315.1C	W	W
Don0761	Mijdrecht	7	16069T,16126C,16231C,73G,150T,195C,263G,295T,315.1C,489C	J2a	J
Don0762	Mijdrecht	7	16519C,263G,309.1C,315.1C	HV*	HV
Don0764	Mijdrecht	7	16519C,152C,263G,315.1C	HV*	HV
Don0765	Mijdrecht	7	16179T,16221T,16356C,16519C,73G,146C,195C,263G,315.1C,499A,524.1A,524.2C	U4	U
Don0769	Barneveld	7	16069T,16126C,16145A,16231C,16261T,73G,150T,152C,195C,215G,263G,295T,310.1T,315.1C,319C,489C,513A	J2a	J
Don0770	Barneveld	7	16223T,16292T,16519C,73G,189G,194T,195C,199C,204C,207A,263G,309.1C,309.2C,315.1C	W	W
Don0771	Barneveld	7	16126C,16163G,16186T,16189C,16274A,16294T,16519C,73G,199C,263G,309.1C,315.1C,512G	T1a	T
Don0772	Barneveld	7	16093C,16129A,16145A,16316G,16519C,152C,236C,263G,315.1C	H27	HV
Don0773	Barneveld	7	16146G,16162G,16342C,73G,263G,282C,309.1C,315.1C	U8a	U
Don0774	Barneveld	7	16223T,16292T,16362C,16519C,73G,189G,194T,195C,204C,207A,263G,315.1C	W	W
Don0775	Barneveld	7	16093I,73G,263G,315.1C,523d,524d	undetermined	undetermined
Don0777	Barneveld	7	16311C,16519C,263G,315.1C	HV*	HV
Don0778	Barneveld	7	16092C,16140C,16311C,195C,263G,315.1C	H11a	HV
Don0780	Barneveld	7	16093C,16224C,16311C,16319A,16463G,16519C,73G,152C,199C,263G,309.1C,315.1C,524.1A,524.2C	K1b	U
Don0781	Barneveld	7	16519C,263G,315.1C	HV*	HV
Don0782	Barneveld	7	16069T,16126C,16193T,16195C,16242A,16326A,73G,150T,152C,263G,295T,309.1C,315.1C,489C	J2b	J
Don0783	Barneveld	7	16222T,16224C,16270T,16311C,16519C,73G,146C,263G,315.1C	K2a	U
Don0785	Barneveld	7	16311C,263G,309.1C,315.1C	HV*	HV
Don0802	Heemskerk	6	16126C,16294T,16296T,16304C,16519C,73G,200G,263G,315.1C	T2b	T
Don0803	Heemskerk	6	16069T,16126C,16188C,16189C,16519C,73G,185A,228A,263G,295T,309.1C,315.1C,462T,489C	J1c	J
Don0820	Hilversum	7	16223T,16292T,16296T,16519C,73G,189G,195C,204C,207A,263G,315.1C	W	W
Don0821	Hilversum	7	16224C,16293G,16311C,16519C,73G,146C,152C,263G,309.1C,315.1C	K2a	U
Don0822	Hilversum	7	152C,263G,309.1C,315.1C	HV*	HV
Don0823	Hilversum	7	16263G,16519C,152C,263G,315.1C,477C	H1c	HV
Don0824	Hilversum	7	16069T,16126C,16189Y,16213A,73G,185A,228A,263G,295T,309.1C,315.1C,462T,482C,489C	J1c	J
Don0827	Hilversum	7	16093C,16224C,16311C,16362C,16400T,16519C,73G,263G,315.1C,497T	K1a	U

Don0828	Hilversum	7	16256T,16270T,16399G,73G,263G,315.1C	U5a1	U
Don0829	Hilversum	7	16129A,16223T,16391A,16519C,73G,199C,204C,207A,250C,263G,309.1C,315.1C,573.1C,573.2C,573.3C,573.4C,573.5C	I2	N1
Don0830	Hilversum	7	16263C,16519C,263G,315.1C,477C,	H1c	HV
Don0831	Hilversum	7	16051G,16162G,16465T,16519C,73G,195C,263G,315.1C,523d,524d	H1a	HV
Don0832	Hilversum	7	16304C,263G,315.1C,456T,523d,524d	H5	HV
Don0833	Hilversum	7	16519C,152C,153G,263G,309.1C,309.2C,315.1C	HV*	HV
Don0834	Hilversum	7	16126C,16294T,16296T,16304C,16519C,73G,263G,315.1C	T2b	T
Don0835	Hilversum	7	16256T,16348G,16380A,16519C,73G,150T,152C,263G,309.1C,315.1C,524.1A,524.2C	U3a	U
Don0836	Hilversum	7	152C,263G,309.1C,309.2C,315.1C	HV*	HV
Don0837	Hilversum	7	16126C,16153A,16183C,16189C,16294T,16519C,41T,73G,150T,247A,263G,315.1C	T2e	T
Don0866	Hoorn	6	16224C,16311C,16519C,73G,146C,152C,263G,309.1C,315.1C	K2a	U
Don0868	Hoorn	6	16519C,152C,263G,315.1C,524.1A,524.2C	HV*	HV
Don0869	Hoorn	6	16224C,16311C,16519C,73G,189G,195C,263G,315.1C,497T	K1a	U
Don0875	Meppel	5	16224C,16311C,16360T,16519C,73G,146C,152C,263G,309.1C,315.1C	K2a	U
Don0876	Meppel	5	16224C,16269C,16311C,16362C,16519C,73G,195C,263G,309.1C,315.1C,524.1A,524.2C,524.3A,524.4C,573.1C,573.2C	K1d	U
Don0878	Meppel	5	16126C,16362C,16519C,263G,309.1C,315.1C	HV*	HV
Don0879	Meppel	5	16242T,195C,263G,309.1C,315.1C,523d,524d	HV*	HV
Don0880	Meppel	5	16304C,234G,263G,315.1C,456T,523d,524d	H5	HV
Don0882	Meppel	5	16129A,16519C,263G,309.1C,315.1C	HV*	HV
Don0883	Meppel	5	16129A,16519C,263G,309.1C,315.1C	HV*	HV
Don0884	Meppel	5	16170G,16243C,16304C,152C,263G,309.1C,315.1C,456T	H5	HV
Don1009	Woerden	7	16129R,16153A,16189C,16519C,263G,309.1C,309.2C,315.1C	HV*	HV
Don1010	Woerden	7	16293G,16519C,263G,315.1C	HV*	HV
Don1011	Woerden	7	16140C,16183C,16189C,16193.1C,16266A,16519C,73G,210G,263G,309.1C,315.1C,523d,524d	B5a	B5
Don1012	Woerden	7	16354T,263G,315.1C,575T	H2a1	HV
Don1013	Woerden	7	16519C,131C,263G,315.1C	HV*	HV
Don1014	Woerden	7	16192T,16256T,16270T,16362C,16399G,16428A,73G,263G,309.1C,315.1C	U5a1	U
Don1015	Woerden	7	16104T,16126C,16163G,16186T,16189A,16294T,16519C,73G,152C,195C,263G,309.1C,315.1C,573.1C,573.2C	T1a	T
Don1016	Woerden	7	16192T,16311C,73G,150T,263G,315.1C	U5b	U
Don1017	Woerden	7	16232M,16519C,146Y,263G,315.1C	HV*	HV
Don1018	Woerden	7	16069T,16126C,16145A,16231C,16261T,73G,150T,152C,195C,215G,263G,295T,310.1T,315.1C,319C,489C,513A	J2a	J
Don1019	Woerden	7	16069T,16093C,16126C,16145A,16172C,16222T,16261T,73G,242T,263G,295T,315.1C,462T,489C	J1b	J
Don1020	Woerden	7	16126C,16163G,16186T,16189C,16294T,16519C,73G,152C,195C,263G,309.1C,315.1C	T1a	T
Don1021	Woerden	7	16051G,16092Y,16129C,16183C,16189C,16362C,16519C,73G,152C,189G,217C,263G,315.1C,340T,508G,524.1A,524.2C	U2e	U
Don1022	Woerden	7	16126C,16163G,16186T,16189C,16294T,16519C,73G,152C,195C,263G,309.1C,315.1C,394T	T1a	T
Don1023	Woerden	7	16290T,16298C,16519C,72C,263G,309.1C,315.1C	HV0	HV
Don1024	Woerden	7	16343G,16390A,16519C,73G,150T,152C,263G,309.1C,315.1C	U3a	U
Don1025	Woerden	7	16354T,263G,309.1C,315.1C,575T	H2a1	HV
Don1026	Woerden	7	16069T,16126C,16145A,16231C,16261T,16355T,16519C,73G,150T,152C,189G,195C,215G,263G,295T,310.1T,315.1C,319C,430C,489C,513A	J2a	J
Don1027	Woerden	7	16140C,16519C,263G,315.1C	HV*	HV
Don1028	Woerden	7	16069T,16126C,73G,185A,228A,263G,295T,315.1C,462T,489C	J1c	J
Don1030	Purmerend	6	16129A,16224C,16311C,16519C,16T,73G,150T,199C,263G,309.1C,315.1C,497T	K1a	U
Don1034	Purmerend	6	16297C,16519C,72G,152C,263G,315.1C	H74	HV
Don1035	Purmerend	6	16223T,16264T,16278T,16390A,73G,93G,146C,150T,152C,182T,195C,198T,263G,309.1C,315.1C,325T,523d,524d	L2c2	L2c
Don1037	Purmerend	6	16129A,16224C,16311C,16497G,16519C,16T,73G,150T,199C,263G,309.1C,315.1C,497T	K1a	U
Don1038	Purmerend	6	16069T,16126C,16243C,73G,185A,188G,228A,263G,295T,309.1C,315.1C,462T,489C	J1c	J
Don1039	Purmerend	6	16069T,16126C,16222T,73G,228A,263G,295T,315.1C,462T,489C	J1c	J
Don1040	Purmerend	6	16265C,16356C,16519C,73G,146C,152C,195C,263G,309.1C,315.1C,499A,524.1A,524.2C	U4	U
Don1041	Purmerend	6	16126C,16294T,16296T,16519C,73G,263G,315.1C	T2*	T
Don1042	Purmerend	6	16311C,16519C,93G,263G,315.1C	HV*	HV

Don1043	Purmerend	6	16126C,16153A,16223T,16294T,16519C,41T,73G,150T,263G,309.1C,315.1C	T2e	T	
Don1044	Purmerend	6	16093C,16192T,16224C,16311C,16519C,73G,263G,315.1C,497T	K1a	U	
Don1045	Purmerend	6	16519C,263G,315.1C	HV*	HV	
Don1046	Purmerend	6	16298C,72C,263G,309.1C,315.1C	HV0	HV	
Don1048	Kollum	8	16129A,16519C,263G,315.1C	HV*	HV	
Don1049	Kollum	8	16069T,16126C,16145A,16172C,16222T,16261T,73G,146C,242T,263G,295T,315.1C,462T,489C	J1b	J	
Don1050	Kollum	8	16293G,16311C,16325C,16519C,263G,309.1C,315.1C	HV*	HV	
Don1051	Kollum	8	16224C,16311C,16519C,73G,114T,195C,263G,309.1C,315.1C,497T	K1a	U	
Don1055	Tubbergen	10	16519C,204Y,263G,315.1C	HV*	HV	
Don1057	Tubbergen	10	16129R,16293G,16311C,152C,195C,263G,309.1C,315.1C	H11a	HV	
Don1060	Dokkum	8	16126C,16163G,16186T,16189C,16294T,16519C,73G,121C,263G,315.1C,548T	T1a	T	
Don1061	Genemuiden	5	16183C,16189C,16193.1C,16240G,16266C,16662C,151T,263G,315.1C,523d,524d	H1b	HV	
Don1062	Genemuiden	5	16129A,16519C,263G,315.1C	HV*	HV	
Don1063	Genemuiden	5	16192T,16294T,16304C,16320T,263G,309.1C,315.1C,456T	H5	HV	
Don1064	Genemuiden	5	16231C,16256T,16270T,16399G,73G,263G,315.1C,	U5a1	U	
Don1065	Genemuiden	5	16209C,16223T,16256T,16270T,16526A,73G,263G,309.1C,315.1C,	U5a2	U	
Don1066	Genemuiden	5	16189C,16519C,263G,309.1C,309.2C,315.1C	HV*	HV	
Don1067	Genemuiden	5	16126C,16153A,16294T,16519C,41T,73G,150T,263G,309.1C,315.1C	T2e	T	
Don1068	Genemuiden	5	16069T,16126C,16311C,16519C,73G,185A,188G,228A,263G,295T,315.1C,462T,489C,523d,524d	J1c	J	
Don1069	Genemuiden	5	16224C,16311C,16519C,73G,146C,152C,263G,315.1C	K2a	U	
Don1070	Genemuiden	5	16104T,16126C,16294T,16304C,16519C,73G,152C,263G,309.1C,315.1C	T2b	T	
Don1071	Leeuwarden	8	16224C,16311C,16362C,16519C,73G,195C,263G,309.1C,315.1C,524.1A,524.2C,524.3A,524.4C,573.1C,573.2C,573.2C,573.3C	K1d	U	
Don1073	Leeuwarden	8	16263G,16519C,263G,315.1C,477C	H1c	HV	
Don1074	Leeuwarden	8	16069T,16126C,16145A,16231C,16261T,73G,150T,152C,195C,215G,263G,295T,309T,315.1C,319C,489C,513A	J2a	J	
Don1075	Leeuwarden	8	16111T,16174T,16189C,16192T,16270T,16311C,73G,150T,263G,309.1C,315.1C,516T	U5b	U	
Don1076	Leeuwarden	8	16519C,152C,263G,309.1C,315.1C,	HV*	HV	
Don1077	Leeuwarden	8	16311Y,162406,16298C,72C,73G,263G,309.1C,315.1C,	HV0	HV	
Don1078	Leeuwarden	8	16093C,16129A,16224C,16311C,16519C,73G,263G,315.1C,497T,524.1A,524.2C,	K1a	U	
Don1079	Leeuwarden	8	16172C,16192T,16256T,16270T,16301T,16526A,41T,73G,207A,263G,309.1C,315.1C	U5a2	U	
Don1080	Leeuwarden	8	16235G,16291T,16297C,16357C,263G,309.1C,315.1C,523d,524d,	H2a2b	HV	
Don1081	Dokkum	8	16162G,16209C,16519C,73G,263G,309.1C,315.1C,	H1a	HV	
Don1082	Dokkum	8	16224C,16239G,16311C,16519C,73G,146C,152C,263G,309.1C,315.1C	K2a	U	
Don1084	Dokkum	8	16126C,16163G,16186T,16189C,16294T,16519C,73G,152C,195C,263G,309.1C,315.1C	T1a	T	
Don1085	Dokkum	8	16162G,16209C,16519C,73G,263G,315.1C	H1a	HV	
Don1086	Dokkum	8	16239G,16519C,263G,315.1C,481T,	HV*	HV	
Don1087	Dokkum	8	16162G,16209C,16519C,73G,263G,315.1C	H1a	HV	
Don1088	Dokkum	8	16270T,315.1C	H1ba	HV	
Don1089	Dokkum	8	16256T,16270T,16399G,73G,152C,195C,263G,309.1C,315.1C,524.1A,524.2C,	U5a1	U	
Don1090	Dokkum	8	16038G,16316G,16519C,189C,263G,309.1C,315.1C,	HV*	HV	
Don1091	Dokkum	8	16192T,16270T,16304C,73G,150T,263G,315.1C,	U5b	U	
Don1119	Heerhugowaard	6	16126C,16163G,16186T,16189C,16294T,16519C,73G,152C,195C,263G,309.1C,315.1C	T1a	T	
Don1121	Heerhugowaard	6	16104T,16126C,16294T,16304C,16519C,73G,152C,263G,315.1C,	T2b	T	
Don1122	Heerhugowaard	6	16224C,16311C,16519C,73G,150T,195C,263G,315.1C,497T,524.1A,524.2C,533G	K1a	U	
Don1123	Heerhugowaard	6	16069T,16126C,16519C,73G,185A,188G,228A,263G,295T,309.1C,315.1C,462T,489C	J1c	J	
Don1125	Heerhugowaard	6	16154C,16311C,263G,309.1C,309.2C,315.1C,455.1T,	HV*	HV	
Don1126	Heerhugowaard	6	16146R,16298C,16318T,16526R,72C,195C,204C,263G,315.1C	HV0	HV	
Don1127	Heerhugowaard	6	16311C,16519C,93G,263G,309.1C,315.1C	HV*	HV	
Don1130	Heerhugowaard	6	16519C,263G,309.1C,315.1C,477C	H1c	HV	
Don1131	Heerhugowaard	6	195C,263G,315.1C,523d,524d	HV*	HV	
Don1132	Heerhugowaard	6	16126C,16163G,16186T,16189C,16294T,16519C,73G,152C,195C,263G,309.1C,315.1C	T1a	T	

Don1133	Heerhugowaard	6	16224C,16311C,16362C,16519C,73G,195C,263G,315.1C,524.1A,524.2C,524.3A,524.4C,573.1C,573.2C	K1d	U	
Don1134	Heerhugowaard	6	16092C,16140C,16293G,16311C,195C,263G,315.1C	H11a	HV	
Don1135	Heerhugowaard	6	16179T,16336C,16519C,73G,150T,195C,263G,315.1C,499A,524.1A,524.2C,524.3A,524.4C	U4	U	
Don1136	Heerhugowaard	6	16261T,16294T,16519C,200G,263G,309.1C,315.1C	HV*	HV	
Don1172	Schagen	6	16299G,16519C,200G,263G,309.1C,309.2C,315.1C	HV*	HV	
Don1173	Schagen	6	16293G,16519C,263G,309.1C,309.2C,315.1C	HV*	HV	
Don1175	Schagen	6	16519C,72G,152C,263G,309.1C,315.1C	H74	HV	
Don1176	Schagen	6	16293G,16519C,263G,309.1C,309.2C,315.1C	HV*	HV	
Don1177	Schagen	6	16114T,16223T,16295T,16362C,16497G,16519C,73G,146C,199C,263G,309.1C,309.2C,315.1C,489C,523d,524d	M7c1	M7c	
Don1178	Schagen	6	16069T,16126C,16527T,73G,228A,263G,295T,309.1C,315.1C,462T,488C	J1c	J	
Don1180	Schagen	6	16519C,263G,309.1C,315.1C	HV*	HV	
Don1181	Schagen	6	16104T,16126C,16294T,16304C,16519C,73G,152C,263G,309.1C,315.1C	T2b	T	
Don1182	Schagen	6	16215G,16224C,16311C,16388C,73G,263G,309.1C,315.1C,385G,497T	K1a	U	
Don1183	Schagen	6	16224C,16256T,16311C,16519C,73G,146C,152C,263G,309.1C,315.1C	K2a	U	
Don1184	Schagen	6	16126C,16163G,16186T,16189C,16294T,16519C,73G,152C,195C,263G,309.1C,315.1C	T1a	T	
Don1185	Schagen	6	16293G,16311C,152C,195C,263G,309.1C,315.1C	H11a	HV	
Don1186	Schagen	6	16069T,16126C,16222T,73G,185A,228A,263G,295T,315.1C,462T,488C	J1c	J	
Don1188	Schagen	6	16051G,16162G,16466T,16519C,73G,195C,263G,315.1C,372C,523d,524d	H1a	HV	
Don1189	Schagen	6	16093C,16263C,16519C,263G,315.1C,477C	H1c	HV	
Don1190	Maarssenbroek	7	16126C,16163G,16186T,16189C,16274A,16294T,16519C,73G,199C,263G,309.1C,315.1C,512G	T1a	T	
Don1191	Maarssenbroek	7	16311C,16519C,263G,309.1C,315.1C	HV*	HV	
Don1192	Maarssenbroek	7	16311C,16519C,263G,315.1C	HV*	HV	
Don1193	Maarssenbroek	7	16223T,16292T,16519C,73G,189G,194T,195C,203A,204C,207A,263G,309.1C,315.1C,523d,524d	W	W	
Don1194	Maarssenbroek	7	16231C,16256T,16270T,16399G,73G,152C,263G,315.1C	U5a1	U	
Don1195	Maarssenbroek	7	16224C,16311C,16519C,73G,146C,152C,263G,315.1C	K2a	U	
Don1196	Maarssenbroek	7	16304C,152C,263G,315.1C,340T,456T	H5	HV	
Don1197	Maarssenbroek	7	16168T,16192T,16256T,16270T,16526A,73G,150T,263G,309.1C,315.1C,455d	U5a2	U	
Don1199	Maarssenbroek	7	16519C,73G,263G,315.1C	undetermined	undetermined	
Don1201	Maarssenbroek	7	16224C,16311C,16519C,73G,146C,152C,204Y,263G,309.1C,315.1C	K2a	U	
Don1203	Maarssenbroek	7	16203G,16293G,16311C,195C,207A,263G,309.1C,309.2C,315.1C,573.1C	H11a	HV	
Don1204	Maarssenbroek	7	16093C,16224C,16311C,16519C,73G,114T,152C,263G,309.1C,315.1C,497T	K1a	U	
Don1205	Maarssenbroek	7	16092C,16126C,16153A,16294T,16519C,41T,73G,150T,263G,309.1C,315.1C	T2e	T	
Don1206	Maarssenbroek	7	16223T,16292T,16362C,16519C,73G,189G,194T,195C,204C,207A,263G,309.1C,315.1C	W	W	
Don1207	Maarssenbroek	7	16126C,16189C,16294T,16296T,16390R,16519C,73G,263G,309.1C,315.1C,	T2*	T	
Don1264	Leeuwarden	8	16069T,16126C,16256T,16519C,73G,185A,188G,198T,228A,263G,295T,309.1C,315.1C,462T,488C	J1c	J	
Don1265	Leeuwarden	8	16224C,16311C,16362C,16519C,73G,152C,195C,263G,315.1C,497T	K1a	U	
Don1266	Leeuwarden	8	16224C,16311C,16362C,16519C,73G,195C,263G,309.1C,315.1C,524.1A,524.2C,524.3A,524.4C,573.1C,573.2C,573.3C	K1d	U	
Don1271	Drachten	8	16224C,16311C,16519C,73G,146C,152C,263G,309.1C,309.2C,315.1C,324T,524.1A,524.2C	K2a	U	
Don1273	Drachten	8	16162G,16209C,16291T,16519C,73G,263G,315.1C	H1a	HV	
Don1283	Hoogeveen	5	16356C,16362C,16519C,73G,195C,263G,315.1C,499A	U4	U	
Don1308	Hoogeveen	5	16093C,16223T,16519C,58C,60.1T,60.2T,65d,71d,73G,189G,194T,204C,207A,263G,315.1C	W	W	
Don1313	Drachten	8	16069T,16126C,73G,185A,189G,263G,295T,309.1C,315.1C,462T,488C	J1c	J	
Don1316	Harlingen	8	16265C,16356C,16519C,73G,146C,152C,195C,263G,309.1C,315.1C,499A,524.1A,524.2C,524.3A,524.4C	U4	U	
Don1317	Harlingen	8	16256T,16270T,16399G,73G,263G,315.1C	U5a1	U	
Don1318	Harlingen	8	16224C,16311C,16519C,73G,146C,152C,263G,309.1C,309.2C,315.1C,324T,524.1A,524.2C	K2a	U	
Don1319	Harlingen	8	16293G,16311C,195C,204Y,207A,263G,315.1C	H11a	HV	
Don1320	Drachten	8	16216G,16269C,16291T,16519C,73G,263G,315.1C	HV0	HV	
Don1322	Harlingen	8	16162G,16209C,16291T,16519C,73G,263G,315.1C	H1a	HV	
Don1326	Harlingen	8	16192T,16256T,16270T,16399G,73G,263G,315.1C	U5a1	U	
Don1328	Harlingen	8	16126C,16294T,16296T,16304C,16519C,73G,151T,195C,263G,315.1C	T2b	T	

Don1329	Harlingen	8	16224C,16311C,16519C,73G,263G,309.1C,315.1C,497T			K1a	U
Don1334	Leeuwarden	8	16093C,16311C,16519C,56.1C,263G,315.1C			HV*	HV
Don1357	Elde	9	16063C,16069T,16126C,73G,228A,263G,295T,315.1C,462T,489C			J1c	J
Don1358	Dokkum	8	16162G,16209C,16519C,73G,263G,315.1C			H1a	HV
Don1361	Delfzijl	9	16069T,16126C,73G,228A,263G,295T,309.1C,315.1C,462T,489C			J1c	J
Don1362	Delfzijl	9	16069T,16126C,73G,150T,185A,263G,295T,309.1C,315.1C,462T,489C			J1c	J
Don1364	Delfzijl	9	16192T,16256T,16270T,16399G,73G,263G,309.1C,315.1C			U5a1	U
Don1365	Delfzijl	9	16293G,16311C,152C,195C,263G,309.1C,315.1C,			H11a	HV
Don1366	Delfzijl	9	16069T,16092Y,16126C,16261T,73G,146C,185A,228A,263G,295T,315.1C,462T,489C			J1c	J
Don1384	Schagen	6	16224C,16311C,16519C,73G,146C,152C,263G,309.1C,315.1C,498d			K1c	U
Don1385	Heemskerk	6	16311C,16519C,93G,263G,315.1C			HV*	HV
Don1386	Heerhugowaard	6	16093C,16221T,16519C,263G,309.1C,315.1C			HV*	HV
Don1388	Heerhugowaard	6	16220G,16519C,195C,263G,309.1C,315.1C,523d,524d			H82	HV
Don1389	Heerhugowaard	6	16239T,16298C,16399G,263G,315.1C			HV0	HV
Don1391	Heemskerk	6	16274A,16304C,16362C,263G,315.1C,456T,523d,524d			H5	HV
Don1416	Mijdrecht	7	16304C,263G,315.1C,456T,523d,524d			H5	HV
Don1417	Maarssenbroek	7	16162G,16209C,16519C,73G,263G,315.1C			H1a	HV
Don1418	Maarssenbroek	7	16304C,16342C,263G,315.1C,456T,523d,524d			H5	HV
Don1421	Purmerend	6	16129A,16316G,16519C,263G,315.1C			H27	HV
Don1424	Emmen	9	16256T,16270T,16399G,73G,263G,315.1C			U5a1	U
Don1425	Emmen	9	16093C,16218Y,16519C,263G,309.1C,315.1C,408A			H3v	HV
Don1433	Delfzijl	9	16093C,16224C,16311C,16519C,73G,114T,263G,315.1C,497T			K1a	U
Don1451	Assen	9	16224C,16311C,16519C,73G,263G,315.1C,524.1A,524.2C,524.3A,524.4C			K*	U
Don1460	Assen	9	16311C,16519C,73G,263G,309.1C,315.1C			undetermined	undetermined
Don1461	Assen	9	16126C,16294T,16296T,16304C,16519C,73G,263G,315.1C			T2b	T
Don1473	Veendam	9	16129A,16224C,16311C,16519C,167,73G,150T,199C,263G,309.1C,315.1C,497T			K1a	U
Don1474	Veendam	9	16146G,16519C,131C,263G,315.1C			HV*	HV
Don1478	Emmen	9	16069T,16126C,16145A,16231C,16261T,73G,150T,152C,195C,215G,263G,295T,310.1T,315.1C,319C,489C,513A			J2a	J
Don1479	Emmen	9	16192T,16311C,73G,150T,263G,315.1C			U5b	U
Don1482	Steenwijk	5	16519C,263G,315.1C			HV*	HV
Don1483	Steenwijk	5	16362C,16482G,239C,263G,309.1C,315.1C			H6	HV
Don1484	Steenwijk	5	16183M,16189C,16261T,16270T,16304C,16398A,73G,150T,210G,263G,315.1C			U5b	U
Don1485	Steenwijk	5	16129A,16519C,263G,309.1C,315.1C			HV*	HV
Don1496	Emmen	9	16093C,16167T,16179T,73G,263G,282C,309.1C,315.1C			U8a	U
Don1499	Hoogeveen	5	16265G,16356C,16362C,16519C,73G,195C,247A,263G,315.1C,499A,524.1A,524.2C,524.3A,524.4C			U4	U
Don1501	Hoogeveen	5	16224C,16311C,16519C,73G,146C,152C,263G,315.1C			K2a	U
Don1511	Emmen	9	16093C,16224C,16234T,16311C,16519C,73G,263G,309.1C,315.1C,497T,524.1A,524.2C			K1a	U
Don1515	Drachten	8	16189C,16192T,16270T,16398A,73G,150T,263G,309.1C,315.1C			U5b	U
Don1533	Kampen	5	16126C,16294T,16296T,16519C,73G,263G,315.1C			T2*	T
Don1534	Kampen	5	16224C,16311C,16519C,73G,150T,195C,263G,315.1C,497T,524.1A,524.2C			K1a	U
Don1535	Kampen	5	16189C,16261T,16270T,16304C,16398A,73G,150T,210G,263G,315.1C			U5b	U
Don1536	Kampen	5	16104T,16362Y,16519C,152C,249d,263G,309.1C,315.1C,477C			H1c	HV
Don1537	Kampen	5	16304C,263G,315.1C,456T,523d,524d			H5	HV
Don1538	Kampen	5	16093Y,16224C,16311C,16519C,73G,114T,152C,263G,315.1C,497T,524.1A,524.2C			K1a	U
Don1540	Assen	9	16093Y,16184T,16519C,263G,309.1C,315.1C			HV*	HV
Don1541	Assen	9	16126C,16294T,16296T,16304C,16519C,73G,263G,309.1C,315.1C,458T			T2b	T
Don1542	Assen	9	16145A,16184T,16224C,16311C,16519C,73G,150T,195C,263G,315.1C,497T,523d,524d			K1a	U
Don1543	Assen	9	16298C,72C,263G,309.1C,315.1C			HV0	HV
Don1550	Genemuiden	5	16223T,16519C,150T,152C,263G,315.1C			undetermined	undetermined
Don1551	Genemuiden	5	16069T,16126C,16145A,16231C,16261T,73G,150T,152C,195C,215G,295T,310.1T,315.1C,319C,489C,513A			J2a	J

Don1552	Hoogeveen	5	16224C,16245T,16311C,16519C,73G,263G,309.1C,315.1C,497T,524.1A,524.2C,524.3A,524.4C,573.1C,573.2C,573.3C,573.4C,573.5C	K1a	U	
Don1553	Beilen	9	16092C,16140C,16293G,16311C,195C,263G,315.1C	H11a	HV	
Don1554	Beilen	9	16382C,16482G,23C,263G,309.1C,315.1C	H6	HV	
Don1556	Beilen	9	16183C,16189C,16223T,16235A,16278T,16519C,73G,153G,195C,200G,225A,227G,263G,315.1C	X2c1c	X	
Don1557	Beilen	9	16519C,152C,263G,315.1C	HV*	HV	
Don1558	Beilen	9	16343G,16390A,16519C,73G,150T,263G,315.1C	U3a	U	
Don1559	Beilen	9	16356C,73G,146C,195C,228A,263G,315.1C,499A	U4	U	
Don1560	Beilen	9	16362C,16519C,263G,309.1C,315.1C	HV*	HV	
Don1599	Meppel	5	16069T,16126C,16256T,16519C,73G,185A,188G,198T,228A,263G,295T,309.1C,315.1C,462T,489C	J1c	J	
Don1600	Meppel	5	16092C,16140C,16293G,16311C,152C,195C,263G,315.1C	H11a	HV	
Don1601	Meppel	5	195C,263G,309.1C,309.2C,315.1C,523d,524d	HV*	HV	
Don1602	Meppel	5	16126C,16294T,16296T,16304C,16519C,73G,263G,309.1C,315.1C,321C	T2b	T	
Don1603	Meppel	5	16093C,16221T,16519C,263G,309.1C,309.2C,315.1C	HV*	HV	
Don1612	Veendam	9	16093C,16189C,16224C,16311C,16519C,73G,152C,263G,315.1C,497T,524.1A,524.2C	K1a	U	
Don1613	Veendam	9	16362C,16482G,23C,263G,309.1C,315.1C	H6	HV	
Don1634	Winschoten	9	16104T,16126C,16294T,16304C,16519C,73G,152C,263G,309.1C,315.1C	T2b	T	
Don1636	Winschoten	9	16519C,263G,315.1C	HV*	HV	
Don1637	Winschoten	9	16519C,263G,315.1C,477C,524.1A,524.2C	H1c	HV	
Don1639	Hoogeveen	5	16519C,263G,309.1C,315.1C	HV*	HV	
Don1640	Hoogeveen	5	16224C,16311C,16360T,16519C,73G,146C,152C,263G,309.1C,315.1C	K2a	U	
Don1648	Eelde	9	16093C,16189C,16270T,73G,150T,263G,315.1C	U5b	U	
Don1649	Eelde	9	16093C,16129A,16189C,16270T,73G,150T,263G,315.1C	U5b	U	
Don1650	Eelde	9	16304C,263G,315.1C,456T,523d,524d	H5	HV	
Don1651	Eelde	9	16224C,16293G,16311C,16519C,73G,146C,152C,263G,309.1C,315.1C	K2a	U	
Don1652	Eelde	9	16260T,16519C,189G,263G,315.1C	HV*	HV	
Don1653	Eelde	9	16326G,16519C,73G,263G,315.1C	undetermined	undetermined	
Don1659	Winschoten	9	16126C,16163G,16186T,16189C,16294T,16519C,73G,152C,195C,263G,309.1C,315.1C,489T	T1a	T	
Don1666	Winschoten	9	16216G,16261T,16298C,16362C,16519C,44.1C,72C,263G,309.1C,315.1C	HV0	HV	
Don1667	Winschoten	9	152C,263G,315.1C	HV*	HV	
Don1668	Winschoten	9	16134T,16356C,16519C,73G,152C,195C,263G,296T,315.1C,499A	U4	U	
Don1683	Denekamp	10	16104T,16126C,16294T,16304C,16519C,73G,152C,263G,309.1C,315.1C	T2b4b	T	
Don1685	Denekamp	10	16224C,16245T,16311C,16519C,73G,146C,263G,309.1C,315.1C,497T,524.1A,524.2C	K1a4a1a2b	U	
Don1686	Denekamp	10	16126C,16294T,16296T,16362C,16519C,73G,263G,315.1C	T2i	T	
Don1687	Denekamp	10	16224C,16311C,16519C,73G,146C,152C,263G,315.1C	K2a	U	
Don1690	Denekamp	10	16126C,16294T,16296T,16519C,73G,263G,309.1C,315.1C	T2*	T	
Don1693	Winschoten	9	195C,263G,315.1C,523d,524d	HV*	HV	
Don1694	Emmen	9	16311C,263G,315.1C	HV*	HV	
Don1704	Kollum	8	16126C,16294T,16296T,16304C,16519C,73G,263G,315.1C	T2b	T	
Don1705	Kollum	8	16069T,16126C,16222T,73G,185A,228A,263G,295T,315.1C,462T,489C	J1c	J	
Don1706	Kollum	8	16069T,16126C,16261T,73G,185A,228A,263G,295T,315.1C,462T,489C	J1c	J	
Don1707	Kollum	8	16069T,16126C,16311C,73G,228A,263G,295T,309.1C,315.1C,462T,489C,544T	J1c	J	
Don1709	Kollum	8	16192T,16256T,16270T,16526A,73G,263G,315.1C	U5a2	U	
Don1713	Hoogeveen	5	16069T,16126C,16256T,16519C,73G,185A,188G,228A,263G,295T,309.1C,315.1C,462T,489C	J1c	J	
Don1714	Hoogeveen	5	16104T,16126C,16294T,16304C,16519C,73G,152C,263G,309.1C,315.1C	T2b	T	
Don1716	Meppel	5	16224C,16290Y,16297N,16311C,16519C,73G,146C,152C,263G,315.1C	K2a	U	
Don1717	Hollum	8	16162G,16209C,16234T,16519C,73G,263G,309.1C,315.1C	H1a	HV	
Don1718	Hollum	8	16293G,16311C,195C,207A,263G,309.1C,315.1C	H11a	HV	
Don1719	Hollum	8	16519C,263G,315.1C	HV*	HV	
Don1722	Ameland	8	16362C,16519C,146C,195C,263G,309.1C,315.1C	HV*	HV	
Don1723	Borger	9	16224C,16311C,16360T,16519C,73G,146C,152C,263G,309.1C,315.1C	K2a	U	

Don1724	Markelo	10	16051G,16162G,16172C,16213A,16286T,16519C,73G,263G,315.1C	H1a	HV
Don1725	Hollum	8	16069T,16092C,16126C,16261T,73G,185A,263G,295T,309.1C,315.1C,462T,489C	J1c	J
Don1726	Borger	9	16126C,16294T,16304C,16519C,73G,152C,263G,315.1C	T2b	T
Don1727	Hollum	8	16126C,16163G,16186T,16189C,16292T,16294T,16362Y,16519C,73G,152C,195C,259R,263G,315.1C	T1a	T
Don1728	Hollum	8	16291Y,16519C,131C,263G,315.1C	HV*	HV
Don1731	Borger	9	16114T,16189C,16356C,16519C,263G,315.1C,523d,524d	H1b1f	HV
Don1734	Tubbergen	10	16069T,16126C,16145A,16231C,16261T,16265C,16278T,16311C,73G,150T,152C,195C,215G,263G,295T,315.1C,319C,489C,513A,524.1A,524.2C	J2a	J
Don1736	Tubbergen	10	16299G,16519C,263G,315.1C	HV*	HV
Don1738	Tubbergen	10	no difference to rCRS	H2a2a1	HV
Don1741	Tubbergen	10	16224C,16245T,16311C,16519C,73G,146C,263G,309.1C,315.1C,497T	K1a4a1a2b	U
Don1743	Tubbergen	10	16129A,16519C,263G,293C,309.1C,315.1C,493G	H3z	HV
Don1744	Tubbergen	10	16126C,16294T,16296T,16519C,73G,263G,315.1C	T2*	T
Don1745	Haringen	8	16304C,16327T,263G,309.1C,315.1C,456T	H5	HV
Don1746	Haringen	8	16162G,16209C,16519C,73G,263G,315.1C	H1a	HV
Don1747	Haringen	8	16256T,16270T,16304C,16526A,73G,263G,315.1C	U5a2	U
Don1781	Dokkum	8	16169T,16519C,263G,315.1C	HV*	HV
Don1782	Dokkum	8	16063C,16069T,16126C,73G,228A,263G,295T,309.1C,315.1C,462T,489C	J1c	J
Don1783	Dokkum	8	16051G,16162G,16519C,73G,263G,315.1C,444G	H1a	HV
Don1784	Dokkum	8	16304C,16399G,16497G,16519C,152C,263G,315.1C,340T,456T	H5	HV
Don1785	Dokkum	8	16311C,16519C,263G,309.1C,315.1C	HV*	HV
Don1786	Dokkum	8	16176T,16214T,16219G,16519C,2576,263G,315.1C,477C	H1c	HV
Don1787	Dokkum	8	16298G,16519C,152Y,263G,315.1C	HV*	HV
Don1788	Dokkum	8	16311C,16519C,263G,315.1C	HV*	HV
Don1850	Hoogeveen	5	16519C,72C,263G,315.1C	HV*	HV
Don1851	Hoogeveen	5	16183C,16189C,16325C,16566A,73G,150T,152C,263G,315.1C,	U5b	U
Don1852	Hoogeveen	5	16192T,16256T,16270T,16294T,16391A,16526A,64T,73G,93G,185A,263G,309.1C,315.1C,524.1,524d	U5a2	U
Don1853	Hoogeveen	5	16256T,16270T,16362C,16399G,73G,185A,204C,263G,309.1C,315.1C	U5a1	U
Don1855	Hoogeveen	5	16145A,16261T,16519C,263G,309.1C,315.1C	HV*	HV
Don1857	Hoogeveen	5	16129A,16519C,73G,263G,309.1C,315.1C	undetermined	undetermined
Don1858	Hoogeveen	5	16126C,16294T,16296T,16304C,16519C,73G,263G,315.1C	T2b	T
Don1859	Hoogeveen	5	16224C,16311C,16360T,16519C,73G,146C,152C,263G,309.1C,315.1C	K2a	U
Don1860	Hoogeveen	5	263G,315.1C	HV*	HV
Don1870	Steenwijk	5	16069T,16126C,16145A,16231C,16261T,73G,150T,152C,195C,215G,263G,295T,310.1T,315.1C,319C,489C,513A	J2a	J
Don1871	Steenwijk	5	16069T,16126C,16519C,73G,185A,188G,228A,263G,295T,315.1C,462T,489C	J1c	J
Don1872	Steenwijk	5	16051G,16129C,16188C,16189C,16193.1C,16362C,16519C,73G,152C,217C,263G,309.1C,315.1C,340T,508G,524.1A,524.2C	U2e	U
Don1873	Steenwijk	5	16192T,16256T,16270T,16399G,73G,263G,309.1C,315.1C	U5a1	U
Don1874	Steenwijk	5	93G,146C,195C,263G,315.1C,523d,524d	HV*	HV
Don1875	Steenwijk	5	16288R,16519C,263G,315.1C	HV*	HV
Don1876	Midsland	8	16162G,16209C,16519C,73G,263G,315.1C	H1a	HV
Don1877	Midsland	8	16519C,263G,309.1C,315.1C	HV*	HV
Don1878	Midsland	8	16162G,16209C,16311C,16519C,73G,263G,315.1C	H1a	HV
Don1879	Midsland	8	16093C,16129A,16316G,16519C,185A,263G,309.1C,315.1C,523d,524d	H27	HV
Don1880	Midsland	8	16192T,16256T,16270T,16526A,73G,263G,309.1C,315.1C,524.1A,524.2C	U5a2	U
Don1881	Midsland	8	16126C,16294T,16296T,16519C,73G,146Y,263G,309.1C,315.1C	T2*	T
Don1884	Midsland	8	16209C,16213R,16304C,16519C,263G,315.1C,456T,523d,524d	H5	HV
Don1885	Midsland	8	16519C,152C,263G,315.1C	HV*	HV
Don1887	Midsland	8	16162G,16209C,16311C,16519C,73G,263G,315.1C	H1a	HV
Don1888	Midsland	8	16174T,16188C,16192T,16270T,16311C,73G,150T,195C,263G,315.1C,516T	U5b	U
Don1889	Midsland	8	16519C,152C,263G,309.1C,309.2C,315.1C	HV*	HV
Don1890	Midsland	8	16126C,16294T,16304C,16519C,73G,152C,263G,309.1C,315.1C	T2b	T

Don1892	Midsland	8	16051G,16162G,16172C,16213A,16266T,16519C,73G,263G,315.1C	H1a	HV
Don1948	Kampen	5	16093C,16111A,16126C,16294T,16296T,16519C,73G,263G,315.1C	T2*	T
Don1949	Kampen	5	16129A,16223T,16278T,16311C,16391A,16519C,73G,199C,204C,250C,263G,309.1C,315.1C,455.1T,524.1A,524.2C,573.1C,573.2C,573.3C,573.4C	I1	N1
Don1951	Kampen	5	16092C,16209C,16256T,16270T,16399G,73G,153G,263G,309.1C,315.1C	U5a1	U
Don1952	Kampen	5	16069T,16126C,16519C,73G,185A,188G,228A,263G,295T,309.1C,315.1C,462T,489C,523d,524d	J1c	J
Don1953	Kampen	5	16069T,16126C,16145A,16231C,16261T,16299G,73G,150T,152C,195C,263G,295T,310.1T,315.1C,319C,489C,513A	J2a	J
Don1954	Kampen	5	16093C,16224C,16311C,16519C,73G,195C,263G,315.1C,497T,524.1A,524.2C,524.3A,524.4C	K1a	U
Don1965	Meppel	5	16069T,16126C,73G,185A,195C,228A,263G,295T,315.1C,462T,489C	J1c	J
Don1966	Meppel	5	16126C,16163G,16186T,16189C,16294T,16519C,73G,152C,195C,263G,309.1C,309.2C,315.1C	T1a	T
Don2009	Kampen	5	16242T,195C,263G,315.1C,523d,524d	HV*	HV
Don2010	Kampen	5	16069T,16126C,16222T,73G,185A,228A,263G,295T,315.1C,462T,489C	J1c	J
Don2011	Kampen	5	16183C,16189C,16193.1C,16519C,263G,315.1C	HV*	HV
Don2012	Kampen	5	16140C,16293G,16311C,195C,263G,315.1C,460C	H11a	HV
Don2014	Kampen	5	16145A,16193T,16362C,73G,263G,309.1C,315.1C,523d,524d	undetermined	undetermined
Don2015	Kampen	5	16093Y,16189C,16192T,16256T,16270T,16362C,16399G,16428A,73G,263G,315.1C	U5a1	U
Don2016	Kampen	5	16126C,16292T,16294T,16519C,73G,146C,263G,279C,309.1C,315.1C	T2c	T
Don2054	Markelo	10	16126C,16292T,16294T,16519C,73G,146C,263G,279C,309.1C,315.1C	T2c	T
Don2055	Markelo	10	16311C,16319A,16519C,131C,152C,263G,309.1C,315.1C	HV9a	HV
Don2056	Markelo	10	16189C,16192T,16270T,16324C,16398A,73G,150T,152Y,263G,315.1C	U5b	U
Don2057	Markelo	10	16209C,64T,263G,315.1C	HV*	HV
Don2058	Markelo	10	16093C,16129A,16189C,16270T,73G,150T,263G,315.1C	U5b	U
Don2059	Markelo	10	16070G,16304C,93G,152C,263G,309.1C,315.1C,456T,524.1A,524.2C	H36	HV
Don2061	Markelo	10	16126C,16163G,16186T,16189C,16294T,16519C,73G,152C,195C,263G,309.1C,315.1C,469T	T1a	T
Don2062	Markelo	10	16218Y,73G,204Y,263G,315.1C,523d,524d	undetermined	undetermined
Don2063	Markelo	10	16092C,16140C,16293G,16311C,195C,263G,315.1C	H11a	HV
Don2064	Markelo	10	16224C,16311C,16519C,73G,263G,309.1C,315.1C,497T,524.1A,524.2C	K1a	U
Don2066	Markelo	10	16069T,16126C,16145A,16261T,16278T,16287T,16519C,73G,263G,271T,295T,309.1C,315.1C,462T,489C,523d,524d	J1b	J
Don2068	Markelo	10	16051G,16293G,16519C,200R,263G,309.1C,309.2C,315.1C	H24	HV
Don2080	Genemuiden	5	16224C,16311C,16519C,73G,146C,152C,217Y,263G,309.1C,315.1C,324T,524.1A,524.2C	K2a	U
Don2091	Genemuiden	5	16093C,16189C,16270T,16519C,73G,150T,151T,263G,315.1C	U5b	U
Don2096	Nes	8	16069T,16126C,16222T,73G,185A,228A,263G,295T,315.1C,462T,489C	J1c	J
Don2099	Nes	8	16126C,16274A,16294T,16296T,16304C,16519C,73G,263G,309.1C,315.1C	T2b	T
Don2100	Nes	8	152C,263G,309.1C,315.1C	HV*	HV

Supplementary Table 1: Control-region profiles of 680 samples from the Netherlands. Variant positions from the rCRS are shown for nucleotide positions 16024 – 16569 and 1 – 576. Haplogroup nomenclature is according to PhyloTree build 16 [12].

Chapter 5.3

Simultaneous Whole Mitochondrial Genome Sequencing with Short Overlapping Amplicons Suitable for Degraded DNA Using the Ion Torrent Personal Genome Machine

Human Mutation 36 (2015) 1236 – 1247

Abstract

Whole mitochondrial (mt) genome analysis enables a considerable increase in analysis throughput, and improves the discriminatory power to the maximum possible phylogenetic resolution. Most established protocols on the different massively parallel sequencing (MPS) platforms, however, invariably involve the PCR amplification of large fragments, typically several kilobases in size, which may fail due to mtDNA fragmentation in the available degraded materials. We introduce a MPS tiling approach for simultaneous whole human mt genome sequencing using 161 short overlapping amplicons (average 200 bp) with the Ion Torrent Personal Genome Machine. We illustrate the performance of this new method by sequencing 20 DNA samples belonging to different worldwide mtDNA haplogroups. Additional quality control, particularly regarding the potential detection of nuclear insertions of mtDNA (NUMTs), was performed by comparative MPS analysis using the conventional long-range amplification method. Preliminary sensitivity testing revealed that detailed haplogroup inference was feasible with 100 pg genomic input DNA. Complete mtgenome coverage was achieved from DNA samples experimentally degraded down to genomic fragment sizes of about 220 bp, and up to 90% coverage from naturally degraded samples. Overall, we introduce a new approach for whole mt genome MPS analysis from degraded and non-degraded materials relevant to resolve and infer maternal genetic ancestry at complete resolution in anthropological, evolutionary, medical, and forensic applications.

Introduction

Previous years have witnessed conspicuous progress in the establishment of the high-copy-number human mitochondrial DNA (mtDNA) as an imperative tool in forensic, anthropological, and medical genetics. This small, circular, double-stranded genome has enthralled fundamental and applied geneticists with its unique features, bolstering its recognized value as a useful molecular marker when analyzing degraded or low-copy-number DNA samples, as often confronted at crime scenes, from old and ancient human remains, and from limited samples available in medical and anthropological applications. The higher mutation rate of mtDNA as compared with nuclear DNA, and the resulting significant sequence variability between maternally unrelated individuals, has made mtDNA an indispensable tool in various applications particularly forensics and anthropology [1-4]. Most applications employ traditional Sanger sequencing of (the hypervariable portions of) the noncoding control region of themt genome to delineate haplotypes that enable maternal lineage identification [e.g., 1,3,5-7]. In addition to being labor-intensive and expensive, the control-region Sanger sequencing approach does not always allow reliable inference of maternal haplogroups, due to the high level of homoplasy in the control region that can obscure phylogenetic signatures [8,9] as well as the fact that many haplogroup-defining variants are located outside the control region. Since not all the major mtDNA haplogroups can be distinguished through control-region sequencing only, several studies [8,10,11] have resorted to the use of multiplex single-base primer extension assays for simultaneous genotyping of a limited number of mtDNA coding-region SNPs to improve maternal haplogroup definition. However, such assays are hampered by technical limitations in terms of multiplexing capacity allowing not more than 20–40 SNPs in a single multiplex assay.

Whole mtDNA genome sequencing would augment the haplotype and haplogroup resolution to themaximumpossible resolution level, thereby greatly enhancing the discriminatory power, and allowing inference of matrilineal biogeographic ancestry at a greater resolution; however, whole mtDNA genome analysis via Sanger sequencing is highly labor-intensive given the >16 kb involved [12,13]. Massively parallel sequencing (MPS) or next-generation sequencing (NGS) technology in principle provides a solution to simultaneous whole mtDNA genome analysis allowing high-throughput analysis at

reduced per-sample costs. It had been pioneered in the field of ancient DNA analysis where degradation problems are most severe. The shotgun-sequencing approach usingMPS technology has facilitated the complete mtDNA genome analysis of extinct non-human species such as mammoths [14,15], cave bears, and others [16-20] as well as extinct human species such as Neanderthals [21]. Recently, Davis et al. [22] used the MPS approach to analyze the hypervariable segments (HVS)-I and II of human mtDNA using short amplicons in tissue and bone samples keeping with the previously discussed limitations of partialmtDNA analysis. However, themost established currently available MPS protocols for whole mtDNA genome sequencing are based on PCR amplification of large mtDNA fragments, typically several kilobases in size [23,24], which fail when applied to degraded DNA as encountered in many mtDNA applications. A very recent publication [25] described a midi-sized amplicon approach for whole mtDNA MPS analysis using 62 PCR amplicons of 300–500 bp (average about 380 bp) in two multiplex assays, which proved useful for human hair analysis. However, many DNA samples encountered in mtDNA testing, especially for forensic and anthropological purposes are more severely degraded resulting in smaller-sized fragments.

Recently, we introduced a MPS tool based on short amplicons (203 bp on average) using the Ion Torrent Personal Genome Machine (PGM) for simultaneous analysis of >530 Y-chromosome SNPs covering the entire phylogenetic Y-chromosome tree that allows classification of Y chromosomes into >430 worldwide Y haplogroups for ultra-high-resolution paternal lineage and paternal ancestry inference [26]. As a maternal counterpart, with this study we introduce a short amplicon-based MPS tiling approach for simultaneous whole mtDNA genome analysis using the Ion Torrent PGM allowing to obtain maximum-resolution maternal lineage and maternal ancestry inference from degraded and nondegraded DNA. We tested the performance of the newly developed method using geographically diverse DNA samples with available whole mtDNA genome data based on Sanger sequencing, and using samples for which we generated whole mtDNA genome data based on an alternative MPS protocol. Additionally, we assessed the efficacy of sample pooling, the sensitivity, and the robustness of our new approach regarding experimental and natural DNA degradation.

Materials and Methods

Primer Design

Primer pairs were designed for a total of 161 partially overlapping amplicons in two separate primer pools (pool 1 with 80 pairs and pool 2 with 81 pairs). Each individual primer pool produces a battery of amplicons separated by gaps across the entire mtgenome. By combining the PCR amplicons resulting from each primer pool, the gaps were complemented by amplicons from the other primer pool and the entire 16.5-kb mtgenome could be covered. Amplicons of comparable sizes were preferred (200 bp); however, some regions proved to be difficult to amplify because of, for example, high GC content or repetitive sequences. Also, highly polymorphic positions need to be avoided to overlap with primer annealing sites as much as possible; to overcome these issues, primer positions had to be shifted leading to more variation in amplicon size. Several rounds of test sequencing and primer redesigning were needed to get to the final protocol. All primers were synthesized by Life Technologies, a part of Thermo Fisher Scientific Inc., using their proprietary AmpliSeq modification, which allows multiplexing hundreds or even thousands of targets in one reation. The primers were all received at a concentration of 614 µM. For each pool, 1 µl of each primer is combined and each pool is then brought to a volume of 3070 µl. Each primer is now at a concentration of 200 nM. For the 20-µl PCR, 10 µl of primer pool is used, so the final concentration in the PCR is 100 nM for each primer. Supplementary Table 1 outlines the primer sequences used in this study without the proprietary modification. In the final version, and with the primer pools 1 and 2 together, the entire human mtDNA genome is covered via 161 overlapping amplicons with fragment size of 144–230 bp (average across amplicons: 200 bp).

DNA samples

With the aim to cover the global variation of the mtDNA phylogeny [27], 20 DNA samples known to belong to widely divergent haplogroups were selected from four different sources: the Centre d'Etude du Polymorphisme Humain (CEPH, Paris, France) Human Genome Diversity Project (HGDP) panel (http://www.ceph.fr/HGDP-CEPH-Panel), individuals from across The Netherlands [28], the commercially available Ethnic Diversity DNA panel (EDP-1) manufactured by the European Collection of Cell Cultures (ECACC) (http://www.phecultureccollections.org.uk/products/dna/ethnicdna.jsp), and distributed by

Sigma, and volunteers recruited at the Erasmus MC with informed consent.

Library Construction

DNA libraries were constructed with 10 ng of input DNA using the Ion Ampliseq Library Kit 2.0 (Life Technologies, a part of Thermo Fisher Scientific Inc., Foster City, CA) and 18 cycles with two separate primer pools, amplifying the 161 amplicons simultaneously, following the manufacturer's recommendations. The quantity of DNA was determined using the Qubit dsDNA BR Quantification Kit and a Qubit 2.0 Fluorometer (Invitrogen, Life Technologies, a part of Thermo Fisher Scientific Inc., Grand Island, NY). In order to test the feasibility of sequencing multiple samples simultaneously, Ion Xpress Barcode Adapters (Life Technologies, a part of Thermo Fisher Scientific Inc.) were used.

To evaluate the sensitivity of the tiling approach, variable amounts of DNA input ranging from 10 ng to 100 pg were run and their sequences were evaluated. Three of the 20 DNA samples, hailing from different populations—Italy (haplogroup, T2b), The Netherlands (haplogroup, V3c), and South Africa (haplogroup, L3d3a1a), were sequenced at 10 and 1 ng, 500, 250, and 100 pg of genomic DNA input. All the DNA dilutions were quantified and confirmed in duplicates with the Quantifiler Human DNA Quantification kit (Applied Biosystems, Foster City, CA) following themanufacturer's guidelines.

The robustness of the assay to successfully type degraded materials was tested by subjecting an additional DNA sample (in this case from an Indian individual belonging to mtDNA haplogroup M35a) to DNase I treatment at different time intervals: 5, 10, 15, 20, 30, and 45 minutes. Three microliter of each of these DNase-treated samples was used for the MPS. Furthermore, sample 15 (from Italy with haplogroup T2b) was subjected to two different degradation methods—exposure to ultraviolet (UV) light for 30 minutes using a Bio-Link (Vilber Lourmat) at a strength of 50 J/cm2 and enzymatic shearing using the Ion Shear Plus Reagents Kit (Life Technologies, a part of Thermo Fisher Scientific Inc.). Three microliter of each of the degraded samples was used for the NGS.

In addition, we applied our whole mtgenome MPS method to DNA extracted from six different bones and teeth samples from naturally degraded human remains found in Poland. Samples 1, 2, 4, 5, and 6 were extracted from teeth, whereas sample 3 was extracted from a femoral bone. Samples 1, 3, 4, and 5 gave a full profile with the AmpFSTR NGM (Applied Biosystems), whereas samples 2 and 6 gave partial profiles with only 4 and 6

loci, respectively, out of the 16 loci (15 STRs plus amelogenin) of a full NGM profile. For samples 4, 5, and 6, the control region was analyzed via the Sanger sequencing (unpublished study). The remains that gave rise to sample 2 are believed to be from an archaeological stand fromthe XV–XVI century. Sample 5 likely comes from a recently deceased individual as decomposed soft tissue was found together with the hard tissues. The remains that gave rise to sample 1 were found in soil and the time since death was reported to be approximately 7–8 months, whereas those of sample 3 were found in a river. The samples 1–6 were PCR-quantified using the Quantifiler® Human DNA Quantification Kit (Applied Biosystems) and their concentrations were 0.05, 0.1, 0.045, 5.4, 1.82, and 0.4 ng/µl, respectively. For MPS, we used 4 µl of each of the samples.

Furthermore, as a comparison to currently available NGS assays, five out of the 20 DNA samples were also sequenced using a different library preparation method. The entire mtDNA genome was amplified with two overlapping >8 kb fragments using the amplification primers from Fendt et al. [12]. The SequalPrep Long PCR Kit with dNTPs (Invitrogen) was used for the long-range amplification following the manufacturer's guidelines. The amplified samples were quantified using theQubit dsDNABR Quantification Kit and a Qubit 2.0 Fluorometer (Invitrogen) and the amplified samples were normalized to 100 ng of input DNA. Libraries were constructed using the Ion Shear Plus Reagents Kit (Life Technologies, a part of Thermo Fisher Scientific Inc.) for enzymatic fragmentation of DNA, followed by barcoded adapter ligation using the Ion Xpress Barcode Adapters (Life Technologies, a part of Thermo Fisher Scientific Inc.) and Ion Plus Fragment Library Kit (Life Technologies, a part of Thermo Fisher Scientific Inc.) as per the manufacturer's protocol. The libraries were size selected using the E-Gel SizeSelect 2% Agarose Gel (Invitrogen, Carlsbad, CA).

Template Preparation

Using emulsion PCR, the generated libraries are attached to beads and further amplified. The concentration of each of the libraries was determined with qPCR using the Ion Library Quantitation Kit (Life Technologies, a part of Thermo Fisher Scientific Inc.), and the template dilution factors were calculated. High-quality templated Ion Sphere particles, containing massively parallel clonally amplified DNA, were prepared for the 200 base-read libraries using the Ion PGM Template OT2 200

Kit and the Ion OneTouch 2 System (Life Technologies, a part of Thermo Fisher Scientific Inc.). The template-positive Ion Sphere particles were enriched with the Ion OneTouch Enrichment System (Life Technologies, a part of Thermo Fisher Scientific Inc.) as per the manufacturer's guidelines.

PGM Sequencing

Semiconductor sequencing on the Ion PGM detects the change in pH when the proton (H+) is released during nucleotide incorporation. The sequencing was conducted on the Ion 318 Chip v2 using the Ion PGM Hi-Q Sequencing Kit (six barcoded samples per chip) coupled with the PGM sequencer (Life Technologies, a part of Thermo Fisher Scientific Inc.). A chlorite cleaning, followed by 18 MΩ water wash (Thermo Fisher Scientific, Millipore, MA) was performed prior to the initialization of the Ion PGM System.

MPS Data Analysis

Commercially available Next GENe software (v2.4.0.2) was used to analyze all the sequences generated on the Ion PGM sequencer. All the sequences were trimmed by 20 bases on both the 3′ and the 5′ ends and aligned to the mtDNA revised Cambridge Reference Sequence (rCRS) [29] (GenBank NC 012920.1). The arbitrary threshold of 50 reads was considered as a minimum required to address full sequencing coverage to reliably report variants/polymorphisms. This threshold is the coverage minimum to be able to see heteroplasmies and to be able to clearly identify stochastic sequencing errors. A BED file that contains the amplicon regions was also uploaded for the analysis. The NextGENe software allows visualization of sequence reads in addition to summarizing the polymorphic sites in a tabular column.

The NextGENe software uses BLAST-like Alignment Tool to align the sequence reads to the reference. Nonetheless, in some cases, manual inspection was needed to confirm certain variants and anomalies, especially regarding length heteroplasmy. It is known that the corrected version of the CambridgeReference Sequence (the rCRS) retained the same nucleotide numbering scheme as the original reference sequence [the CRS; 30], and to correct one of the erroneous positions, an "N" was recorded at position 3107 to represent a deletion at that position [29]. The NextGENe software lists it as 3107d in any tested sample, but as this deletion does not represent a genuine sequence variant, it shall be ignored. Additionally,

this software does not adhere to the nomenclature recommendations detailed by the forensic community [31,32]. In particular, rather than reporting the common dinucleotide deletion in the dimeric repeat region between 514 and 524 as m.523A>del and m.524C>del (placing the deletion at the most 3′ position possible), NextGENe® employs a 5′ indel alignment and notes these variants as m.513G>del and m.514G>del. Furthermore, we noticed that the two sequential transitions that sometimes co-occur at positions 151 and 152 were instead reported as a deletion at np 151 and an insertion at 152, following the recommendations byWilson et al. [33,34].

With the mtDNA sequence variants detected, the haplogroups were inferred using the Web-based bioinformatics tool, MitoTool [35],which uses the most recent version of PhyloTree (http://www.phylotree.org; Build 16) [27].

Results and Discussion

In recent years, DNA sequencing technology has evolved from traditional Sanger sequencing of single reads to sequencing thousands of reads with high coverage in a massively parallel fashion with MPS technologies such as for analyzing human mtDNA [22-25]. Contrary to the large-fragment sequencing strategy employed mostly so far for MPS analysis of whole mt genomes [23,24], here we introduce an MPS approach for analyzing the entire human mtDNA genome using 161 pairs of short amplicons (144–230 bp, average 200 bp) in two different primer pools. We demonstrate the whole mt genome coverage of our approach via analyzing worldwide DNA samples belonging to different mtDNA haplogroups, perform quality control via comparison with data obtained with alternative methods in the same samples and provide preliminary data on sensitivity testing, degraded DNA testing using artificially and naturally degraded DNA, and multiple DNA sample testing using barcoded adapters; the DNA variants were determined using the NextGENe v2.4.0.2 software.

Method Application to Worldwide Samples and Quality Comparison with Alternative Methods

To test the performance of the newly designed method for simultaneous whole mt genome sequencing using the Ion Torrent PGM, we applied it to 20 carefully selected DNAsamples that belong to different global mtDNA haplogroups (Table 1). For these 20 samples, the average sequencing coverage ranged from 1302 to 5637 reads with an average of 3293 reads across all samples. The total number of reads per sample ranged from 211908 to 641486, with

an average across samples of 366327, of which 88% were successfully aligned to the rCRS (325349 reads). Using a coverage threshold of 50 reads, 100% mt genome coverage was achieved in all the samples except three (samples 14, 17, and 20), which each miss one small piece of mtDNA sequence with our MPS approach (Table 1). In sample 17, the 9-bp deletion at np 8281–8289 (which is the defining mutation of haplogroup B, among others) was not detected by the NextGENe software in the final variant table, elucidating the current limitation in the software package used. Further improvements in the bioinformatics pipelines are needed to deal with such data. A discreet manual inspection was done to confirm the deletion. In samples 14 and 20, fragments of 66 bp (np 10468–10533) and 61 bp (np 10269–10329), respectively, were missed when using the 50-reads threshold applied. Figure 1 illustrates the average coverage of each of the 161 amplicons across all the 20 samples. The mtDNA variants detected for all the 20 samples analyzed are detailed in Supplementary Table 2.

Further, the whole mt genome data we generated via our MPS approach from the samples belonging to the CEPH–HGDP panel (samples 1–11) were compared with those previously reported by Hartmann et al. [36] based on whole mt genome sequencing using the Sanger approach. The remaining nine samples (12–20) we analyzed via MPS were subjected to de novo Sanger sequencing of the control region (unpublished study, data not shown). Some of the positions in the coding region were further compared with the results previously obtained from the five SNaPshot multiplex assays established earlier (data not shown) [37]. The data obtained via our PGM method were then compared with the data generated by alternative methods for quality control purposes. The 9-bp deletion at np 8281–8289 in sample 17 was not tabulated in the variant table generated by NextGENe. However, this deletion was evident from results obtained with the SNaPshot multiplex assay and also apparent in the NextGENe viewer. Several mis-calls by NextGENe were revised after a meticulous manual inspection of the sequence in the NextGENe viewer. Disparities observed between the Sanger sequencing/SNaPshot and MPS are tabulated in Supplementary Table 3. Some of the discrepancies were observed in calling of the variants involving length heteroplasmy in the polycytosine stretches of the HVS-I and HVS-II. In samples 1, 6, 11, and 17, due to an erroneous shift in the bases, the variant was reported as m.16183A>M (an apparent heteroplasmy), instead of m.16183A>C (a homoplasmic transversion) (Supplementary Figure 1A). In the HVS-II region, for samples 8, 13, 16, 19, and 20, only

one insertion of the base C in the np 303–309 region was reported with Sanger sequencing. However, with MPS, two insertions of the base C were noted (Supp. Fig. S1B). Such differenceswill in practice not have any influence on haplogroup determination since the insertions at np 309, 315, 16,193, AC indels at 515–524, m.16182A>C, m.16183A>C, andm.16519T>C are usually not considered for phylogenetic reconstruction and are therefore not included in the PhyloTree [27]. The variant reported as 357M in the samples 2, 7, 14, and 15 is specious and appears due to a shift in the base position, possibly because of the polyadenine stretches as shown in Supp. Figure S2. Similarly, in samples 8, 12, and 17, the variant m.13128C>M is spurious due to the shift in the base because of the polycytosine stretches as shown in Supp. Figure S3.

Furthermore, spurious gaps were observed in the samples 8 and 12 at position 13,128 (Supplementary Figure 3 boxed region). Samples 9, 10, and 14 report an insertion 539.1C (Supplementary Figure 4),which was not evident in the Sanger sequencing data. It has been described earlier that the PGMproduces a high frequency of homopolymer sequencing errors and indels [38,39]. In samples 2 and 4, a spurious deletion was noticed at position 5824 that was not reported in the Sanger sequencing data of Hartmann et al. [36] (Supplementary Figure 5). Further, differences to the Sanger sequencingwere observed in sample 9 at position 16209 and in sample 10 at position 204 (Supplementary Table 3, Supplementary Figure 6). Spurious single-bp deletions were observed at some sequence positions in some reads in few of the samples (Supplementary Figure 6 boxed region) but not reported in the variant table. Such gaps did not influence the final consensus sequence and therefore did not affect the final haplogroup determination.

One potential disadvantage of using a tiling approach based on small amplicons over a large fragment amplification approach is the potential detection of nuclear copies of mt sequences(NUMTs)with the tiling approach that likely are not detected with the long-range approach simply because NUMTs typically aremuch smaller. To test for this, we generated comparative whole mt genome sequencing data using the previously described two overlapping 8.5-kb amplification method [12] in five of the samples (samples 1, 2, 3, 4, and 20) we used for PGM sequencing. It is noteworthy that the variants detected were in concordance with those detected when sequenced with the short overlapping fragments, except in sample 2. In sample 2, the polymorphisms m.10664C>Y, m.8251G>R, and m.8252C>M (Supplementary Figure 7; Table 2) were discordant with the tiling approach (m.10664C>T, m.8251G>A, and

no variant at position 8,252, respectively). The erroneous calling at positions 8,251 and 8,252 could be due to the base position shifts as seen in Supplementary Figure 7. The disparities observed in the HVS-I and HVS-II due to the length heteroplasmy were also noticed with the long-range amplification. Hence, from the samples analyzed, we have no evidence that our tiling MPS approach picks up NUMTs.

Sample ID	Total Reads	Aligned Reads	Percent of Aligned Reads	Maximum Coverage	Average Coverage	Percent of Coverage
1	641486	535003	83.40	67847	5593	100%
2	502725	427031	84.94	54668	4852	100%
3	268055	224250	83.66	68677	4494	100%
4	408522	343567	84.10	44394	4285	100%
5	355691	305619	85.92	50539	3804	100%
6	493881	485970	98	32082	3143	100%
7	289696	258720	89.31	43687	2607	100%
8	264828	232310	87.72	23412	1488	100%
9	264231	204023	77.21	47124	3777	100%
10	352588	279159	79.17	53959	3496	100%
11	476939	426050	89.33	66262	5637	100%
12	268968	235011	87.38	37458	2560	100%
13	249338	198790	79.73	46631	1840	100%
14	433566	389219	89.77	52840	3358	99.59%
15	541295	532291	98.34	44460	3488	100%
16	211908	194369	91.72	25927	1302	100%
17	307441	278527	90.60	30230	1865	100%*
18	324954	299098	92.04	48897	2006	100%
19	420730	416278	98.94	38839	3644	100%
20	249701	241699	96.80	37051	2628	99.62%
Average	366327	325349	88	45749	3293	

* The 9-bp deletion at np 8281-8289 in Sample 17 is the defining mutation of haplogroup B. Hence the coverage for sample 17 was considered to be 100% despite the 9-bp deletion.

Table 1: Performance Summary of the 20 Geographically Diverse DNA Samples for Whole mt Genome Sequencing with the MPS Tiling Approach via 161 Short Overlapping Amplicons

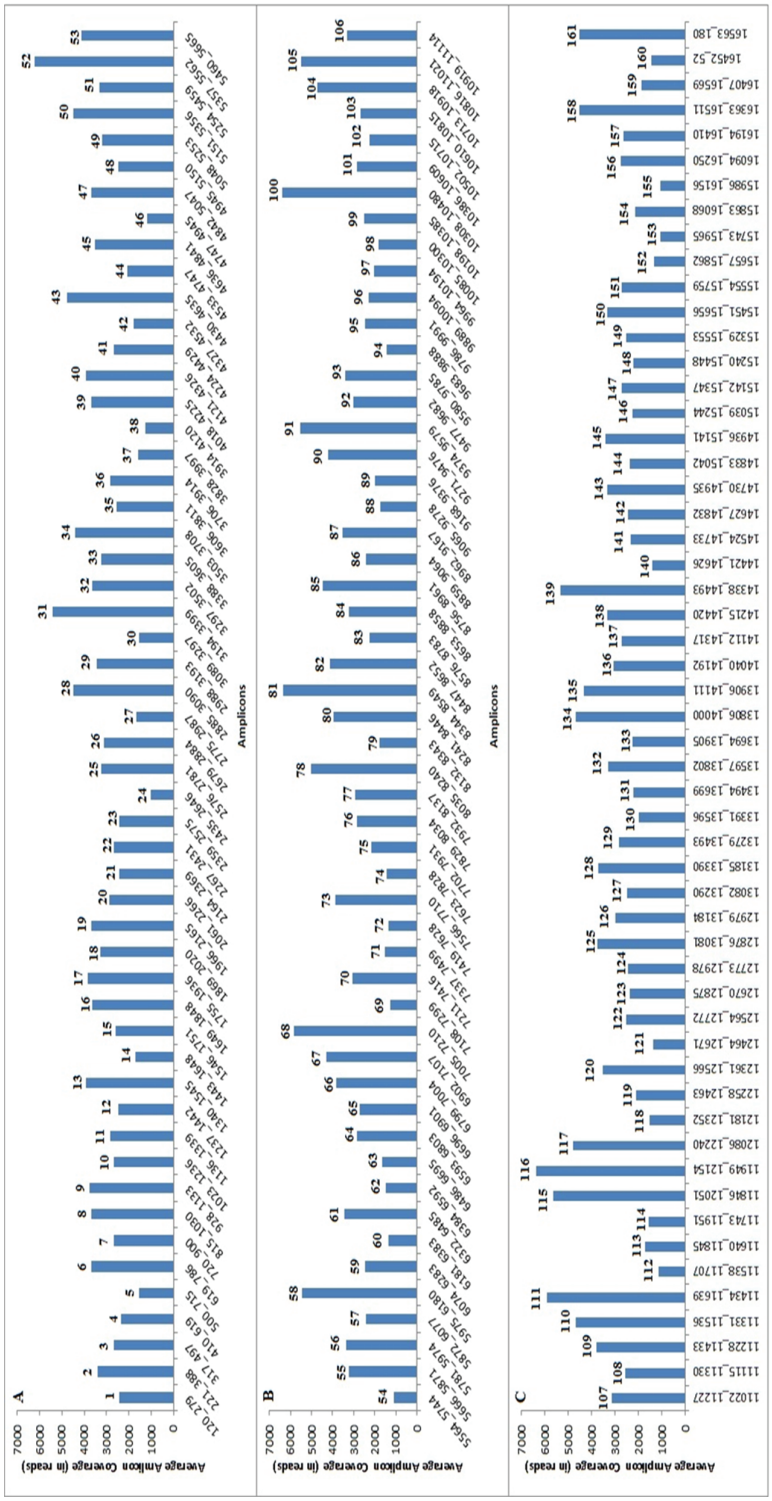


Figure 1: The average amplicon coverage, across all the 20 samples tested with the newly developed MPS tiling approach for whole mt genome sequencing, are presented in three plots. A: Amplicons 1–53. B: Amplicons 54–106. C: Amplicons 107–161.

		NGS Data from NextGENe® software					
		%A	%C	%G	%T	%Insertions	%Deletions
Long Range	m.10664C>Y	0	17.46	0	78.31	0	4.23
Tiling	m.10664C>T	0	0.34	0	99.35	0	0.3
Long Range	m.8251G>R	71.43	0	28.57	0	0	0
	m.8252C>M	29.37	70.63	0	0	2.1	0
Tiling	m.8251G>A	97.27	0	2.1	0	0	0.63

Table 2: Differences Between the Previously Developed Long-Range Amplification MPS Approach and the Newly Introduced MPS Tiling Approach forWhole mt Genome Analysis Both Obtained via the PGM in Sample 2

It has been described earlier [24,40] that false deletions were observed in the mtDNA sequencing data generated using the PGM. However, no such deletions were observed when the variants reported with our tool and the long-range PCR approach was compared. Except for the discordant calls in sample 2, no other differences in variant calling were observed. Nonetheless, this may change when more samples are tested, as in this study only a few samples have been tested.

Point heteroplasmies at 22 positions in 14 samples were reported in the NextGENe variant table: sample 1: m.8155G>R; sample 2: m.1048C>Y, m.357A>M; sample 3: m.14423G>S; sample 4: m.4788G>R, m.14921G>R; sample 5: m.1171A>R, m.13020T>Y, m.15326A>R, m.16189T>Y; sample 6: m.7759T>Y; sample 7: m.771A>R, m.4248T>Y, m.357A>M; sample 8: m.3296T>Y, m.13128C>M; sample 9: m.10185C>Y, m.15326A>R, m.16209T>Y; sample 10: m.204T>Y, m.13928G>S, m.15326A>R, m.14831G>R; sample 12: m.13128C>M; sample 14: m.357A>M; sample 15: m.357A>M; and sample 17: m.13128C>M. Heteroplasmy threshold was set at 20% (of total coverage) [24]. Out of the 22, only seven of the point heteroplasmies (sample 1: m.8155G>R; sample 3: m.14423G>S; sample 4: m.4788G>R, m.14921G>R; sample 5: m.1171A>R, m.16189T>Y; sample 6: m.7759T>Y) were also observed in the Sanger sequencing data of Hartmann et al. [36]. A heteroplasmy at position 3296 was not reported by Hartmann et al. [36], contrary to the m.3296T>Y from this study. However, in a recent publication, wherein some samples from the CEPH-HGDP panel were resequenced on the Illumina platform, a variant at 3296 was

recorded for the sample 8 as m.3296T>N [41], suggesting that this heteroplasmy could be genuine.

In general, the availability of a viewer in the NextGENe software proved advantageous in manually evaluating the discrepancies and solving them accordingly, especially in assessing the true heteroplasmy.

mtDNA Haplogroup Assignment and Maternal Ancestry Inference

The haplogroups for the 20 samples were inferred from the obtained whole mtDNA genome data using MitoTool [35] and are presented in Table 3 together with the geographic sampling origin and the previously reported geographic region of haplogroup origin indicative of maternal biogeographic ancestry. All the haplogroups determined for the 20 samples were in agreement with their biogeographic origin. It is important to assert that mtDNA construes only the matrilineal ancestry information of an individual. To achieve comprehensive biogeographic ancestry inference from DNA, in addition to the global matrilineal biogeographic ancestry assignment, information about paternal ancestry using male-specific Y-chromosomal DNA (in the case of males) and from biparental ancestry using ancestry-informative autosomal DNA markers have to be considered [42].

Sample ID	Haplogroup	Broad Haplogroup	Known Sampling Region	Main geographic region of the (broad) haplogroup origin	References
1	L3d3a1a	L3*(xM,N)	South Africa	Africa, West Asia	Behar et al., 2008
2	L0d1a1a (199 missing)	L0	South Africa	Southern Africa	Behar et al., 2008; Barbieri et. al., 2013
3	H1c	H	Russia (Caucasus)	West Eurasia, Northern Africa	Loogväli et al., 2004; Achilli et al, 2004; Roostalu et al., 2007
4	U7a2	U7	Israel	West Eurasia, Central Asia, Southern Asia	Palanichamy et al., 2004; Brisighelli et al., 2009
5	P1d1	P	New Guinea	Oceania (Papuan, Melanesian and Australian Aborigines)	Friedlaender et al., 2007; Hudjashov et al., 2007
6	V	HV*(xH)	Algeria	West Eurasia, Northern Africa	Achilli et al., 2005; Álvarez-Iglesias et al., 2009
7	A1a1 (missing 235)	A*(xA2)	China	East Asia	Kong et al., 2006; Derenko et. al., 2007
8	J2b1a	J	Italy	West Eurasia	Pala et al., 2012
9	M7c1a2a	M*(xM1,C,D)	China	South Asia, East Asia, Southeast Asia	Kong et al., 2006; Derenko et. al., 2007
10	F4a1a	R9	China	East Asia, Southeast Asia	Kong et al., 2006
11	X2b5	X	Orkney Islands	West Eurasia, Northern Africa, Americas	Reidla et al., 2003; Achilli et al., 2008
12	J2b1(J2b1a1: missing16278)	J	Italy	West Eurasia	Pala et al., 2012
13	P	P	Australia (Aborigine)	Oceania (Papuan, Melanesian and Australian Aborigines)	Friedlaender et al., 2007; Hudjashov et al., 2007
14	M72a	M*(xM1,C,D)	Thailand	South Asia, East Asia, Southeast Asia	Tabbada et al., 2010; Peng et al., 2010
15	T2b	T	Italy	West Eurasia	Pala et al., 2012
16	K1a12	U	Netherlands	West Eurasia	Achilli et al., 2005; Behar et al., 2006
17	B5a1a	B5	Netherlands	East Asia	Kong et al., 2006
18	W5a1a	W	Netherlands	West Eurasia	Finnilä et al. 2001; Palanichamy et al., 2004
19	V3c	HV*(xH)	Netherlands	West Eurasia, Northern Africa	Achilli et al., 2005; Álvarez-Iglesias et al., 2009
20	H23	H	Netherlands	West Eurasia, Northern Africa	Loogväli et al., 2004; Achilli et al, 2004; Roostalu et al., 2007

Table 3: Haplogroups of the 20 DNA Samples Whole mt Genome Sequenced with the MPS Tiling Tool and Interpreted Using MitoTool

Preliminary Sensitivity Testing

In order to determine the limit of detection provided the 50 reads coverage threshold used and given the sensitivity of the developed PGM assay, preliminary sensitivity tests were performed with differing starting amounts of DNA at 100, 250, and 500 pg, and 1 and 10 ng (measured as genomic DNA). For this, we used three of the 20 DNA samples originating from different populations with different haplogroups—South Africa with haplogroup L3d3a1a (sample 1); Italy with T2b (sample 15); and The Netherlands with V3c (sample 19). The results of the sensitivity study on the samples are summarized in Supplementary Table 4. For all the sample dilutions, it was possible to detect the correct haplogroup down to merely 100 pg input genomic DNA, even though the input manufacturer's recommendation for library construction is 10 ng. Regarding whole mt genome coverage, we achieved 100% for all sample dilutions, except for sample 19 at 250 pg (99.89% at 576 reads on average), and for all three samples at 100 pg (99.93% at 2591 average reads, 99.91% at 1175 average reads, and 99.74% at 462 average reads for samples 1, 15, and 19, respectively). Notably, samples were further diluted to 50 pg input, but could not proceed to sequencing as the template dilution factors calculated were below one, whereas one is recommended by the manufacturer, thus implying that the amount of DNA library generated from such low quantity DNA samples was not sufficient to perform optimal emulsion PCR [26]. Previous studies have shown that it is possible to achieve successful sequencing results with the Sanger protocol using 50 pg of input DNA, and sometimes even at 10 pg of input DNA with minimal failures [43,44]. However, to achieve full mtDNA coverage at a high resolution with the Sanger approach, several individual sequences are required, as opposed to the parallel nature of the MPS where a high volume of data is generated in relatively short time. Though it was possible to achieve correct haplogroup information at 100 pg, increasing the number of PCR cycles could result in good coverage with input below 100 pg of DNA. Hence, further testing needs to be done to show the full-sensitivity limits of this system.

Preliminary Analysis of Experimentally and Naturally Degraded DNA Samples

The robustness of our MPS assay regarding DNA degradation was tested by sequencing experimentally and naturally degraded DNA samples. Aliquots of a sample (1 ng genomic DNA), belonging to haplogroup M35a1, was subjected to DNase treatment at different time intervals: 5, 10, 15, 20, 30, and 45 minutes. The effect of DNA degradation in these samples

was first monitored by analyzing them with the AmpFlSTR Identifiler PCR Amplification Kit (Applied Biosystems) targeting 15 autosomal STRs plus the amelogenin sex typing system, which is routinely used for human identification purposes (Figure 2). From Figure 2, it is evident that the DNA degradation-induced STR locus dropouts are clearly correlated with PCR fragment size, as expected. Particularly, at 5 min of DNase treatment, there were complete locus dropouts at three STRs CSF1PO (allelic fragment length 330 bp), FGA (280 bp), and D2S1338 (310 bp); at 10 minutes, additional dropout at D7S820 (271 bp), D16S539 (280 bp), and D18S51 (270 bp); at 15 minutes, additional dropouts at TPOX (238 bp) and D13S317 (222 bp); at 20 minutes, additional dropout at D21S11 (206 bp), TH01 (172 bp), and vWA (178 bp); at 30 minutes, additional dropout at D5S818 (154 bp), D8S1179 (146 bp), and D19S433 (128 bp). After 45 minutes of DNase treatment, all of the 16 loci covering allelic fragment sizes of about 128–330 bp dropped out, except the two loci with shortest alleles, amelogenin (108 bp) and D3S1358 (116 bp). This analysis provides a rough idea on DNA fragmentation in these experimentally degraded samples as relevant for the subsequent mtDNA genome sequencing using our MPS tiling approach.

Table 4 enlists the performance summary of our whole mtgenome tiling MPS approach in the DNase-treated samples at different time intervals. Most notable, for the samples degraded for 5, 10, and 15 minutes, we obtained 100% mt genome coverage using a 50x threshold, at 6810, 3886, and 3007 average reads, respectively. Prior STR analysis showed that eight loci with allelic fragment length from 222 to 330 bp already dropped out in these three degraded samples.For samples degraded for 20, 30, and 45 minutes, the coverage dropped down to 94.4%, 82.6%, and 19.25%, respectively, at 2073, 1183, and 487 average reads, respectively. Regarding the total number of mtDNA variants detected in these samples, after 5 min of DNase treatment, it was 37 as well as after 10 and 15 minutes where we obtained 100% mt genome coverage, whereas it dropped down to 30, 24, and 13 in the samples that received DNase treatment for 20, 30, and 45 minutes, respectively. However, despite the loss of reads, fragments, and thus DNA variants, it was still possible to determine the correct mtDNA haplogroup (M35a1) in all degraded DNA samples from 5 to 45 minutes of DNase treatment. Figure 3 elucidates the average amplicon coverage across all amplicons for the different time intervals of enzymatic degradation clearly showing the loss of sequence reads with increased degradation time, as expected. Figure 4 illustrates the

amplicon coverage according to amplicon length for the different time intervals of enzymatic degradation, clearly demonstrating the effect of DNA degradation on amplicon length and number of reads. It is to be noted that successful sequencing achieving 100% mtgenome coverage was obtained from degraded DNA samples with a maximal fragment size of about 220 bp (Figure 2).

To test two additional experimental degradation methods, aliquots of 1 ng genomic DNA of sample 15, belonging to haplogroup T2b, were exposed to UV radiation for 30 minutes using a Bio-Link (Vilber Lourmat) at a strength of 50 J/cm2 and enzymatic shearing using the Ion Shear Plus ReagentsKit (Life Technologies, a part of Thermo Fisher Scientific Inc.). The performance summary of the samples is depicted in Table 4. At 1 ng unexposed DNA, the total number of mtDNA variants detected was 41 at 100% mtgenome coverage (1865 reads on average); when the sample was exposed to UV for 30 minutes, only 38 of the 41 variants were detected (missed variants were m.7310T>C, m.8697G>A, and m.10463T>C) with a mt genome coverage of 88.65% at 1,362 reads on average. At position 13,368, the variant was reported as m.13368G>R (49.36% A and 50.61% G), whereas the true variant was m.13368G>A. In contrast, all the 41 variants were correctly detected when the sample was subjected to enzymatic shearing, even though only 89.9% mt genome coverage was obtained at 975 reads on average. The variant m.357A>M, which was explained earlier as a sequencing error probably because of a base shift position due to the poly-A stretch, was detected in both degradation approaches. However, both degradation methods allowed our MPS approach to determine the correct haplogroup of sample 15 (i.e., T2b).

Additionally, DNA extracts from six human remains (i.e., teeth and bones) were sequenced with our MPS tiling approach to investigate naturally degraded DNA (Table 4). Notably, samples 1, 3, 4, and 5 delivered a complete 16 loci (15 STRs plus amelogenin) DNA profile with the AmpFlSTR NGM Kit (Thermo Fisher Scientific) regularly used for human identification, whereas samples 2 and 6 delivered only a partial NGM profile with 12 and 10 of the 16 loci missing, respectively.With our MPS approach, we obtained mtgenome coverage of close to 90% for samples 1, 3, and 4, whereas for samples 6, 5, and 2, our method delivered 75%, 60%, and 50% of mtgenome coverage (Table 4). Notably, sample 2 with the lowest observed mtgenome coverage of 50% (225 reads on average) likely originates from the XV–XVI century. For samples 4, 5, and 6, mtDNA control region HVI and HVII data were obtained successfully via Sanger sequencing and compared with those of the tiling MPS

approach. Sanger sequencing of these samples was performed as part of another unpublished study and hence the protocol was not described here. These data were comparable, except that the variant m.16519T>C, which was clearly evident in the tiling MPS approach, but was not reported in the Sanger sequencing data for all the three samples. However, Sanger sequencing mtDNA control region data were not available for samples 1, 2, and 3.

Although mtgenome coverage was only obtained to the degree of 50%–90% from these six naturally degraded DNA samples, our MPStiling approach allowed haplogroup assignment for all of them: U5b2b, H4a1, H, T2b, U4a2, and T1a for samples 1–6, respectively.

The AmpliSeq-based MPS approach we developed here for complete human mt genome analysis is designed for and suited to nondegraded and mildly to considerably degraded DNA leading to DNA fragmentation down to a size range of around 200 bp and larger, as often confronted with in forensic, medical, and anthropological studies. Our approach is not suitable for more severely degraded DNA as often confronted with in ancient DNA studies. For such strongly fragmented DNA, hybridization capture methods are more suitable than AmpliSeq-based approaches and have been developed in the field of ancient DNA research [45 - 48]. The AmpliSeq system we used in our approach is optimized for this kind ofmultiplexed amplification and the workflow requires fewer steps,which reduces the chance of contamination and mix-up.

Additionally, more targets are generated with theAmpliseq PCR system than with a hybridization capture-based system, allowing for deeper coverage, mitigating sequencing errors.

		Total Reads	Aligned Reads	Percent of aligned reads	Maximum Coverage	Average Coverage	Percent of Coverage	Haplogroup
Sample 21 Dnase treated at different time intervals	**5min**	463854	452505	97.55	51989	6810	100%	M35a1
	10min	242109	234156	96.72	45167	3886	100%	
	15min	163854	152505	93.07	38249	3007	100%	
	20min	60988	59085	96.88	31262	2073	94.40%	
	30min	83565	80367	96.17	10731	1183	82.60%	
	45min	15642	13606	86.98	9625	487	19.25%	
Sample 15	**UV for 30 min**	258179	255277	98.88%	21901	1362	88.65%	T2b
	Enzymatic shearing	315030	278648	88.45%	5827	975	89.90%	
Ancient and degraded bone and teeth samples	**Sample 1**	542889	525140	96.73	9572	4042.52	87.87%	U5b2b
	Sample 2	95396	58311	61.13	2379	225.57	50.36%	H4a1
	Sample 3	538548	512826	95.22	7273	3321.23	90.12%	H
	Sample 4	529929	480856	90.74	6048	3141.47	88.32%	T2b
	Sample 5	477578	450506	94.33	8218	2924.11	59.59%	U4a2
	Sample 6	175642	140021	79.72	3535	612.48	75.25%	T1a

Sample 15 subjected to three different degradation methods and six different highly degraded bone and teeth samples.

Table 4: Performance Summary of PGM-BasedWhole mt Genome Sequencing of DNase-Treated Sample at Different Time Intervals

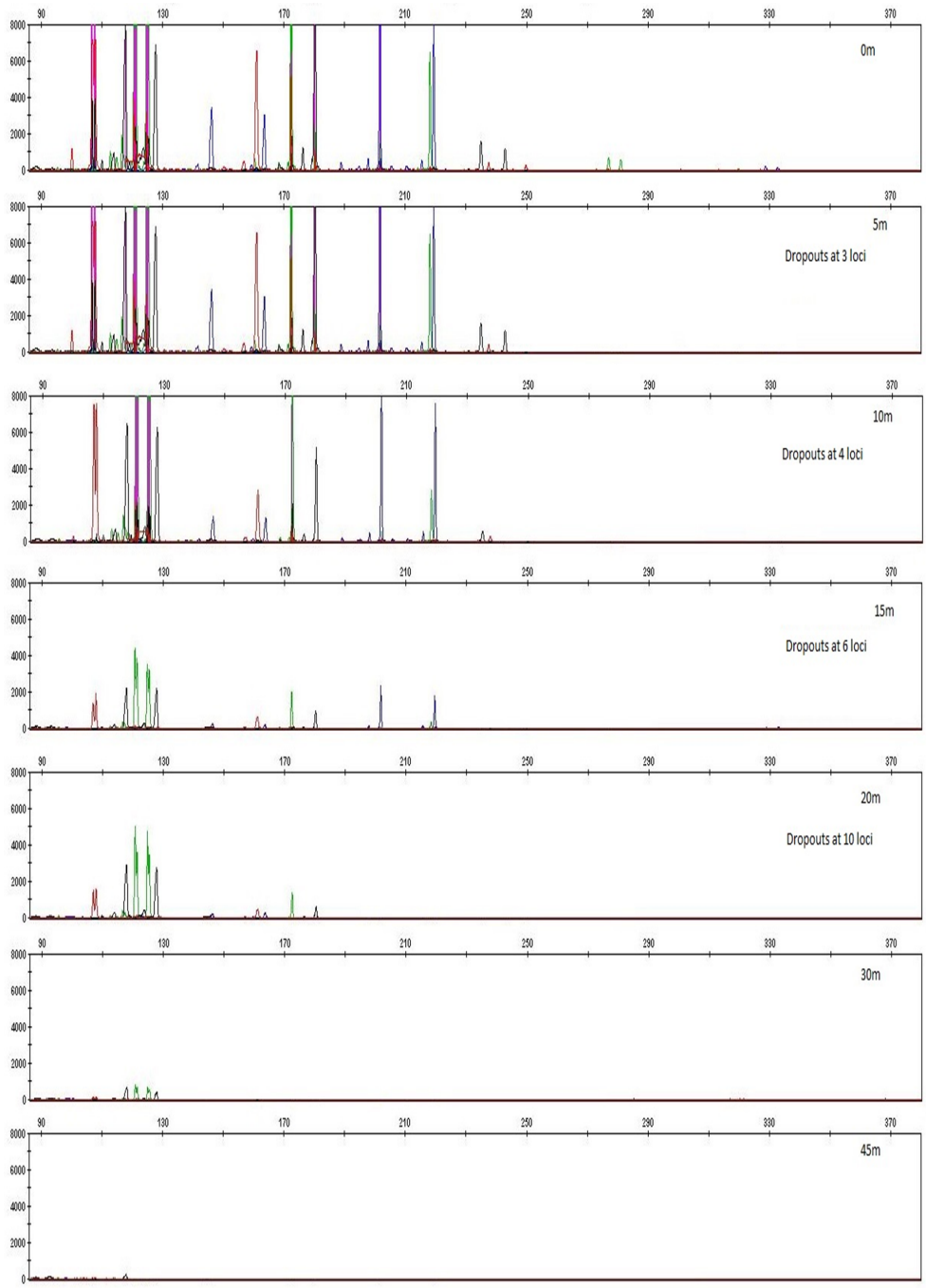


Figure 2: STR profiles from the AmpFlSTR R Identifiler R PCR Amplification Kit (Applied Biosystems) targeting 15 autosomal STRs plus amelogenin to illustrate the degree of DNA fragmentation for the sample treated with DNase at different time intervals: 5, 10, 15, 20, 30, and 45 min.

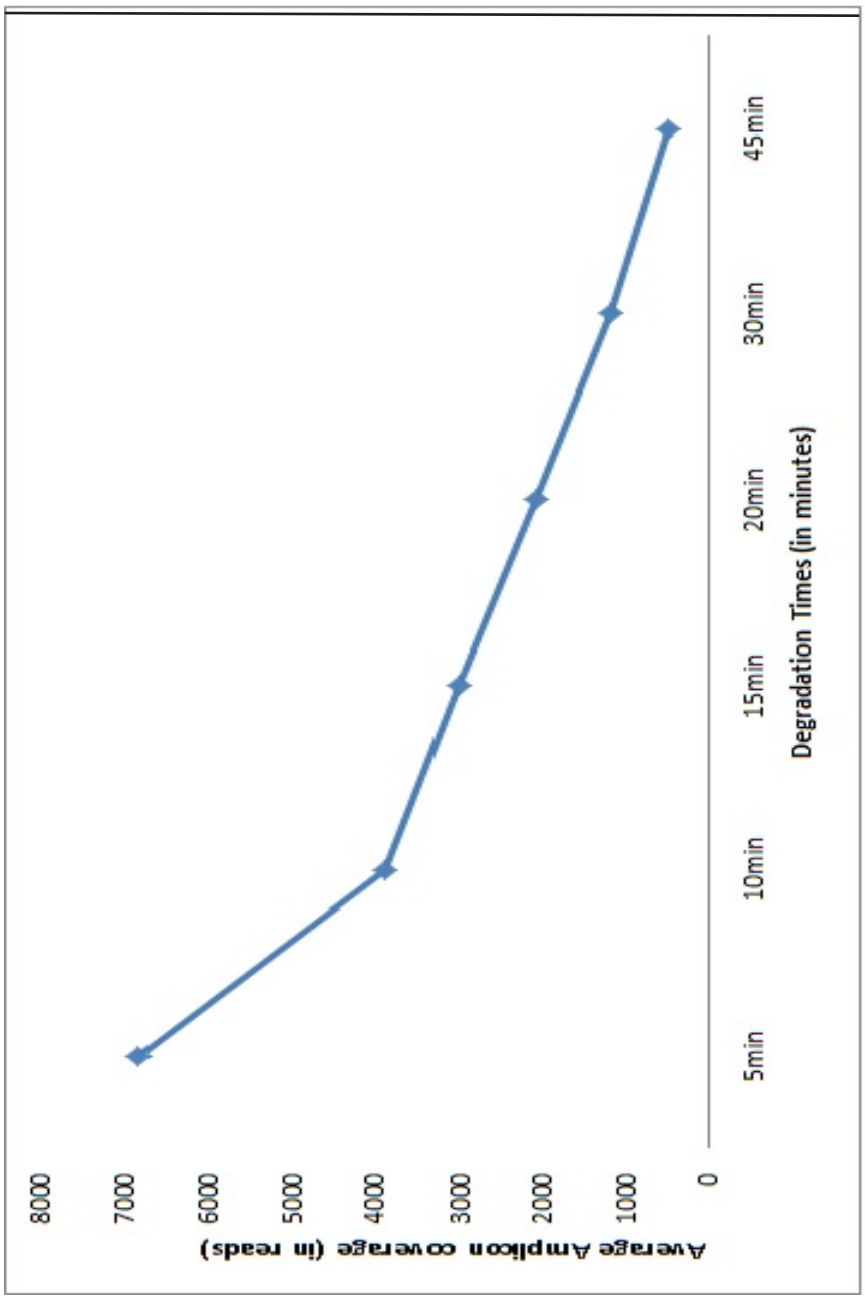


Figure 3: The average mt DNA amplicon coverage across all amplicons for the DNase-treated sample at the different time intervals as obtained with the MPS tiling approach.

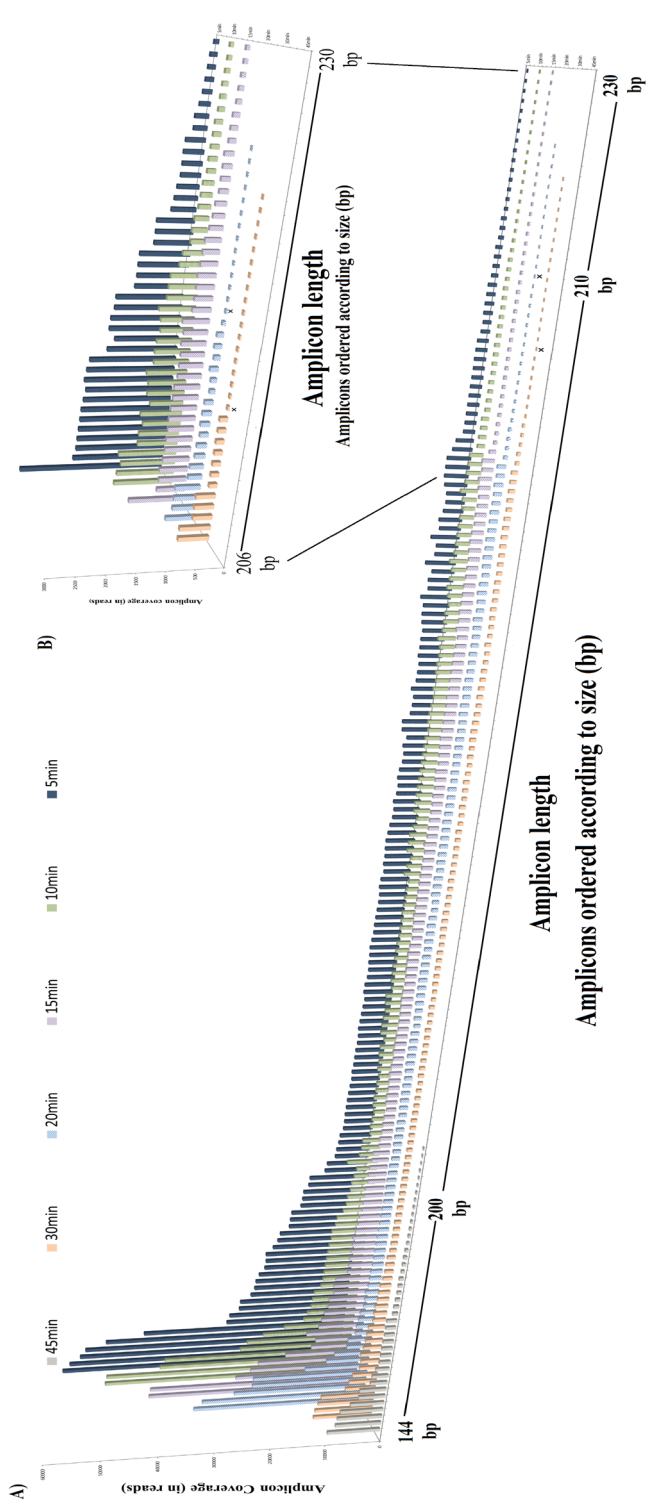


Figure 4.: (A): The amplicon coverage (number of times amplicons were observed in the MPS data) of all the 161 amplicons used to obtain complete mt genome coverage with our MPS approach, arranged according to amplicon length from the shortest amplicon used (144 bp) on the left-hand side to the largest amplicon used (230 bp) on the right-hand side, for the different time intervals of enzymatic DNA degradation: 5, 10, 15, 20, 30, and 45 min. All the amplicons below X represents the coverage below the 50 reads threshold. (B): Additional zoomed-in image of the longer amplicons from amplicon lengths of 206 bp (left-hand side) to 230 bp (right-hand side).

Preliminary Testing of Sample Multiplexing via DNA Barcoding

In order to test the feasibility of sequencing multiple samples simultaneously in one sequencing run, barcode adapters were used. Six DNA samples were pooled (12 libraries, as two primer pools were needed) per each sequencing run and barcoded, using the Ion Xpress Barcode Adapters (Life Technologies, a part of Thermo Fisher Scientific Inc.). The ability to multiplex six samples via barcoding is a valuable asset with regard to the increase in throughput and cost-effectiveness. Further research should involve sequencing more than six samples in a single run on one chip.

Conclusions

Our study shows that analyzing the complete human mt genome in a simultaneous way via a tiling approach targeting short amplicons is feasible. With the short overlapping fragments we employed in covering the entire mtgenome, and supported by our preliminary data on experimentally and naturally degraded DNA samples, we expect our approach to be particularly useful for analyzing degraded materials. As this was a preliminary study introducing the new method, future studies with a larger sample set including more degraded samples need to be conducted to explore the complete value of this tool.However, based on the preliminary data presented here, we already expect our method to be highly useful in many mtDNA applications using degraded and non-degraded DNA where maternal lineage identification and maternal ancestry inference on the maximum possible resolution level is appreciated such as in forensic, anthropological, and medical genetics.

In the longer run, we envision this MPS tool for complete mtgenome analysis allowing maximal maternal lineage determination and maternal ancestry inference being combined with MPS tools for ultra-high-resolution paternal lineage and paternal ancestry identification, such as the PGM tool we recently introduced for simultaneous analysis of >530 Y-chromosomal SNPs allowing to detect >430 worldwide Y haplogroups [26]. Since with the current whole mt genome tool using 161 short overlapping amplicons (144–230 bp, average 200 bp), the sequence capacity limits of the PGM (or alternative devices such as MiSeq) are not reached, further enlargements of combined targeted MPS tools may be expected from future work. For instance, we envision the additional addition of autosomal ancestry-informative SNPs into combined tools together with Y and mtDNA allowing to

not only detect lineages (as possible with mtDNA and Y analysis) but moreover to resolve individual genetic admixture, and to obtain an overall and more extensive estimate of an individual's biogeographic ancestry from all three ancestry components: maternal, paternal, and biparental. For some applications, such as to answer forensic and anthropological questions, the additional incorporation of phenotypic markers into comprehensive ancestry tool(s) such as SNPs predictive for human pigmentation traits or for other externally visible characteristics if available [49] would further add to the investigative value available with such comprehensive targeted MPS tools.

Acknowledgments

The authors thank Thierry Jurado and Philipp Habermeier (both from Life Technologies, a part of ThermoFisher Scientific Inc.) for their help in setting up and hosting this work.

References

[1] Wilson MR, DiZinno JA, Polanskey D, Replogle J, Budowle B. 1995a. Validation of mitochondrial DNA sequencing for forensic casework analysis. Int J Legal Med 108:68–74

[2] Wilson MR, Polanskey D, Butler J, DiZinno JA, Replogle J, Budowle B. 1995b. Extraction, PCR amplification and sequencing of mitochondrial DNA from human hair shafts. Biotechniques 18:662–669.

[3] Holland, MM and Parsons TJ. 1999. Mitochondrial DNA sequences analysis—validation and use for forensic casework. Forensic Sci. Rev 11: 21– 49.

[4] Kivisild T. 2015. Maternal ancestry and population history from whole mitochondrial genomes. Investig Genet 10; 6:3.

[5] Gill P, Ivanov PL, Kimpton C, Piercy R, Benson N, Tully G, Evett I, Hagelberg E, Sullivan K. 1994. Identification of the remains of the Romanov family by DNA analysis. Nat Genet. 6:130-5.

[6] Brandstätter A, Niederstätter H, Pavlic M, Grubwieser P, Parson W. 2007. Generating population data for the EMPOP database – an overview of the mtDNA sequencing and data evaluation processes considering 273 Austrian control region sequences as example. Forensic Sci Int 2;166: 164-75.

[7] Palo JU, Hedman M, Söderholm N, Sajantila A. 2007. Repatriation and identification of the Finnish World War II soldiers. Croat Med J. 48:528-35.

[8] Schlebusch CM, Naidoo T, Soodyall H. 2009. SNaPshot minisequencing to resolve mitochondrial macro-haplogroups found in Africa. Electrophoresis 30: 3657-64.

[9] King JL, LaRue BL, Novroski NM, Stoljarova M, Seo SB, Zeng X, Warshauer DH, Davis CP, Parson W, Sajantila A, Budowle B. 2014. High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. Forensic Sci Int Genet. 12:128-35.

[10] van Oven M, Vermeulen M, Kayser M. 2011. Multiplex genotyping system for efficient inference of matrilineal genetic ancestry with continental resolution. Investig Genet 23;
2 - 6.

[11] Ballantyne KN, van Oven M, Ralf A, Stoneking M, Mitchell RJ, van Oorschot RA, Kayser M. 2012. MtDNA SNP multiplexes for efficient inference of matrilineal genetic ancestry within Oceania. Forensic Sci Int Genet 6:425-36.

[12] Fendt L, Zimmermann B, Daniaux M, Parson W. 2009. Sequencing strategy for the whole mitochondrial genome resulting in high quality sequences. BMC Genomics 10: 139.
[13] Ramos A, Santos C, Alvarez L, Nogués R, Aluja MP. 2009. Human mitochondrial DNA complete amplification and sequencing: a new validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification. Electrophoresis 30:1587-93.
[14] Gilbert MT, Tomsho LP, Rendulic S, Packard M, Drautz DI, Sher A, Tikhonov A, Dalén L, Kuznetsova T, Kosintsev P, Campos PF, Higham T et al. 2007. Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. Science 317:1927-30.
[15] Gilbert MT, Drautz DI, Lesk AM, Ho SY, Qi J, Ratan A, Hsu CH, Sher A, Dalén L, Götherström A, Tomsho LP, Rendulic S et al. 2008. Intraspecific phylogenetic analysis of Siberian woolly mammoths using complete mitochondrial genomes. Proc Natl Acad Sci U S A. 105: 8327-32.
[16] Miller W, Drautz DI, Janecka JE, Lesk AM, Ratan A, Tomsho LP, Packard M, Zhang Y, McClellan LR, Qi J, Zhao F, Gilbert MT et al. 2009. The mitochondrial genome sequence of the Tasmanian tiger (Thylacinus cynocephalus). Genome Res. 19: 213-20.
[17] Stiller M, Knapp M, Stenzel U, Hofreiter M, Meyer M. 2009. Direct multiplex sequencing (DMPS)--a novel method for targeted high-throughput sequencing of ancient and highly degraded DNA. Genome Res. 19: 1843-8.
[18] Willerslev E, Gilbert MT, Binladen J, Ho SY, Campos PF, Ratan A, Tomsho LP, da Fonseca RR, Sher A, Kuznetsova TV, Nowak-Kemp M, Roth TL et al. 2009. Analysis of complete mitochondrial genomes from extinct and extant rhinoceroses reveals lack of phylogenetic resolution. BMC Evol Biol. 9: 95.
[19] Lindqvist C, Schuster SC, Sun YZ, Talbot SL, Qi J, Ratan A, Tomsho LP, Kasson L, Zeyl E, Aars J, Miller W, Ingólfsson Ó et al. 2010. Complete mitochondrial genome of a Pleistocene jawbone unveils the origin of polar bear. Proc. Natl. Acad. Sci. U. S. A. 107: 5053-5057.
[20] Ho SYW, Gilbert MTP. 2010. Ancient mitogenomics. Mitochondrion (Kidlington) 10: 1-11.
[21] Green RE, Malaspinas AS, Krause J, Briggs AW, Johnson PL, Uhler C, Meyer M, Good JM, Maricic T, Stenzel U, Prüfer K, Siebauer M et al. 2008. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell. 134: 416-26.

[22] Davis C, Peters D, Warshauer D, King J, Budowle B. 2015. Sequencing the hypervariable regions of human mitochondrial DNA using massively parallel sequencing: Enhanced data acquisition for DNA samples encountered in forensic testing. Leg Med (Tokyo). 17:123-7.

[23] Sosa MX, Sivakumar IK, Maragh S, Veeramachaneni V, Hariharan R, Parulekar M, Fredrikson KM, Harkins TT, Lin J, Feldman AB, Tata P, Ehret GB et al. 2012. Next-generation sequencing of human mitochondrial reference genomes uncovers high heteroplasmy frequency. PLoS Comput Biol 8: e1002737.

[24] Parson W, Strobl C, Huber G, Zimmermann B, Gomes SM, Souto L, Fendt L, Delport R, Langit R, Wootton S, Lagacé R, Irwin J. 2013. Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). Forensic Sci Int Genet 7: 543-9.

[25] Parson W, Huber G, Moreno L, Madel MB, Brandhagen MD, Nagl S, Xavier C, Eduardoff M, Callaghan TC, Irwin JA. 2015. Massively parallel sequencing of complete mitochondrial genomes from hair shaft samples. Forensic Sci Int Genet 15:8-15.

[26] Ralf A, van Oven M, Zhong K, Kayser M. 2015. Simultaneous Analysis of Hundreds of Y-Chromosomal SNPs for High-Resolution Paternal Lineage Classification using Targeted Semiconductor Sequencing. Hum Mutat 36:151-9.

[27] van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30: 386-94.

[28] Lao O, Altena E, Becker C, Brauer S, Kraaijenbrink T, van Oven M, Nurnberg P, de Knijff P, Kayser M. 2013. Clinal distribution of human genomic diversity across the Netherlands despite archaeological evidence for genetic discontinuities in Dutch population history. Investig Genet 4:9.

[29] Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147.

[30] Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ et al. 1981. Sequence and organization of the human mitochondrial genome. Nature 290:457-65

[31] Carracedo A, Bär W, Lincoln P, Mayr W, Morling N, Olaisen B, Schneider P, Budowle B, Brinkmann B, Gill P, Holland M, Tully G et al. 2000. DNA commission of the international society for forensic genetics: guidelines for mitochondrial DNA typing. Forensic Sci Int

110:79-85.

[32] Bandelt HJ, Parson W. 2008. Consistent treatment of length variants in the human mtDNA control region: a reappraisal. Int J Legal Med 122: 11-21.

[33] Wilson MR, Allard MW, Monson KL, Miller KWP, Budowle B. 2002a. Recommendations for consistent treatment of length variants in the human mtDNA control region. Forensic Sci Int 129:35–42 25.

[34] Wilson MR, Allard MW, Monson KL, Miller KWP, Budowle B. 2002b. Further discussion of the consistent treatment of length variants in the human mitochondrial DNA control region. Forensic Sci Commun 4:#4.

[35] Fan L, Yao YG. 2011. MitoTool: a web server for the analysis and retrieval of human mitochondrial DNA sequence variations. Mitochondrion 11:351-6.

[36] Hartmann A, Thieme M, Nanduri LK, Stempfl T, Moehle C, Kivisild T, Oefner PJ. 2009. Validation of microarray-based resequencing of 93worldwidemitochondrial genomes. Hum Mutat 30:115–122.

[37] Chaitanya L, van Oven M, Weiler N, Harteveld J, Wirken L, Sijen T, de Knijff P, Kayser M. 2014. Developmental validation of mitochondrial DNA genotyping assays for adept matrilineal inference of biogeographic ancestry at a continental level. Forensic Sci Int Genet 11: 39-51.

[38] Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ.2012. Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30: 434-9.

[39] Seo SB, King JL, Warshauer DH, Davis CP, Ge J, Budowle B. 2013. Single nucleotide polymorphism typing with massively parallel sequencing for human identification. Int J Legal Med. 127:1079-86.

[40] Seo SB, Zeng X, King JL, Larue BL, Assidi M, Al-Qahtani MH, Sajantila A, Budowle B. 2015. Underlying Data for Sequencing the Mitochondrial Genome with the Massively Parallel Sequencing Platform Ion TorrentTM PGMTM. BMC Genomics 16(Suppl 1):S4.

[41] Lippold S, Xu H, Ko A, Li M, Renaud G, Butthof A, Schröder R, Stoneking M.2014. Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. Investig Genet 5:13.

[42] Kayser M, de Knijff P. 2011. Improving human forensics through advances in genetics, genomics and molecular biology. Nat Rev Genet 12: 179-92.

[43] Lyons EA, Scheible MK, Sturk-Andreaggi K, Irwin JA, Just RS. 2013. A high-throughput Sanger strategy for human mitochondrial genome sequencing. BMC Genomics 16:14–881.

[44] Just RS, Scheible MK, Fast SA, Sturk-Andreaggi K, Higginbotham JL, Lyons EA, Bush JM, Peck MA, Ring JD, Diegoli TM, R¨ock AW, Huber GE, Nagl S, Strobl C, Zimmermann B, Parson W, Irwin JA. 2014. Development of forensic-quality full mtGenome haplotypes: success rates with low template specimens. Forensic Sci Int Genet 10:73-79.

[45] Noonan JP, Hofreiter M, Smith D, Priest JR, Rohland N, Rabeder G, Krause J, Detter JC, Pääbo S, Rubin EM. 2005. Genomic sequencing of Pleistocene cave bears. Science 309: 597-600.

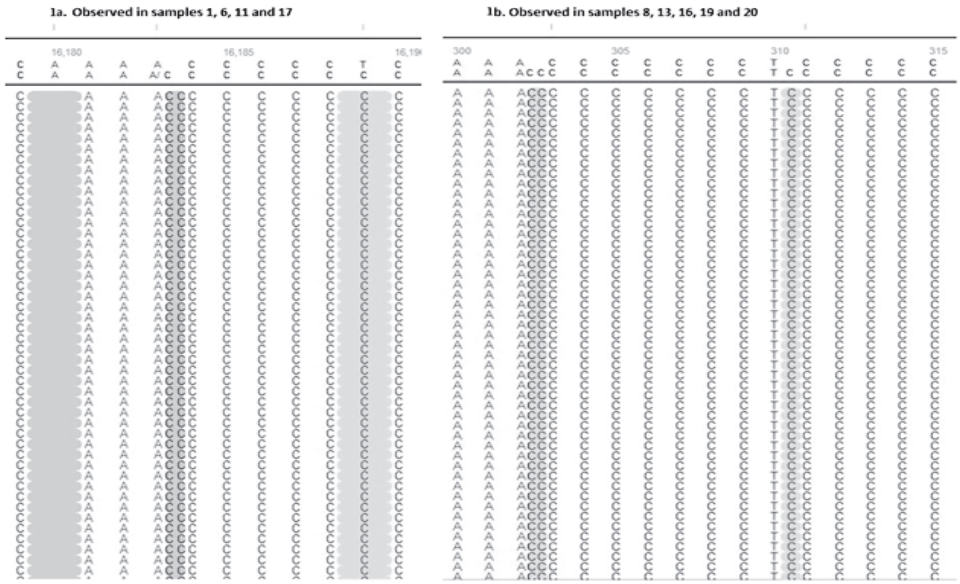
[46] Anderung C, Persson P, Bouwman A, Elburg R, Götherström A. 2008. Fishing for ancient DNA. Forensic Sci. Int. Genet 2: 104-107.

[47] Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Lalueza-Fox C, Rudan P, Brajković D, Kucan Ž, Gušić I, Schmitz R et al. 2009. Targeted Retrieval and Analysis of Five Neandertal mtDNA Genomes. Science 325: 318-321.

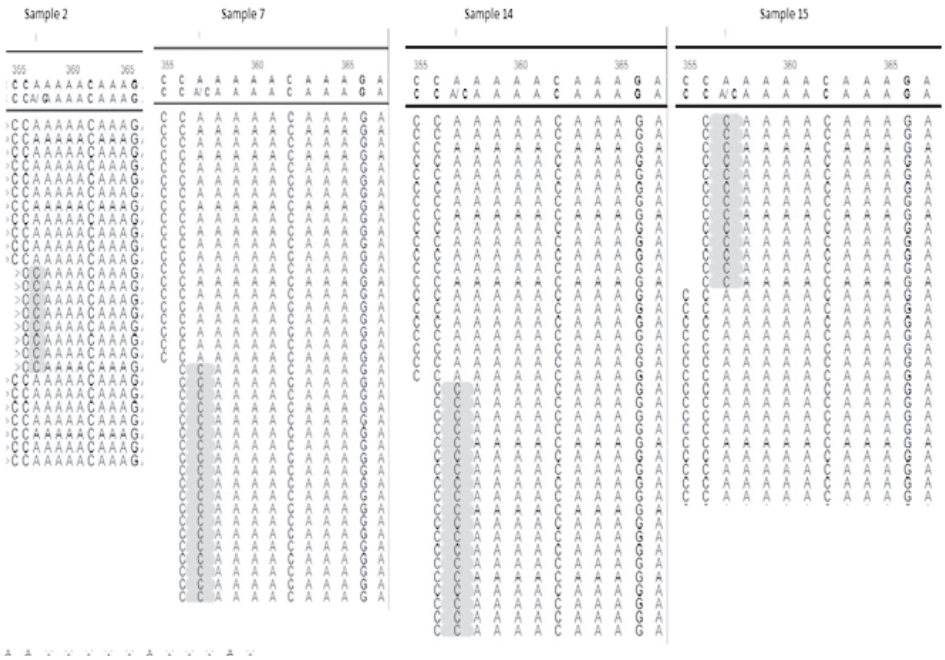
[48] Templeton JE, Brotherton PM, Llamas B, Soubrier J, Haak W, Cooper A, Austin JJ. 2013. DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification. Investig Genet 4: 26.

[49] Kayser M. 2015. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. Forensic Sci Int Genet pii: S1872-4973(15)00032-0.

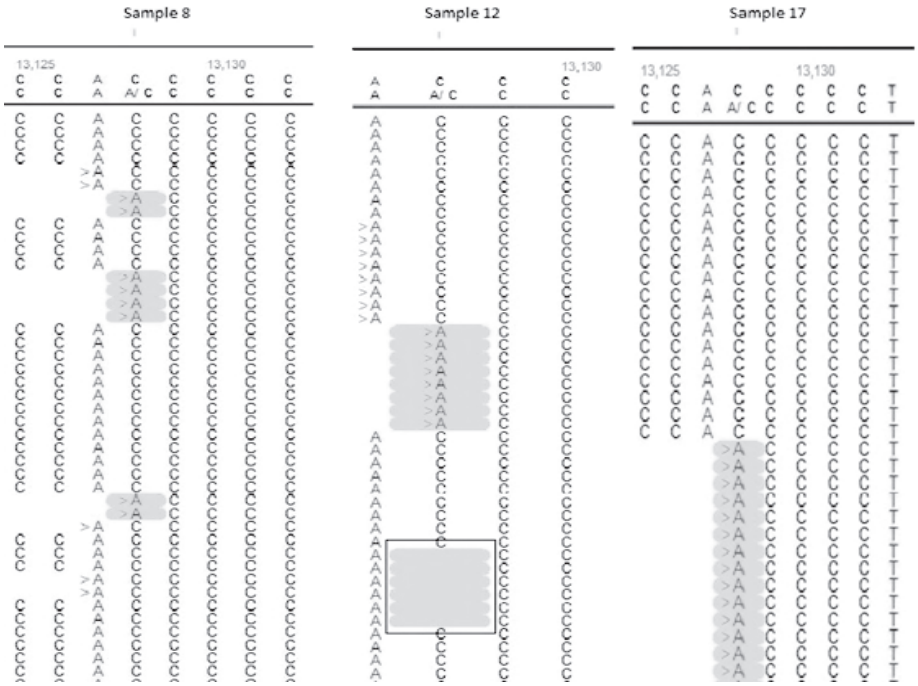
Supplementary materials



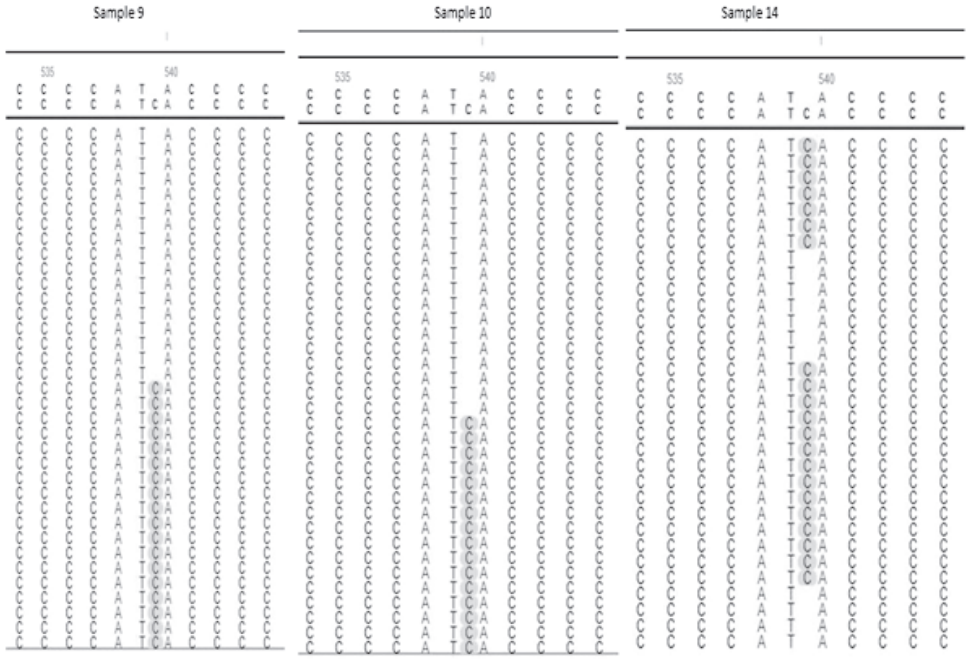
Supplementary Figure 1: Discrepancies were observed in calling of the variants involving length heteroplasmy in the polycytosine stretches of the HVS-I and HVS-II. 1a) Due to an erroneous shift in the bases, the variant was reported as m.16183A>M instead of m.16183A>C in samples 1, 6, 11 and 17. 1b) In the HVS-II region, two insertions of the base C in the np 303-309 were noted for samples 8, 13, 16, 19 and 20.



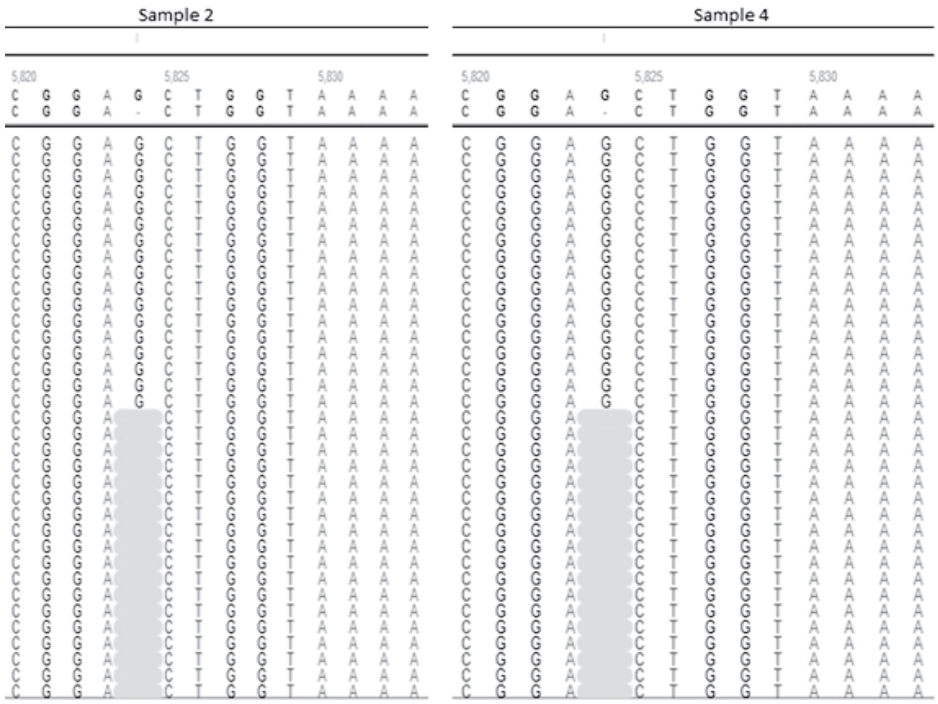
Supplementary Figure 2: In the samples 2, 7, 14 and 15, the variant reported as m.357A>M is specious and appears due to a shift in the base position, possibly because of the polyadenine stretches.



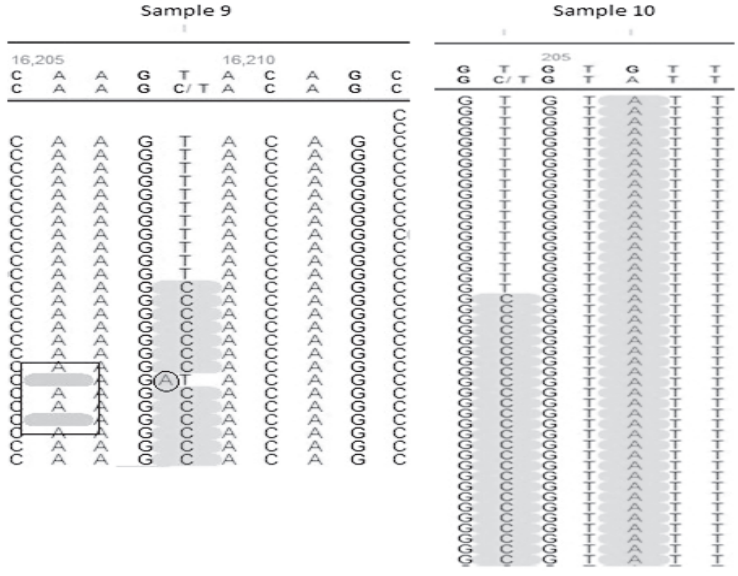
Supplementary Figure 3: In samples 8, 12 and 17, the variant m.13128C>M is spurious due to the shift in the base because of the polycytosine stretches.



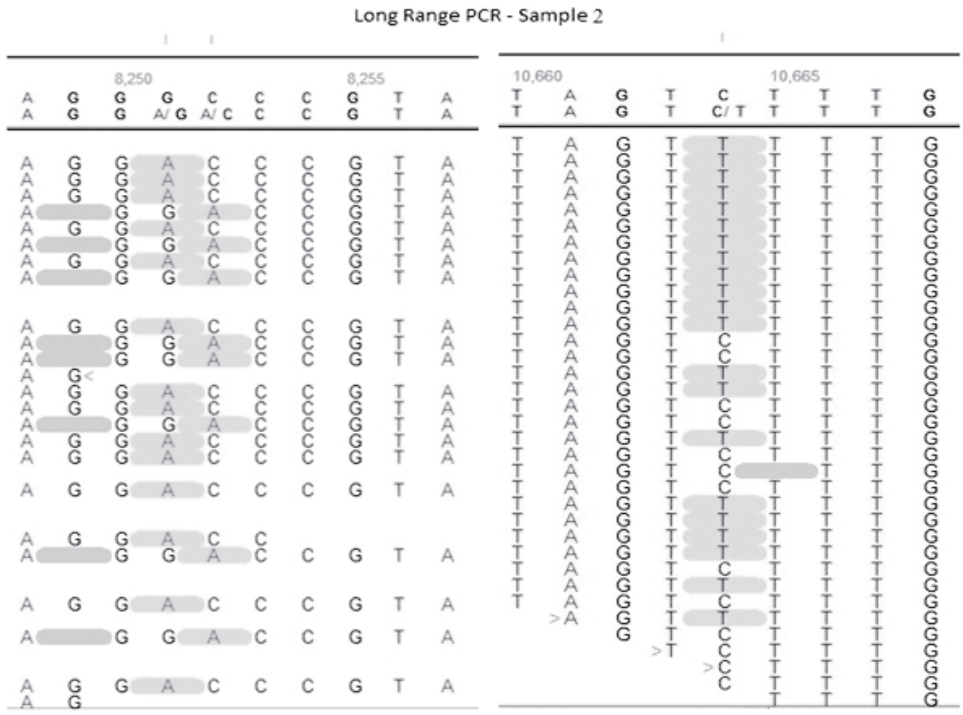
Supplementary Figure 4: Samples 9, 10 and 14 report an insertion m.539T>ins1C (Supp. Figure S4), which was not evident in the Sanger sequencing data.



Supplementary Figure 5: A spurious deletion was noticed at position 5824, in samples 2 and 4, that was not reported in the Sanger sequencing data of Hartmann et al., 2009.



Supplementary Figure 6: Differences to the Sanger sequencing observed in sample 9 at position 16209 and in sample 10 at position 204. Spurious single-bp deletions were observed at some sequence positions in some reads (boxed region) but not reported in the variant table.



Supplementary Figure 7: In sample 2, the polymorphisms m.10664C>Y, m.8251G>R and m.8252C>M were discordant with the tiling approach (m.10664C>T, m.8251G>A and no variant at position 8252, respectively). The erroneous calling at positions 8251 and 8252 could be due to the base position shifts.

Amplicon Number	Primer Pool	Primer Location	Primer	Degenerate Primer 1	Degenerate Primer 2	Degenerate Primer 3
1	Pool_2_F	120_279	CGCAGTATCTGTCTTTGATTCCT			
2	Pool_1_F	221_388	GCTTGTAGGACATAATAATAACAATTGA	GCTTGTAAGACATAATAATAACAATTGA	GCTTGTAAGACATAGTAATAACAATTGA	GCTTGTAGGACATAGTAATAACAATTGA
3	Pool_2_F	317_497	CTTCTGGCCACAGCACTTAAAC			
4	Pool_1_F	410_619	GCGGTATGCACTTTTAACAGTCA			
5	Pool_2_F	500_715	CGCCCATCCTACCCAGCAC			
6	Pool_1_F	619_786	TTAGACGGGCTCACATCACC			
7	Pool_2_F	720_900	AGTTCACCCTCTAAATCACCACGA			
8	Pool_1_F	815_1030	CAGCAGTGATTAACCTTTAGCAATAAAC			
9	Pool_2_F	928_1133	AAGCCGGCGTAAAGAGTGTTTTAG			
10	Pool_1_F	1023_1236	CGAAAGTGGCTTTAACATATCTGAACAC			
11	Pool_2_F	1136_1339	CAGAACACTACGAGCCACAGCTT			
12	Pool_1_F	1237_1442	ACCACCTCTTGCTCAGCCTA			
13	Pool_2_F	1340_1545	CCCATGAGGTGGCAAGAAATG			
14	Pool_1_F	1443_1648	TAGAGTGCTTAGTTGAACAGGGC			
15	Pool_2_F	1546_1751	ATAGAGGAGACAAGTCGTAACATGG			
16	Pool_1_F	1649_1848	CAACTTAACTTGACCGCTCTGAG			
17	Pool_2_F	1755_1936	AGAAATTGAAACCTGGCGCAATAGA			
18	Pool_1_F	1869_2020	AGAAATAACTTTGCAAGGAGAGCC			
19	Pool_2_F	1966_2165	GTGGGAAGATTTATAGGTAGAGGC			
20	Pool_1_F	2061_2266	CAGAACCCTCTAAATCCCCTTG			
21	Pool_2_F	2164_2369	CCCATAGTAGGCCTAAAAGCAG			
22	Pool_1_F	2267_2431	TGGACCAATCTATCACCCTATAGAAG			
23	Pool_2_F	2359_2575	CTGAACTGACAATTAACAGCCCAATAT			
24	Pool_1_F	2435_2646	CAGGCATGCTCATAAGGAAAGGTTAA			
25	Pool_2_F	2576_2781	ACCCTAACCGTGCAAAGGTAG			
26	Pool_1_F	2679_2884	TTGACCTGCCCGTGAAGAGG			
27	Pool_2_F	2775_2987	AGGTCCTAAACTACCAAACCTGCATT			
28	Pool_1_F	2885_3090	TATACTCAATTGATCCAATAACTTGACC			
29	Pool_2_F	2988_3193	CGATGTTGGATCAGGACATCC			
30	Pool_1_F	3089_3297	CCAGGTCGGTTTCTATCTACTTCAA			
31	Pool_2_F	3194_3399	TAGTATTATACCCACACCCACCCAA			
32	Pool_1_F	3297_3502	TCTTAACAACATACCCATGGCCAAC			
33	Pool_2_F	3388_3605	CTAGGCTATATACAACTACGCAAAGG			
34	Pool_1_F	3503_3708	CTACCATCACCCTCTACATCAC			
35	Pool_2_F	3606_3811	AGGCCTCCTATTTATTCTAGCCAC			
36	Pool_1_F	3706_3914	CGAGCAGTAGCCCAAACAATCTC			
37	Pool_2_F	3828_3997	ACTCCTGCCATCATGACCCT			
38	Pool_1_F	3914_4120	GGGAGTCCGAACTAGTCTCAGG			
39	Pool_2_F	4018_4225	ACCACTACAATCTTCCTAGGAACAAC			
40	Pool_1_F	4121_4326	GAATTCGAACAGCATACCCCC			
41	Pool_2_F	4224_4429	CATACCCATTACAATCTCCAGCATTC			
42	Pool_1_F	4327_4532	TTCTAGGACTATGAGAATCGAACCCA			
43	Pool_2_F	4430_4635	CCCATACCCCGAAAATGTTGGT			
44	Pool_1_F	4533_4747	CTAAGCTCGCACTGATTTTTTACCTG			
45	Pool_2_F	4636_4841	CCATCAAGTATTTCCTCACGCAA			
46	Pool_1_F	4747_4945	TACTCATCATTAATAATCATAATAGCTATAGC	TACTCATCATTAATAATCATAATGGCTATAGC		
47	Pool_2_F	4842_5047	ACATCCGGCCTGCTTCTTCT			
48	Pool_1_F	4945_5150	TCTTATCCATCATAGCAGGCAGTTG			
49	Pool_2_F	5048_5253	TCTACCGTACAACCCTAACATAACC			
50	Pool_1_F	5151_5356	CTACTACTATCTCGCACCTGAAACA			
51	Pool_2_F	5254_5459	CCAAATGGGCCATTATCGAAGAATTC			
52	Pool_1_F	5357_5562	AATCTACTCCACCTCAATCACACTAC			
53	Pool_2_F	5460_5665	GCCCTTACCACGCTACTCCT			
54	Pool_1_F	5564_5744	GCAATACTTAATTTCTGCAACAGCTAAGGA	GCAATACTTAATTTCTGTAACAGCTAAGGA		
55	Pool_2_F	5666_5871	TGGGACTTAAACCCACAAACACTTAG			

56	Pool_1_F	5781_5974	CTGCTTCTTCGAATTTGCAATTCAATA				
57	Pool_2_F	5872_6077	CTTCACTCAGCCATTTTACCTCAC				
58	Pool_1_F	5975_6180	ATGACTGGAGTCCTAGGCA				
59	Pool_2_F	6074_6283	CGTCACAGCCCATGCATTTGTAAT				
60	Pool_1_F	6181_6383	CGTTTCCCCGCATAAACAACATAAG				
61	Pool_2_F	6322_6485	GGAGCCTCCGTAGACCTAACC				
62	Pool_1_F	6384_6592	GCCATCAATTTCATCACAACAATTATC				
63	Pool_2_F	6486_6695	CTTCTCCTATCTCTCCCAGTCCTAG				
64	Pool_1_F	6593_6803	ATACCAACACCTATTCTGATTTTTCG				
65	Pool_2_F	6696_6901	AAAGAACCATTTGGATACATAGGTATG				
66	Pool_1_F	6799_7004	TAGACACGAGCATATTTCACCTC				
67	Pool_2_F	6902_7107	ATGATCTGCTGCAGTGCTCTGA				
68	Pool_1_F	7005_7210	CACGACACGTACTACGTTGTAG				
69	Pool_2_F	7108_7299	GCTACACCCTAGACCAAACC				
70	Pool_1_F	7211_7416	GCCCCGACGTTACTCGGA				
71	Pool_2_F	7337_7499	CGAAGCGAAAAGTCCTAATAGTAGAA				
72	Pool_1_F	7419_7628	CGAAGAACCCGTATACATAAAATCTAGACA				
73	Pool_2_F	7566_7710	GGCTAAATCCTATATATCTTAATGGCAC				
74	Pool_1_F	7623_7828	CTTCCCCTATCATAGAAGAGCTTATC				
75	Pool_2_F	7702_7931	GTATGCCCTTTTCCTAACACTCACAAC				
76	Pool_1_F	7829_8034	CGCATCCTTTACATAACAGACGAG				
77	Pool_2_F	7932_8137	TAATCTTCAACTCCTACATACTTCCC				
78	Pool_1_F	8035_8240	TCGTATAATAATTACATCACAAGACGTC				
79	Pool_2_F	8132_8343	ACTTTCACCGCTACACGACCG				
80	Pool_1_F	8241_8446	TTGAAATAGGGCCCGTATTTACCCT				
81	Pool_2_F	8344_8549	ACCAACACCTCTTTACAGTGAAATGC				
82	Pool_1_F	8447_8652	ATATTTAAACACAAACTACCACCTACCT				
83	Pool_2_F	8576_8783	CTACCCGCCGCAGTACTGAT				
84	Pool_1_F	8653_8858	ATCACCACCCAACAATGACTAATCAAA				
85	Pool_2_F	8756_8961	TTTTTATTGCCACAACTAACCTCCTC				
86	Pool_1_F	8859_9064	CACAGTGATTATAGGCTTTCGCTCTA				
87	Pool_2_F	8962_9167	ACCATCAGCCTACTCATTCAACCA				
88	Pool_1_F	9065_9278	CAATATCAACCATTAACCTTCCCTCTA				
89	Pool_2_F	9168_9376	CACACTTCTAGTAAGCCTCTACCTGCAC				
90	Pool_1_F	9271_9476	TCTCAGCCCTCCTAATGACC				
91	Pool_2_F	9374_9579	ATGATGGCGCGATGTAACACGA				
92	Pool_1_F	9477_9682	GTTTTTTTCTTCGCAGGATTTTTCTGA				
93	Pool_2_F	9580_9785	ATCCCCTAGAAGTCCCACTC				
94	Pool_1_F	9683_9888	AATTCAAGCACTGCTTATTACAATTTTAC				
95	Pool_2_F	9786_9991	GGCTCAACATTTTTTGTAGCCACAG				
96	Pool_1_F	9889_10094	CATCCAAACATCACTTTGGCTTCG				
97	Pool_2_F	9964_10194	ATGTCTCCATCTATTGATGAGGGTCTTAC				
98	Pool_1_F	10085_10300	CAACACCCTCCTAGCCTTACTA				
99	Pool_2_F	10198_10385	CCCGCGTCCCTTTCTCCATA				
100	Pool_1_F	10308_10480	CCTGCCACTAATAGTTATGTCATCCC				
101	Pool_2_F	10386_10609	GGATTAGACTGAACCGAATTGGTAT				
102	Pool_1_F	10502_10715	ATTTACCATCTCACTTCTAGGAATACTAGT				
103	Pool_2_F	10610_10815	AACCCTCAACACCCACTCCC				
104	Pool_1_F	10713_10918	TCCAACACATATGGCCTAGACTAC				
105	Pool_2_F	10816_11021	AAAACACATAATTTGAATCAACACAACCAC				
106	Pool_1_F	10919_11114	CCAACCTTTTCCTCCGACCC				
107	Pool_2_F	11022_11227	CACTATCACGAAAAAAAACTCTACCTCT				
108	Pool_1_F	11115_11330	ATATCTTCTTCGAACCACACTTATCC				
109	Pool_2_F	11228_11433	TCCCTTCCCCTACTCATCGC				
110	Pool_1_F	11331_11536	CCAACAACTTAATATGACTAGCTTACA				

111	Pool_2_F	11434_11639	CGCTGGGTCAATAGTACTTGC			
112	Pool_1_F	11538_11707	CCCTTCCTTGTACTATCCCTATGAG			
113	Pool_2_F	11640_11845	TAGCCCTCGTAGTAACAGCC			
114	Pool_1_F	11743_11951	CCTAGCAAACTCAAACTACGAACG			
115	Pool_2_F	11846_12051	AGCCTCGCTAACCTCGCCTT			
116	Pool_1_F	11949_12154	GACTCAACATACTAGTCACAGCC			
117	Pool_2_F	12086_12240	CCCCATTCTCCTCCTATCCCT			
118	Pool_1_F	12181_12352	ACAGAGGCTTACGACCCCTTATTT			
119	Pool_2_F	12258_12463	CTTTCTCAACTTTTAAAGGATAACAGC			
120	Pool_1_F	12361_12566	ACCCTAACCCTGACTTCCCT			
121	Pool_2_F	12464_12671	CCTTTATTATCAGTCTCTTCCCCAC			
122	Pool_1_F	12564_12772	CTCCCTAAGCTTCAAACTAGACTACTTC			
123	Pool_2_F	12670_12875	CCAAACATTAATCAGTTCTTCAAATATCT			
124	Pool_1_F	12773_12978	GCGTAGGAATTATATCCTTCTTGCT			
125	Pool_2_F	12876_13081	CGGTTTCATCCTCGCCTTAG			
126	Pool_1_F	12979_13184	GGCCTCCTCCTAGCAGCA			
127	Pool_2_F	13082_13290	CAAGCACTATAGTTGTAGCAGGAATC			
128	Pool_1_F	13185_13390	CACCACTCTGTTCGCAGCAG			
129	Pool_2_F	13279_13493	GTTACAATCGGCATCAACCAACC			
130	Pool_1_F	13391_13596	ATGAACAAGATATTCGAAAAATAGGAGG			
131	Pool_2_F	13494_13699	CACAGGTTTCTACTCCAAAGACCA			
132	Pool_1_F	13597_13802	GCCTATAGCACTCGAATAATTCTTCTC			
133	Pool_2_F	13694_13905	CCATTAAACGCCTGGCAGC			
134	Pool_1_F	13806_14000	CCCTCGCTGTCACTTTCCTAGG			
135	Pool_2_F	13906_14111	AACATACTCGGATTCTACCCTAGC			
136	Pool_1_F	14040_14192	AGCACCAAATCTCCACCTCCA			
137	Pool_2_F	14112_14317	CTTCCCACTCATCCTAACCC			
138	Pool_1_F	14215_14420	TACTAATCAACGCCCATAATCATACAA			
139	Pool_2_F	14338_14493	CCACCCCATCATACTCTTTCACCC			
140	Pool_1_F	14421_14626	CTGACCCCCATGCCTCAG			
141	Pool_2_F	14524_14733	ATAACCTCCCCCAAAATTCAGAATAAT			
142	Pool_1_F	14627_14832	AACCCCATTACTAAACCCACACTC			
143	Pool_2_F	14730_14935	CAACTACAAGAACACCAATGACCC			
144	Pool_1_F	14833_15042	ATGATGAAACTTCGGCTCACTCCT			
145	Pool_2_F	14936_15141	TCATCAATCGCCCACATCACTC			
146	Pool_1_F	15039_15244	TCGGGCGAGGCCTATATTAC			
147	Pool_2_F	15142_15347	CCTCCCGTGAGGCCAAATAT			
148	Pool_1_F	15240_15448	GACGCAGGCTACTCAGTAGACAGTCC			
149	Pool_2_F	15329_15553	CTCCACCTCCTATTCTTGCACG			
150	Pool_1_F	15451_15656	CCTTCTCTCCTTAATGACATTTAACACT			
151	Pool_2_F	15554_15759	CCCGAATGATATTTCCTATTCGCCT			
152	Pool_1_F	15657_15862	TCCCCATCCTCCATATATCCAAAC			
153	Pool_2_F	15743_15965	CTCATTCTAACCTGAATCGGAGGAC			
154	Pool_1_F	15863_16068	GAAAACAAAATACTCAAATGGCCTGT			
155	Pool_2_F	15986_16156	GCACCCAAAGCTAAGATTCTAAT			
156	Pool_1_F	16094_16250	TCGTACATTACTGCCAGCC	TCGTACATTACTGCCAGTC		
157	Pool_2_F	16194_16410	ATGCTTACAAGCAAGCACAGCAA	ATGCTTACAAGCAAGTACAGCAA		
158	Pool_1_F	16363_16511	CCCCATGGATGACCCC			
159	Pool_2_F	16407_16569	CCTCCGTGAAATCAATATCCCG			
160	Pool_2_F	16452_52	TCCGGGCCCATAACACTTG			
161	Pool_1_F	16563_180	CACGATGGATCACAGGTCTATC			

1	Pool_2_R	120_279	ATGTCTGTGTGGAAAGCGGC	ATGTCTGTGTGGAAAGTGGC		
2	Pool_1_R	221_388	TGGTTAGGCTGGTGTTAGGGT	GGGGTTGTGTTTGATGAGATTAGTAG		
3	Pool_2_R	317_497	GGGGTTGTATTGATGAGATTAGTAG			
4	Pool_1_R	410_619	AACATTTTCAGTGTATTGCTTTGAG			
5	Pool_2_R	500_715	GGAACGGGGATGCTTGCATG			
6	Pool_1_R	619_786	CGTTTTGAGCTGCATTGCTGC			
7	Pool_2_R	720_900	GGTGGCTGGCACGAAATTGA			
8	Pool_1_R	815_1030	CACTTTCGTAGTCTATTTTGTGTCAAC			
9	Pool_2_R	928_1133	GAGCAGTTTTGTTGATTTAACTGTTGA			
10	Pool_1_R	1023_1236	GAGGTTGATCGGGGTTTATCGA			
11	Pool_2_R	1136_1339	CTACACCTTGACCTAACGTCTTTAC			
12	Pool_1_R	1237_1442	CTCTTAGTTTTACTGCTAAATCCACCTT			
13	Pool_2_R	1340_1545	ATAAATGCGTAGGGGTTTTAGTTAAATG			
14	Pool_1_R	1443_1648	AAATCTCCTAAGTGTAAGTTGGGTGC			
15	Pool_2_R	1546_1751	TCGCCTATACTTTATTTGGGTAAATGG			
16	Pool_1_R	1649_1848	AAGGTATAGGGGTTAGTCCTTGCT			
17	Pool_2_R	1755_1936	AGCTGTTCTTAGGTAGCTCGTCTG			
18	Pool_1_R	1869_2020	CAGCTATCACCAGGCTCGGT			
19	Pool_2_R	1966_2165	GGTGTTAAATTTTTTACTCTCTCTACAAG			
20	Pool_1_R	2061_2266	ATTGGGTGTGAGGAGTTCAGTTATAT			
21	Pool_2_R	2164_2369	TGTCAGTTCAGTGTTTTAATCTGACG			
22	Pool_1_R	2267_2431	GTTGGGTTGACAGTGAGGGTAA			
23	Pool_2_R	2359_2575	ACCGCGGCCGTTAAACATGT			
24	Pool_1_R	2435_2646	TCGTGGAGCCATTCATACAGGT			
25	Pool_2_R	2576_2781	AGGACCTGTGGGTTGTTAGGT			
26	Pool_1_R	2679_2884	GTAGTTCGCTTTGACTGGTGAAGT			
27	Pool_2_R	2775_2987	AGGTCGTAAACCCTATTGTTGATATG			
28	Pool_1_R	2885_3090	CTGGATTACTCCGGTCTGAACT			
29	Pool_2_R	2988_3193	AGTTGAGATGATATCATTTACGGGGG			
30	Pool_1_R	3089_3297	AGAGGAATTGAACCTCTGACTGTAAA			
31	Pool_2_R	3194_3399	TATATAGCCTAGAATTTTTCGTTCGGT			
32	Pool_1_R	3297_3502	ATGTGGCGGGTTTTAGGGGC			
33	Pool_2_R	3388_3605	AGGTTGAGGTTGACCAGGGG			
34	Pool_1_R	3503_3708	TCGCAGTGCGCCGATCAG			
35	Pool_2_R	3606_3811	GTGTTGTGATAAGGGTGGAGAG			
36	Pool_1_R	3706_3914	CCTTCGGCAAGGTCGAAGG			
37	Pool_2_R	3828_3997	TGTTTGTGTATTCGGCTATGAAGAATA			
38	Pool_1_R	3914_4120	ATAAGAACACGGAGGTTAGAAGTAGGGT			
39	Pool_2_R	4018_4225	TGGAGACATATCATATAAGTAATGCTAGG			
40	Pool_1_R	4121_4326	ATAAGGGGTTTAAGCTCCTATTATTT			
41	Pool_2_R	4224_4429	CCCGATAGCTTATTTAGCTGACCT			
42	Pool_1_R	4327_4532	CGCTGTGATGAGTGTGCCTG			
43	Pool_2_R	4430_4635	CAGCTTCTGTGGAACGAGGG			
44	Pool_1_R	4533_4747	TATTGATTGGTAGTATTGGTTATGGTTC			
45	Pool_2_R	4636_4841	CAGAGGGGTGCCTTGGGTAA			
46	Pool_1_R	4747_4945	TTGAGAGAGTGAGGAGAAGGCTT			
47	Pool_2_R	4842_5047	ACTGCTATTTATTCATCCTATGTGGG			
48	Pool_1_R	4945_5150	GGTCGTGGTGCTGGAGTTTA			
49	Pool_2_R	5048_5253	GCAAAAAGCCGGTTACGGG			
50	Pool_1_R	5151_5356	AGGCGTAGGTAGAAGTAGAGGTT			
51	Pool_2_R	5254_5459	GATGAGTGTGGGGAGGAATG			
52	Pool_1_R	5357_5562	AACTTACTGAGGGCTTTGAAGGC			
53	Pool_2_R	5460_5665	TTGGTCTAGTAAGGGCTTAGCTTAA			
54	Pool_1_R	5564_5744	GGCGGCGGGAGAAGTAGATT			
55	Pool_2_R	5666_5871	CATTGGACTGTAAATCTAAAGACAGG			

56	Pool_1_R	5781_5974	GCGCCGAATAATAGGTATAGTGTTC				
57	Pool_2_R	5872_6077	GACGATAACGTTGTAGATGTGGTC				
58	Pool_1_R	5975_6180	CCATATCGGGGGCACCGATT				
59	Pool_2_R	6074_6283	GTTCAACCTGTTCCTGCTCC				
60	Pool_1_R	6181_6383	CCCTAAGATAGAGGAGACACCT				
61	Pool_2_R	6322_6485	AGGACTGCTGTGATTAGGACGG				
62	Pool_1_R	6384_6592	AGAATGGGGTCTCCTCCTCC				
63	Pool_2_R	6486_6695	TTTTCCGGAGTAGTAAGTTACAATATGG				
64	Pool_1_R	6593_6803	GTCTACGTCTATTCCTACTGTAAATATATGG				
65	Pool_2_R	6696_6901	TTCATATTGCTTCCGTGGAGTGT				
66	Pool_1_R	6799_7004	TAGTACGATGTCTAGTGATGAGTTTG				
67	Pool_2_R	6902_7107	CTGAGAAATAGGGGAAATCAGTGAATG				
68	Pool_1_R	7005_7210	ATTCCGGATAGGCCGAGAAAGT				
69	Pool_2_R	7108_7299	TTACTGCTGTTAGAGAAATGAATGAGC				
70	Pool_1_R	7211_7416	ATGTGTGGTAGGGTGGGG				
71	Pool_2_R	7337_7499	CCATGGGGTTGGCTTGAAAC				
72	Pool_1_R	7419_7628	GGAAGTAGCGTCTTGTAGACCTACTTG				
73	Pool_2_R	7566_7710	GGGCATACAGGACTAGGAAGCA				
74	Pool_1_R	7623_7828	TACGGGATGGGAGGGCGATGA				
75	Pool_2_R	7702_7931	GTCCGCCCGTAGTCGGTGT				
76	Pool_1_R	7829_8034	ATGGGGGCTTCAATCGGGAG				
77	Pool_2_R	7932_8137	GAAAGTGGTTTGGTTTAGACGTCC				
78	Pool_1_R	8035_8240	AGATTTTTAGGGGAATTAATTCTAGGAC				
79	Pool_2_R	8132_8343	TCTCTTAATCTTTAACTTAAAAGGTTAATGC				
80	Pool_1_R	8241_8446	TTTTAGTTGGGTGATGAGGAATAGTGT				
81	Pool_2_R	8344_8549	GAAGCGAACAGATTTTCGTTCATTTTG				
82	Pool_1_R	8447_8652	TAGTCGGTTGTTGATGAGATATTTGGA				
83	Pool_2_R	8576_8783	CGAGGAGGTTAGTTGTGGCA				
84	Pool_1_R	8653_8858	CCCGCTCATAAGGGGATGG				
85	Pool_2_R	8756_8961	TTCGATAATAACTAGTATGGGGATAAG				
86	Pool_1_R	8859_9064	CTAGGGTGGCGCTTCCAATT				
87	Pool_2_R	8962_9167	AAAACGTAGGCTTGGATTAAGGC				
88	Pool_1_R	9065_9278	GGCTGAGAGGCCCCTGT				
89	Pool_2_R	9168_9376	CATTGGTATATGGTTAGTGTGTTGGTTAG				
90	Pool_1_R	9271_9476	TTCTGAGGTAATAAATAGGATTATCCC				
91	Pool_2_R	9374_9579	TTAGCGGGTGATGCCTGTT				
92	Pool_1_R	9477_9682	ATTTGGTTTCGGTTGTTTTCTATTAGAC				
93	Pool_2_R	9580_9785	GTAGATGCCGTCGGAAATGG				
94	Pool_1_R	9683_9888	TAAAGTGAAATATTAGTTGGCGGATGA				
95	Pool_2_R	9786_9991	TAAGACCCTCATCAATAGATGGAGA				
96	Pool_1_R	9889_10094	GAGGGTGTTGATTATTAAAATTAAGGC				
97	Pool_2_R	9964_10194	GGGATATAGGGTCGAAGCCG				
98	Pool_1_R	10085_10300	GTTTGTAGGGCTCATGGTAGG				
99	Pool_2_R	10198_10385	TTTTTGTAGTCACTCATAGGCCAGA				
100	Pool_1_R	10308_10480	TGAGGGGCATTTGGTAAATATGATTATCA				
101	Pool_2_R	10386_10609	ATGAGAGTAGCTATAATGAACAGCGAT				
102	Pool_1_R	10502_10715	GGAGATTGAGACTAGTAGGGCTAGG				
103	Pool_2_R	10610_10815	TTGGAAAGTCATGTCAGTGGTAGTA				
104	Pool_1_R	10713_10918	GGAACAGCTAAATAGGTTGTTGTTGAT				
105	Pool_2_R	10816_11021	GTTCACTGGATAAGTGGCGTTG				
106	Pool_1_R	10919_11114	AAAATATGATTAGTTCTGTGGCTGTG				
107	Pool_2_R	11022_11227	GCCTACTAGGGTGTAGAATAGGAA				
108	Pool_1_R	11115_11330	CTCAGGAGTTTGATAGTTCTTGG				
109	Pool_2_R	11228_11433	ATGGGGGCTTCGACATGGG				
110	Pool_1_R	11331_11536	GTAGGCTATGTGTTTTGTCAGGG				

111	Pool_2_R	11434_11639	TGTGGCTGATTGAAGAGTATGCAATG				
112	Pool_1_R	11538_11707	ATGAGAATGACTGCGCCGGT				
113	Pool_2_R	11640_11845	TGCTAGAAGTCATCAAAAAGCTATTAG				
114	Pool_1_R	11743_11951	GTCCTGTAAGTAGGAGAGTGATATTTTGATC				
115	Pool_2_R	11846_12051	GTGAATGAGGGTTTTATGTTGTTAATG				
116	Pool_1_R	11949_12154	GGTTAAACTATATTTACAAGAGGAAAACC				
117	Pool_2_R	12086_12240	GGGCATGAGTTAGCAGTTCTTGT				
118	Pool_1_R	12181_12352	AGTGTGCATGGTTATTACTTTTTATTTGG				
119	Pool_2_R	12258_12463	TGGATGCGACAATGGATTTTACATAAT				
120	Pool_1_R	12361_12566	GAGAGCTGGGTTGTTTGGGT				
121	Pool_2_R	12464_12671	GGGTCTGAGTTTATATATCACAGTGAGAA				
122	Pool_1_R	12564_12772	CCTCTCAGCCGATGAACAGT				
123	Pool_2_R	12670_12875	ATATCGCCGATACGGTTGTATAGG				
124	Pool_1_R	12773_12978	TAGTAGTGGGGTGAGGCTTG				
125	Pool_2_R	12876_13081	AGTGGATGTAGGGCTGAGACT				
126	Pool_1_R	12979_13184	ATAGCGCCTAAGCATAGTGTTAGAGT				
127	Pool_2_R	13082_13290	GCCGATTGTAACTATTATGAGTCCTAGTTG				
128	Pool_1_R	13185_13390	TGTTAAGGTTGTGGATGATGGACC				
129	Pool_2_R	13279_13493	AGGAAAGGTATTCCTGCTAATGCTAG				
130	Pool_1_R	13391_13596	GCTTGTCAGGGAGGTAGCGA				
131	Pool_2_R	13494_13699	TAATGGGGTTTAGTAGGGTGGG				
132	Pool_1_R	13597_13802	GTGAGTTTTAGGTAGAGGGGGA				
133	Pool_2_R	13694_13905	GGAGAAATAAAATGTGCATAGTGGGG				
134	Pool_1_R	13806_14000	GGTCTAGGAGGAGTAGGGGCA				
135	Pool_2_R	13906_14111	AAGAAAGAGAGGAAGTAAAGTTTAATTATG				
136	Pool_1_R	14040_14192	TGTTGGTGTATATATTGTAATTGAGATTGC				
137	Pool_2_R	14112_14317	TAGTGTAGGAAGCTGAATAATTTATGAAG				
138	Pool_1_R	14215_14420	GGGTTGAGGTCTTGGTGAGT				
139	Pool_2_R	14338_14493	GGGGAATGATGGTTGTCTTTGGA				
140	Pool_1_R	14421_14626	TGTGGGGTTTTTCTTCTAAGCCTTC				
141	Pool_2_R	14524_14733	GTTGAAATACAACGATGGTTTTTCATATC				
142	Pool_1_R	14627_14832	GCGGAGATGTTGGATGGGGT				
143	Pool_2_R	14730_14935	AAAGGCGGTTGAGGCGTCTG				
144	Pool_1_R	14833_15042	CCGATGTGTAGGAAGAGGCAGATAAAG				
145	Pool_2_R	14936_15141	ACATAGCCTATGAAGGCTGTTGCT				
146	Pool_1_R	15039_15244	TCCTCAGATTCATTGAACTAGGTCTG				
147	Pool_2_R	15142_15347	GCAAGAATAGGAGGTGGAGTGT				
148	Pool_1_R	15240_15448	GAGAAGTAAGCCGAGGGCGTCT				
149	Pool_2_R	15329_15553	CTTGATGTGCGGAGGGGTGT				
150	Pool_1_R	15451_15656	TTATTGCTAGGATGAGGATGGATAGTA				
151	Pool_2_R	15554_15759	ATTCAGGTTAGAATGAGGAGGTCT				
152	Pool_1_R	15657_15862	AATTAGGGAGATAGTTGGTATTAGGAT				
153	Pool_2_R	15743_15965	TTTTCTCTGATTTGTCCTTGGAAAAAG				
154	Pool_1_R	15863_16068	AGTCAATACTTGGGTGGTACCCA				
155	Pool_2_R	15986_16156	GCACCCAAAGCTAAGATTCTAAT	CTACAGGTGGTTAAGTATTTATGGTAC			
156	Pool_1_R	16094_16250	GAGTTGCAGTTGATGTATGATAG	GAGTTGCAGTTGATGTGTGTGATAG			
157	Pool_2_F	16194_16410	ATGCTTACAAGCAAGCACACAGCAA	ATGCTTACAAGCAAGTACAGCAA			
158	Pool_1_R	16363_16511	GTAGGAACCAGATGTCGGATAC				
159	Pool_2_R	16407_16569	CATCGTGATGTCTTATTTAAGGGGAA				
160	Pool_2_R	16452_52	AAATGCATGGAGAGCTCCC				
161	Pool_1_R	16563_180	AATATTGAACGTAGGTGCGATAAAT				

Supplementary Table 1: The primer sequences used in this study without the proprietary AmpliSeq™ modification.

Sample ID	1	2		3	4	5	6	7	8	9	10
Haplotypes	m.73A>G	m.73A>G	m.10398A>G	m.263A>G	m.73A>G	m.73A>G	m.72T>C	m.73A>G	m.73A>G	m.73A>G	m.73A>G
	m.152T>C	m.146T>C	m.10589G>A	m.315ins1C	m.151C>T (m.151C>del)	m.212T>C	m.263A>G	m.152T>C	m.150C>T	m.146T>C	m.146T>C
	m.263A>G	m.195T>C	m.10664C>T	m.477T>C	m.152T>C (152.1C)	m.263A>G	m.315ins1C	m.263A>G	m.152T>C	m.199T>C	m.152T>C
	m.315ins1C	m.247G>A	m.10688G>A	m.750A>G	m.263A>G	m.309C>ins2C	m.750A>G	m.315ins1C	m.199T>C	m.263A>G	m.207G>A
	m.523A>del(m.514C>del)	m.315ins1C	m.10810T>C	m.1438A>G	m.309C>ins1C	m.315ins1C	m.1438A>G	m.523A>del(m.514C>del)	m.263A>G	m.309C>ins1C	m.249A>del
	m.524C>del(m.515A>del)	m.498C>del	m.10873T>C	m.3010G>A	m.315ins1C	m.709G>A	m.2706A>G	m.524C>del(m.515A>del)	m.295C>T	m.315ins1C	m.263A>G
	m.750A>G	m.719G>A	m.10915T>C	m.3107del	m.523A>del(m.514C>del)	m.750A>G	m.3107del	m.663A>G	m.309C>ins1C	m.489T>C	m.309C>ins1C
	m.921T>C	m.750A>G	m.11719G>A	m.4769A>G	m.524C>del(m.515A>del)	m.1438A>G	m.4580G>A	m.750A>G	m.315ins1C	m.523A>del(m.514C>del)	m.315ins1C
	m.1438A>G	m.769G>A	m.11914G>A	m.6671T>C	m.750A>G	m.2706A>G	m.4769A>G	m.771A>G (771R)	m.489T>C	m.524C>del(m.515A>del)	m.709G>A
	m.1719G>A	m.825T>A	m.12007G>A	m.8860A>G	m.980T>C	m.3107del	m.6227T>C	m.1438A>G	m.750A>G	m.750A>G	m.750A>G
	m.2416T>C	m.1018G>A	m.12121T>C	m.14423G>S	m.1438A>G	m.4769G>A>G	m.6773C>T	m.1442G>A	m.1438A>G	m.1438A>G	m.1438A>G
	m.2706A>G	m.1048C>T	m.12142A>G	m.15326A>G	m.1811A>G	m.5969C>T	m.7028C>T	m.1736A>G	m.2706A>G	m.1462G>A	m.2706A>G
	m.3107del	m.1243T>C	m.12696T>C	m.16519T>C	m.2706A>G	m.6077C>T	m.7759T>Y	m.2706A>G	m.3107del	m.2706A>G	m.3107del
	m.3498C>G	m.2758G>A	m.12705C>T		m.3107del	m.7028C>T	m.8440>G	m.3107del	m.3296T>Y	m.3107del	m.3290T>C
	m.4688T>C	m.2885T>C	m.12720A>G		m.3741C>T	m.8860A>G	m.8860A>G	m.4248T>C(m.4248T>Y)	m.4216T>C	m.3882G>A	m.3970C>T
	m.4769A>G	m.3107del	m.12798C>T		m.4502T>C	m.9033A>G	m.15326A>G	m.4769A>G	m.4769A>G	m.4071C>T	m.4769A>G
	m.5147G>A	m.3438G>A	m.13105A>G		m.4769A>G	m.9055G>A	m.15904C>T	m.4824A>G	m.5633C>T	m.4769A>G	m.5263C>T
	m.7028C>T	m.3516C>A	m.13276A>G		m.4788G>R	m.9554G>A	m.16183A>del(m.16183A>M)	m.4928T>C	m.7028C>T	m.4850C>T	m.6392T>C
	m.7424A>G	m.3594C>T	m.13506C>T		m.5256A>G	m.10118T>C	m.16189T>C	m.7028C>T	m.7476C>T	m.5442T>C	m.7028C>T
	m.8155G>R	m.3756A>G	m.13650C>T		m.5360C>T	m.11711G>R	m.16193C>ins1C	m.8794C>T	m.8860A>G	m.6053C>T	m.7861T>C
	m.8251G>A	m.4104A>G	m.13759G>A		m.7028C>T	m.11719G>A	m.16298T>C	m.8860A>G	m.10172G>A	m.6455C>T	m.8860A>G
	m.8618T>C	m.4232T>C	m.14766C>T		m.8137C>T	m.12191C>T		m.9713G>A	m.10286C>G	m.7028C>T	m.10310G>A
	m.8701A>G	m.4312C>T	m.15326A>G		m.8684C>T	m.12346C>T		m.9756T>G	m.10398A>G	m.8701A>G	m.10915T>C
	m.8860A>G	m.4769A>G	m.15466G>A		m.8860A>G	m.13020T>C		m.11719G>A	m.11251A>G	m.8860A>G	m.11038A>G
	m.9540T>C	m.5442T>C	m.15930G>A		m.10142C>T	m.14766C>T		m.12705C>T	m.11719G>A	m.9540T>C	m.11719G>A
	m.10398A>G	m.6185T>C	m.15941T>C		m.11467A>G	m.15326A>G		m.14766C>T	m.12612A>G	m.9824T>C	m.12153C>T
	m.10873T>C	m.6266A>G	m.16129G>A		m.11719G>A	m.15355G>A		m.15326A>G	m.13708G>A	m.10398A>G	m.12396T>C
	m.10899A>G	m.6815T>C	m.16187C>T		m.12094C>T	m.15607A>G		m.16223C>T	m.14766C>T	m.10400C>T	m.12408T>C
	m.11404A>G	m.7028C>T	m.16189T>C		m.12308A>G	m.16176C>T		m.16249T>C	m.15257G>A	m.10873T>C	m.12630G>A
	m.11719G>A	m.7146A>G	m.16230A>G		m.12372G>A	m.16189T>Y		m.16290C>T	m.15326A>G	m.11665C>T	m.13602T>C
	m.12705C>T	m.7256C>T	m.16234T>C		m.13500T>C	m.16266C>T		m.16319G>A	m.15452C>A	m.11719G>A	m.13928G>C (m.13928G>S)
	m.13105A>G	m.7389T>C	m.16243T>C		m.14533C>T	m.16270C>T		m.16362T>C	m.15812G>A	m.12091T>C	m.14766C>T
	m.13886T>C	m.7521G>A	m.16266C>A		m.14569G>A	m.16319G>A			m.16069C>T	m.12705C>T	m.14831G>R
	m.14284C>T	m.8113C>A	m.16311T>C		m.14766C>T	m.16357T>C			m.16126T>C	m.12804T>C	m.15326A>G (15326R)
	m.14766C>T	m.8152G>A	m.16519T>C		m.14921G>R				m.16193C>T	m.14755A>T	m.15670T>C
	m.15061A>G	m.8251G>A			m.14927A>G				m.16278C>T	m.14766C>T	m.16207A>G
	m.15208A>G	m.8468C>T			m.15326A>G					m.14783T>C	m.1630T>C
	m.15301G>A	m.8508A>G			m.16129G>A					m.15043G>A	m.16362T>C
	m.15326A>G	m.8566A>G			m.16318A>T					m.15301G>A	m.16399A>G
	m.16124T>C	m.8655C>T			m.16519T>C					m.15326A>G	m.16497A>G
	m.16183A>del(m.16183A>M)	m.8701A>G								m.16038A>G	
	m.16189T>C	m.8860A>G								m.16223C>T	
	m.16193C>ins1C	m.9042C>T								m.16294C>T	

Sample ID	11	12	13	14	15	16	17	18	19	20
Haplotypes	m.73A>G	m.73A>G	m.73A>G	m.73A>G	m.73A>G	m.73A>G	m.73A>G	m.73A>G	m.44C>ins1C	m.263A>G
	m.150A>G	m.150C>T	m.152T>C	m.263A>G	m.263A>G	m.195T>C	m.210A>G	m.189A>G	m.72T>C	m.309C>ins1C
	m.195T>C	m.152T>C	m.150A>G	m.315ins1C	m.315ins1C	m.263A>G	m.263A>G	m.194C>T	m.263A>G	m.315ins1C
	m.225G>A	m.263A>G	m.215A>G	m.489T>C	m.709G>A	m.309C>ins1C	m.309C>ins1C	m.195T>C	m.309C>ins1C	m.750A>G
	m.226T>C	m.295C>T	m.263A>G	m.750A>G	m.750A>G	m.315ins1C	m.315ins1C	m.204T>C	m.315ins1C	m.1438A>G
	m.263A>G	m.315ins1C	m.309C>ins1C	m.1438A>G	m.930G>A	m.497C>T	m.523A>del(m.514C>del)	m.207G>A	m.750A>G	m.3107del
	m.315ins1C	m.489T>C	m.315ins1C	m.1872T>C	m.1438A>G	m.524C>ins4AC(m.513A>ins3AC)	m.524C>del(m.515A>del)	m.263A>G	m.1438A>G	m.4769A>G
	m.750A>G	m.750A>G	m.523A>del(m.514C>del)	m.2706A>G	m.2706A>G	m.750A>G	m.709G>A	m.309C>ins1C	m.2706A>G	m.8860A>G
	m.1438A>G	m.1438A>G	m.524C>del(m.515A>del)	m.3107del	m.3107del	m.1189T>C	m.750A>G	m.315ins1C	m.3107del	m.10211C>T
	m.1719G>A	m.2706A>G	m.750A>G	m.3565A>G	m.4216T>C	m.1438A>G	m.1438A>G	m.709G>A	m.4580G>A	m.12033A>G
	m.2706A>G	m.3107del	m.1438A>G	m.4769A>G	m.4769A>G	m.1811A>G	m.2706A>G	m.750A>G	m.4769A>G	m.12732T>C
	m.3107del	m.4216T>C	m.2706A>G	m.7028C>T	m.4823T>C	m.2706A>G	m.3107del	m.1243T>C	m.7028C>T	m.13889G>R
	m.4769G>A>G	m.4769G>A>G	m.3107del	m.7931C>T	m.4917A>G	m.3107del	m.3537A>G	m.1438A>G	m.8568C>A	m.15326A>G
	m.6221T>C	m.5633C>T	m.3834G>C	m.8701A>G	m.5147G>A	m.3480A>G	m.4769G>A>G	m.2706A>G	m.8860A>G	m.16519T>C
	m.6371C>T	m.7028C>T	m.4385A>G	m.8848T>C	m.6935C>T	m.4769A>G	m.6960C>T	m.3107del	m.12810A>G	
	m.7028C>T	m.7476C>T	m.4769A>G	m.8860A>G	m.7028C>T	m.5460G>A	m.7028C>T	m.3505A>G	m.15326A>G	
	m.8393C>T	m.8860A>G	m.6041C>T	m.9302C>T	m.7310T>C	m.7028C>T	m.8281-8289del	m.4363T>C	m.15904C>T	
	m.8860A>G	m.10172G>A	m.7028C>T	m.9438G>A	m.8697G>A	m.7245A>G	m.8584G>A	m.4769G>A>G	m.16216A>G	
	m.10532A>G	m.10398A>G	m.7657T>C	m.9540T>C	m.8860A>G	m.8860A>G	m.8860A>G	m.5046G>A	m.16261C>T	
	m.11719G>A	m.11251A>G	m.8860A>G	m.10398A>G	m.9007A>G	m.9055G>A	m.9950T>C	m.5460G>A	m.16298T>C	
	m.12705C>T	m.11719G>A	m.9509T>C	m.10400C>T	m.10463T>C	m.9698T>C	m.10398A>G	m.6528C>T	m.16362T>C	
	m.13251C>T	m.12612A>G	m.11719G>A	m.10873T>C	m.10580A>G	m.10398A>G	m.11719G>A	m.7028C>T	m.16519T>C	
	m.13708G>A	m.13708G>A	m.12172A>G	m.11719G>A	m.11251A>G	m.10550A>G	m.13145G>A	m.8251G>A		
	m.13966A>G	m.14569G>A	m.12831C>T	m.12705C>T	m.11719G>A	m.11299T>C	m.13395A>G	m.8860A>G		
	m.14470T>C	m.14766C>T	m.13368G>A	m.12753A>G	m.11812A>G	m.11467A>G	m.14766C>T	m.8994G>A		
	m.14766C>T	m.15257G>A	m.14094T>C	m.14233A>G	m.12070G>A	m.11719G>A	m.15235A>G	m.10097A>G		
	m.15326A>G	m.15326A>G	m.14766C>T	m.14766C>T	m.13368G>A	m.12308A>G	m.15326A>G	m.10410T>C		
	m.15927G>A	m.15452C>A	m.15326A>G	m.14783T>C	m.14233A>G	m.12372G>A	m.15565T>C	m.11674C>T		
	m.16183A>del(m.16183A>M)	m.15812G>A	m.15607A>G	m.15043G>A	m.14766C>T	m.14167C>T	m.16140T>C	m.11719G>A		
	m.16189T>C	m.16069C>T	m.16193C>T	m.15301G>A	m.14905G>A	m.14766C>T	m.16183A>C (m.16183A>M)	m.11947A>G		
	m.16193C>ins1C (16192Y)	m.16126T>C	m.16357T>C	m.15326A>G	m.15326A>G	m.14798T>C	m.16189T>C	m.12414T>C		
	m.16223C>T	m.16193C>T	m.16519T>C	m.15644A>G	m.15452C>A	m.15115T>C	m.16266C>A	m.12705C>T		
	m.16278C>T	m.16301C>T		m.15820C>T	m.15607A>G	m.15326A>G	m.16519T>C	m.14470T>C		
	m.16519T>C			m.15838C>T	m.15928G>A	m.16093C>T		m.14766C>T		
				m.15939C>T	m.16126T>C	m.16163A>G		m.15326A>G		

**Human mitochondrial genome, rCRS (GenBank NC_012920.1)*

Supplementary Table 2: The mtDNA variants detected for all the 20 samples analysed.

				NGS Data from NextGENe® software					
Sample ID	rCRS	Sanger	NGS	%A	%C	%G	%T	%Insertions	%Deletions
1	C	m.16183A>del	m.16183A>M	62.62	35.05	0.23	0.12	0.93	1.99
	-	m.16193C>ins1C	-	-					
2	A	m.357A	m.357A>M	86.75	11.99	0	0	0	1.26
	C	m.1048C>T	m.1048C>Y	0	12.66	0	87.34	0.1	0
	G	m.5824G	m.5824G>del	0.5	0	62.41	0	0.03	37.09
4	G	m.5824G	m.5824G>del	0.56	0	61.17	0.03	0	38.24
5	T	m.13020T>C	m.13020T>Y	0	80.18	0	19.7	0.16	0.12
	A	m.15326A>G	m.15326A>R	15.08	0	84.92	0	0.1	0
6	C	m.16183A>del	m.16183A>M	58.46	36.41	0	0	0	5.13
	-	m.16193C>ins1C	-	-					
7	A	m.357A	m.357A>M	75.78	22.36	0	0	0.62	1.86
	A	m.771A>G	m.771A>R	16.15	0	83.1	0	0	0.75
	T	m.4248T>C	m.4248T>Y	0	85.67	0.11	13.9	7.22	0.32
8	-	m.309C>ins1C	m.309C>ins2C	0	98.64	0	0	82.34	1.36
	C	m.13128C	m.13128C>M	13.99	83.33	0	0	1.53	2.67
9	A	m.540A	m.539T>ins1C	99.85	0	0	0	36.98	0.15
	T	m.15326A>G	m.15326A>R	26.92	0.07	72.93	0	0.15	0.07
	T	m.16209T	m.16209T>Y	0	25	0	73.33	1.67	1.67
10	T	m.204T	m.204T>Y	0.06	12.85	0	87.09	0.95	0
	A	m.540A	m.539T>ins1C	99.73	0	0.13	0	47	0.13
	G	m.13928G>C	m.13928G>S	0	45.41	54.59	0	0	0
	A	m.15326A>G	m.15326A>R	44.77	0	55.23	0	0.11	0
11	C	m.16183A>del	m.16183A>M	82.69	15.07	0	0.2	0	2.04
	-	m.16193C>ins1C	-	-					
12	C	m.13128C	m.13128C>M	14.86	83.17	0	0	0.19	1.97
13	-	m.309C>ins1C	m.309C>ins1C	0.19	99.25	0	0	87.45	0.56
14	A	m.357A	m.357A>M	53.99	39.63	0.07	0	0.42	6.32
	A	m.540A	m.539T>ins1C	99.92	0	0.08	0	83.15	0
15	A	m.357A	m.357A>M	70.3	24.04	0.1	0	0	5.56
16	-	m.309C>ins1C	m.309C>ins2C	0.19	99.81	0	0	85.91	0
17	C	m.13128C	m.13128C>M	33.31	64.87	0	0.1	0.35	1.72
19	-	m.309C>ins1C	m.309C>ins2C	0	100	0	0	90.99	0
20	-	m.309C>ins1C	m.309C>ins2C	0	99.57	0	0	89.46	0.43

**Human mitochondrial genome, rCRS (GenBank NC_012920.1)*

Supplementary Table 3: Differences between the previously developed Long-Range amplification MPS approach and the newly introduced MPS tiling approach for whole mt genome analysis both obtained via the PGM in sample 2.

Sample ID	Total Reads	Aligned Reads	Percent of Aligned Reads	Maximum Coverage	Average Coverage	Percent of Coverage
1_10ng	641486	535003	83.40	17847	3593	100%
1_1ng	471340	462629	98.15%	13050	2985.10	100%
1_500pg	462206	443056	95.86%	11385	2821.92	100%
1_250pg	457948	447719	97.77%	11876	2622.88	100%
1_100pg	408501	401726	98.34%	9947	2591.70	99.93%
15_10ng	541295	532291	98.34%	14460	3487.86	100%
15_1ng	478246	459469	96.07%	16360	2991.91	100%
15_500pg	413768	409022	98.85%	11768	2660.91	100%
15_250pg	335769	289732	86.29%	8079	1241.88	100%
15_100pg	284550	282884	99.42%	5923	1175.5	99.91%
19_10ng	420730	416278	98.94%	8839	2644.43	100%
19_1ng	245109	242407	98.90%	4867	1568.94	100%
19_250pg	200738	176739	88.05%	5134	1198.71	100%
19_500pg	113875	95517	83.88%	4525	576.49	99.89%
19_100pg	93205	87189	93.55%	6796	462.08	99.74%

Supplementary Table 4: Performance summary from the sensitivity tests of the PGM tool for whole mt genome sequencing on three samples belonging to different haplogroups with genomic DNA inputs ranging from 10 ng to 100 pg.

Chapter 6

General Discussion

In the subfield of Forensic DNA Phenotyping, considerable research is ongoing invariably to determine how detailed the information on appearance and bio-geographic ancestry can be retrieved from human biological materials such as crime scene traces. Moreover, there are constant developments in genotyping methods, chemistries and platforms for SNPs used for EVC and ancestry prediction aimed to successfully deal with DNA obtained from crime scene stains, which typically is of low quantity and low quality providing challenges on genotyping methods and technologies. When the DNA profile in question fails to match any of the suspects' DNA profiles, or the DNA profiles in the criminal offender DNA database, most of the cases remain unsolved. The burgeoning field of Forensic DNA Phenotyping provides invaluable information from DNA samples that can reduce the potential number of suspects helping to find unknown perpetrators unable to be identified with standard DNA profiling.

Whenever a new method has been developed, it is important to test the performance proficiency and the limitations of the new methodology before it is introduced for routine forensic DNA analysis. To ensure that the novel methodology is highly reliable and accurate, it is imperative to validate the new method before implementing it for routine forensic casework. Validation involves various laboratory experiments to assert that the new DNA method developed is highly robust and reliable. It is necessary to assure that the new technology and methods can easily be reproduced and applied by other forensic laboratories worldwide, even by those with minimal experience. The Scientific Working Group on DNA Analysis Methods (SWGDAM) has set forth guidelines for the developmental validation of the newly developed DNA methods to ensure their quality and reliability. In this thesis, I describe forensic validation studies on previously developed SNP-based DNA-based intelligence approaches for predicting eye and hair color as well as maternal ancestry as a prerequisite for future forensic applications. Moreover in this thesis, I demonstrate the new development and the forensic developmental validation of a new DNA phenotyping system for skin color in combination with eye and hair color and of a new MPS-based approach of sequencing the entire mitochondrial genome from degraded DNA.

Prediction of eye, hair and skin colour from DNA

Extensive research has elucidated the association of DNA variants with human pigmentation traits. The ability to predict the aspects of an unknown individual's appearance from their DNA using EVC SNPs could provide information comparable to an eye-witness's statement. Several approaches, such as, likelihood association ratio methods [1], a prediction guide process [2] and a Bayesian classifier based on likelihood ratios [3], and amongst others are available for predicting the eye colour. However, the IrisPlex system that uses model-based probability prediction approach [4-6] for prediction of blue and brown eye colour has established itself as a most sought tool in the forensic community for ascertaining the most probable eye colour using 6 DNA variants. As of May 2012, the Dutch government has allowed the use of the validated [4-6] IrisPlex eye colour prediction system in routine forensic casework, which is done at the Netherlands Forensic Institute (NFI). In **Chapter 2**, I demonstrated the ease of implementation of the IrisPlex system across forensic laboratories worldwide, with varying levels of experience and expertise of the laboratory personnel. It was deduced from this technical exercise that the IrisPlex system could be easily implemented in less-experienced laboratories, demonstrating the robustness of the assay. Despite the laboratories using different methods for extraction and quantification, it was possible to achieve a high success rate in genotyping leading to high success rates of eye colour phenotyping. For Tasks 1 and 2 in the collaborative exercise, the focus was on prediction of either blue or brown, keeping in mind the general limitations in genetically explaining the non-blue and non-brown eye colours with existing DNA markers in general and not only those included in IrisPlex [4-7]. However, in Task 3 no sampling constraints were applied as participants were free to choose their own volunteers based on any eye colour they wanted. For tasks 1 and 2 together, 99.2% (1875) of the 1890 genotypes were called correctly and of the 15 (0.8%) incorrect genotype calls, only 2 (0.1%) resulted in incorrect eye colour phenotypes. The voluntary Task 3 involved 19 laboratories and across the 100 samples, 96% of the eye colour phenotypes were correctly inferred. Through the data provided by the study collaborators, I successfully assessed the reliability, consistency and the overall performance of the IrisPlex system for predicting blue and brown eye colour, when used by participants with absolutely no SNP typing experience as well as by those with SNaPshot experience. Several publications raised concerns about the marker content and the model outcomes of the

IrisPlex system [8–10]. However, this EDNAP exercise was carried-out to demonstrate how the IrisPlex assay performed in different laboratories with different pre-existing experiences as typical for any EDNAP exercises to provide practical evidence to those colleagues who plan to implement the system into their laboratory routine. Since, for the purpose of FDP, the analysis does not stop with producing the genotype but continues in translating the genotypic outcomes into phenotypic outcomes, this second step using the IrisPlex prediction model was added as well. However, this step simply includes inputting the obtained genotypes into a prediction spreadsheet as the provided eye colour probabilities are automatically generated by using our previous established model parameters. As has been previously established based on different datasets and was already evident during initial marker ascertainment, the IrisPlex system is suitable for accurately predicting blue and brown eye colours but is unable to predict the intermediate/green eye colours to the same degree due to general limitations in genetically explaining such non-blue and non-brown eye colours with existing DNA markers and not only those included in IrisPlex [4–7]. To overcome this limitation, fundamental work in the genetics of eye colour is needed. DNA variants with similarly high prediction effect on non-blue/non-brown eye colours as the IrisPlex SNPs have on blue and brown eye colour remain to be identified.

At present the IrisPlex system is the best assay combined model available for the prediction of blue and brown eye colour on a categorical level, ultimately with further research into the continuous basis of eye colour and the addition of more SNPs that contribute significantly to the intermediate category, this will improve eye colour knowledge and its individual-based prediction overall. In an effort to address the majority of the issues or suggestions seen with regards to the 6-SNP IrisPlex system, a new online tool with an updated eye colour phenotype database of >9100 individuals from 8 countries (north and south) of Europe, and a model that caters for incomplete profiles has been generated as described in **Chapter 3.1** of this thesis. Moreover, combining the knowledge available from the publications on the IrisPlex system [4,5] and other previous publications regarding the ability to predict hair color from DNA variants [11–14], the HIrisPlex DNA test system, a single multiplex system targeting 24 DNA variants for simultaneous prediction of eye and hair colour, was established recently [15]. This model-based probability prediction system allows predicting red, blond, black and brown hair colours with varying accuracies and additionally provides information about the shade of the hair colour (light or dark). The HIrisPlex system

is the first forensic tool that simultaneously predicts both the eye and hair colours from statistical models. In **Chapter 3.1**, I demonstrated how the HIrisPlex assay surpassed the rigorous guidelines set forth by the SWGDAM, which typically is used for forensic developmental validation testing of newly developed DNA tools as prerequisite for casework applications, in terms of sensitivity, specificity, stability, mixture studies, reproducibility, precision and accuracy, casework samples and population studies. Complete 24 SNP profiles were achieved with a minimum DNA input of 63 pg, thus reaching sensitivity levels comparable to those published earlier during the tool development [15]. The HIrisPlex system displayed sensitivity comparable to that of several STR systems, such as Powerplex ESX systems from Promega [16] and higher than that obtained with the 18-plex SNP assay for identification of highly degraded human DNA [17], but lower than the AmpFlSTR MiniFiler kit [18]. The HIrisPlex assay delivered consistent results across internal and external laboratories. Mixture studies were also performed to assess the ability of the HIrisPlex assay to identify mixtures. It is to be noted that mixture determination can be very difficult with biallelic SNP markers [19]. There is a possibility to detect the presence of a mixture in samples with mixture ratios of 1:5 and 1:10 from the average peak heights and the peak height ratios. However, standard STR typing is the method of choice for mixture determination. STR typing is routinely done on samples prior to any other assay, such as IrisPlex or HIrisPlex. Furthermore, accurate results were obtained from extremely degraded samples and from simulated casework samples, such as blood, semen, saliva and hair. Together with the enhanced prediction tool for eye color based on >9100 individuals and on hair colour based on >1600 individuals, an online web-based tool was introduced, and has been made publically available for use in routine forensic casework. With this tool it is now possible to achieve the eye and hair colour probabilities not only from full profiles but also from partial profiles, however with the exception of those with missing *HERC2s* rs12913832 and all 11 *MC1Rs* DNA markers, which are vital for hair color prediction so that their missing cannot be tolerated.

The HIrisPlex system was earlier applied to DNA samples extracted from ancient and contemporary skeletal remains [20]. The feasibility of the HIrisPlex assay was exemplified by analysing the tooth collected from the corpse General Władysław Sikorski., a historical figure in Polish history who died in 1943 in an air-plane crash [20]. The blue eye and blond hair colour obtained from HIrisPlex were positively confirmed with reliable documented reports [20]. Recently, the HIrisPlex DNA markers and prediction models were

also used to predict the blue eye colour and blond hair colour information from skeletal remains that were identified to be those of King Richard III of England (1452-1485) [21]. In **Chapter 3.2**, I illustrated the suitability of the HIrisPlex system to genotype DNA extracted from skeletal remains of World War II victims, excavated from six different sites in Slovenia. This chapter focusses on testing the suitability of the HIrisPlex system as a robust tool to retrieve eye and hair colour information from DNA samples from human remains that are almost 70 years old and have been buried in the ground for most of that time. Basic anthropological research alone cannot be used to retrieve information about pigmentation; reliable tools such as HIrisPlex can be used to bring colour of the deceased remains back by accurately predicting the eye and hair colour. In the present study, complete 15-loci autosomal STR (plus amelogenin) profiles were obtained in 44 of the 49 samples (89.8%). It was possible to obtain complete HIrisPlex profiles for all 49 samples, with the exception of the *MC1R* DNA marker N29insA, which dropped out in 41 samples (83.7%). Two skeletal samples originated from two brothers as confirmed by STR genotyping when analysed together with a living sister's reference sample. The living sister of the two brothers confirmed their HIrisPlex-predicted eye and hair colors. STR profiling revealed that 5 pairs of skeletal remains belonged to the same individuals. Of these 5 pairs, the HIrisPlex genotyping reported the same eye and hair colour for 4 of the pairs. The quality and quantity of the DNA extracted from the old and degraded samples, as well as the preservation and storage conditions of the sample also influence the results obtained with the HIrisPlex system. Hence, it must be noted that it may not always be possible to achieve successful results when the HIrisPlex system is applied to very old or ancient samples. However, with this study we anticipate that the HIrisPlex system, which was designed to cope with DNA fragmentation by targeting small amplicons, may be applied for the purpose of disaster victim and missing person identification. With the present study and the previously published applications [26,27] of the HIrisPlex system, one can foretell that the HIrisPlex tool could be used to answer questions not only in forensics, but also in historical, anthropological and even evolutionary perspective, as it started to do already [20,21].

Compared to the prediction of eye colour, hair colour is a more variable and complex trait and hence there are more difficulties in its prediction. One of the apparent issues is the age dependent change in hair colour from blond during early childhood to dark blond or brown in late childhood and adulthood [22]. This is the main reason why blond and

brown hair color is less accurately predicted with the HIrisPlex system (or any other current available DNA markers) than red and black [15]. The molecular mechanism underlying the darkening of hair during adolescence and why it is observed in only some blonds is still obscure. Hormonal changes/fluctuations during adolescence could be a plausible explanation for such age dependent changes in hair colour [22]. Hence, the bio-marker that can identify such change in hair colour is yet to be identified. The HIrisPlex model predicts the childhood hair colour of these individuals, i.e., light blond, in adults that remained blond in adulthood as well as those that darkened during adulthood into brown [15]. Another issue with relatively little knowledge is the age-dependent loss of hair colour i.e. whitening or graying. Recently, it has been found that the SNP rs12203592 in the *IRF4* gene is associated with graying of hair [23]. Future research lies in identification of bio-markers for age-dependent hair color change that may be included in an extended version of HIrisPlex. These limitations of the HIrisPlex system when it comes to age-dependent hair colour changes have to be considered and kept in mind when using HIrisPlex for forensic investigations. In particular, when blond is predicted with HIrisPlex, the conclusion on the hair colour phenotype of the individual tested must include blond as well as the possibility of brown in case of age-dependent hair colour change.

With the development of eye and hair colour prediction systems such as IrisPlex and HIrisPlex, prediction of EVCs has become popular and the next logical step was to add skin colour. Several studies have published SNPs associated with skin colour [2,13,14]. Using the available knowledge on skin colour associated SNPs, an extension to the HIrisPlex system was developed which I described in **Chapter 4**. The HIrisPlex–S system (S for skin) consists of a second set of 17 SNPs targeted with one multiplex, in addition to the multiplex targeting the 24 SNPs from HIrisPlex, and three prediction models for eye, hair and skin colour, respectively, from one test outcome. Fitzpatrick scale that is the numerical classification of the human skin color was used for the phenotyping. The six categories of the Fitzpatrick scale (Type I–Type VI) were categorized into five categories for the HIrisPlex-S assay: very pale, pale, intermediate, dark and dark/black. Prediction of skin colour is more complex as its variation is between populations in different continents such as Europe, Asia, Africa, whereas, the eye and hair colour varies mainly in European populations. Studies on skin colour using GWAS can be done within continental groups but since the variation between the continental groups was less, the list of genes is also quite limited. Nevertheless,

the skin color prediction model of the HIrisPlex-S achieved accuracies of 0.75. for very pale, 0.73 for pale, 0.75 for medium, 0.84 for dark and 0.98 for dark/black expressed as AUC [Walsh et al., in preparation]. We thus envision the HIrisPlex-S system as a start but expect that future genetic research will deliver more DNA predictors of skin colour, which will be included in extended versions of the HIrisPlex-S tool allowing improved prediction accuracy of skin colour categories. A total of 36 SNPs from 15 genes were used for the skin color prediction. These 36 markers included all the 17 SNPs in the newly developed set and 19 of the 24 SNPs (excluding N29insA, rs1805005, rs1805009, Y152OCH and rs4959270) from the HIrisPlex assay.

The HIrisPlex-S assay was designed with the intention to cope with low template and degraded DNA, a standard concern when genotyping forensic casework samples. Therefore care was taken to ensure small PCR amplicon sizes of <150 bp for all of the 17 DNA variants considered. All the 17 SNPs from 7 genes: rs3114908 (*MC1R*), rs1800414 (*OCA2*), rs10756819 (*BNC2*), rs2238289 (*HERC2*), rs17128291 (*SLC24A4*), rs6497292 (*HERC2*), rs1129038 (*HERC2*), rs1667394 (*HERC2*), rs1126809 (*TYR*), rs1470608 (*OCA2*), rs1426654 (*SLC24A5*), rs6119471 (*ASIP*), rs1545397 (*OCA2*), rs6059655 (*ASIP*), rs12441727 (*HERC2*), rs3212355 (*MC1R*) and rs8051733 (*MC1R*) were designed using Primer3Plus and tested for primer-primer interactions using the AutoDimer software. Several rounds of re-designing was done for the SNPs rs1470608 and rs6059655 as they were the most difficult to optimise. Compared to the other 16 SNPs, the peak height of rs6059655 was lower and hence there are chances that rs6059655 may be one of the first to drop out in case of low quality/quantity DNA. Similar to the previously published IrisPlex and HIrisPlex assays, an Excel macro for statistical prediction of the eye, hair and skin color simultaneously was developed (Walsh et al., in preparation). The model for the skin colour prediction will soon be added to the online web-based tool available at www.erasmusmc.nl/genetic_identification/resources/Irisplex_HIrisPlex/ that currently predicts eye and hair colour from full as well as partial profiles. In addition to developing the HIrisPlex–S tool, I also validated the tool for routine forensic applications following the SWGDAM guidelines. As described in **Chapter 4**, the minimum DNA input for the HIrisPlex–S system to produce full profile was 63 pg. This is comparable to the HIrisPlex assay and several other previously published STR and SNP assays [22-26]. Similar to the IrisPlex and HIrisPlex, the mixture detection and interpretation was difficult. It was possible to deduce the probable presence of a mixture with mixture ratios

of 1:5 and 1:10. As mentioned earlier in **Chapter 3.1**, STR typing is always the method of choice for mixture interpretation prior to other assays Rigorous testing of simulated forensic casework samples such as blood, semen, saliva stains, and low quantity touch (trace) DNA samples revealed the robustness and the efficiency of the HIrisPlex-S assay..

One major concern in prediction of EVCs, in general, is the interpretation problems that may exist with the current categorical approach of all the three systems: IrisPlex, HIrisPlex and HIrisPlex-S regarding eye, hair, and skin colour [4,15,24]. For example, consider the interpretation outcome is mentioned as blue eyes, brown hair and medium skin colour. There is a likely chance that different investigators have different shades of these colours in mind and carry on their search in different directions. A probable solution is the shift to quantitative phenotyping of the three pigmentation traits and their use in genetic studies that would further improve the current available genetic knowledge of the eye, hair and skin colours. Shifting from categorical phenotyping to quantitative phenotyping would increase the level of detail in eye, hair and skin color and also may answer the questions yet to be answered. In general the issues concerning prediction of the intermediate eye color category have still not been resolved however research is ongoing, including several publications that propose epistatic interactions as a factor in eye colour prediction, in particular, the intermediate category or non-blue/non-brown eye colour [16]. As more research using quantitative assessment of eye, hair and skin colour and its prediction will uncover additional information for such predictive tools, however it shall improve individual based prediction immensely where eye, hair and skin colour can be predicted using a type of colour value or scale, as similar to the phenotypic scale suggested by Liu et al [25] or Anderson et al [26]. With this knowledge, in addition to the categorical outcome, a print-out with the probable colours could be provided and this would avoid any ambiguities in the perception of the colour categories by the investigators [15].

Another general concern is that the visible physical traits may not always be genuine or inherent, as individuals can alter their physical features [24]. There are several ways to alter the physical appearance of an individual, such as using coloured contact lenses, hair dying, hair styling, sun- and self-tanning of skin, cosmetic surgeries, amongst others [24]. However, it is to be noted that once the appearance is feigned, the perpetrators would need to maintain the fabricated look throughout, at least till the end of the investigation to remain free from suspicion. Additionally, all the documents which have information about the EVCs,

such as passports, ID cards, driver's license need to be forged to match with the feigned look, which is not very easy and sometimes impractical [24].

Prediction of matrilineal biogeographical ancestry from mtDNA

Inferring biogeographical ancestry (BGA) is another aspect of DNA intelligence. Ancestral origin inference provides exciting clues and extra dimension to police investigations, such as diverting the DNA dragnet for the police or narrowing down the potential suspects [27,28]. The most likely forensic use of mtDNA SNPs is for missing person cases or for disaster victim identifications (DVI) due to the suitability of high copy number mtDNA to survive DNA degradation processes better than genomic DNA markers. However, maternal ancestry inference from mtDNA SNPs is also useful for finding unknown perpetrators in forensic investigation, particularly when combined with paternal ancestry obtained from Y-chromosomal ancestry informative markers (AIMs) and bi-parental ancestry inferred from autosomal AIMs (both not investigated in this thesis).

Extensive worldwide mtDNA studies including recent whole mtgenome sequencing has increased the understanding of the mtDNA phylogeny, divulging a large number of different mtDNA haplogroups as well as their geographic origin. Although the utility of mtDNA for inference of BGA is clearly evident from population studies, mtDNA SNP assays for BGA prediction have been slower to develop. This may be attributed to the conviction that mtDNA polymorphisms demonstrate reduced geographical informativeness compared to those in the Y-chromosome due to the widespread practice of patrilocality, which has favoured the analysis of non-recombining Y-chromosome (NRY) SNPs as the preferred markers for BGA [29]. Nevertheless, inferences of maternal ancestry from mtDNA SNPs are also informative and may provide the sole source of information for compromised and highly degraded samples such as for DVI purposes. Several ancestry informative SNPs, including those able to differentiate close geographic regions, have been published after intensive research into worldwide distribution of mtDNA variation [30–32]. In **Chapter 5.1**, I described the validation of the hierarchically organized SNaPshot multiplex assays, that were previously developed to discern 70 major mtDNA haplogroups typical for African, American, Western Eurasian, Eastern Eurasian, Australian and Oceanian populations, targeting 62 differentiating mtDNA SNPs [33,34]. It is to be noted that the (most likely) region of origin for each haplogroup is provided, not its present-day distribution. For instance,

a sample belonging to haplogroup H is indicated to have a maternal biogeographic origin in West Eurasia. In reality, H can be found almost everywhere in America and even in Asia or North Africa. As mentioned above, the detected haplogroups indicate their region of origin. Almost any mtDNA haplogroup can currently be found in the American continent due to the high degree of admixture of its present-day inhabitants (with three major components: European, African and Native American). Nevertheless, the presence of haplogroup H in an individual living in America still points to a Western Eurasian/Northern African ancestry of that individual, at least regarding his/her maternal line.

Few modifications were applied to the multiplex design when compared to the previous publications [33,34]. The total number of multiplexes was reduced to five from six. The multiplex 5 contained only five markers that resolved the haplogroup Q into four sub-haplogroups Q1, Q2, Q2a and Q3. The markers from the multiplex 5 were allocated into multiplex 4 and 6, based on the compatibility of them markers. Hence the overall number of multiplexes was now five and this reduced the amount of sample and reagents used, and the time spent on genotyping of the assays. As the multiplexes were designed in a hierarchical fashion, the samples need not be genotyped by all the five multiplexes in case of any previous information about the questioned sample. For instance, if there is prior knowledge indicating that the sample under investigation could have Oceanian maternal ancestry, then the sample may be directly genotyped with multiplex 4 (skipping the genotyping with the Multiplexes 1–3), and if needed genotype with multiplex 6 for further resolution. Similarly, if a sample is suspected to have European origin, the sample may be genotyped with Multiplex 2 directly (skipping Multiplex 1). All the five multiplex assays included in this system met the rigorous conditions set by the SWGDAM guidelines in terms of sensitivity, precision, mixtures, reproducibility, specificity, stability, simulated case samples and population studies. The developed assays are highly sensitive and it was possible to obtain full profiles with as little as 1 pg of DNA input, comparable to previously published results [30,35], and are more sensitive than several other genotyping assays [36-38]. As mentioned earlier, detection of mixtures is rather difficult with biallelic markers such as SNPs [19]. It is highly impossible to detect mixtures if the contributors have the same geographic origin and belong to the same haplogroup. However, when the sample donors represent different haplogroups, there are chances for these multiplex genotyping assays to identify mixture ratios of 1:10. It was possible to achieve full profiles from DNA samples up to 30 min of

ultraviolet (UV) light exposure. Further, the adeptness of these assays was demonstrated by genotyping the DNA samples from 92 individuals, with diverse known biogeographic origin, from the European Collection of Cell Cultures (ECACC) Ethnic Diversity DNA panel. The haplogroups and the maternal biogeographic origin detected by the five multiplex assays were in accordance with the information provided. Through this study I demonstrated the ease and robustness of these multiplex assays in determining the matrilineal BGA, and their suitability for forensic applications.

A common issue regarding BGA inference is the interpretation of data obtained from people who are of mixed genetic ancestry, involving ancestors descending from various geographic regions [33]. Due to the lack of recombination, Y and mtDNA variation is conserved in both the paternal and maternal lineages, respectively and has strong correlations with the continental regions [33]. Because of population admixture, there are chances of detecting atypical lineages and misinterpreting an individual's overall ancestry [33]. Limited database coverage of certain populations may also hamper the accurate interpretation of the Y and mtDNA data. To acquire a more comprehensive depiction of an individual's ancestry, it is recommended to combine the information obtained from the lineage markers (Y and mtDNA) with the autosomal ancestry-informative SNPs to detect lineages, to resolve genetic admixture and to obtain an extensive estimate of the an individual's overall biogeographic ancestry information [33]. With the use MPS, one can envision the possibility to generate large amounts of information that can significantly increase the marker depth and also aid in the interpretation of admixtures.

There is a need for construction of high quality mtDNA databases to assess the rarity of mtDNA haplotype frequencies. Besides maternal ancestry, maternal lineage identification is the major use of mtDNA for which reference databases are needed to statistically interpret matches. It has been reported that certain haplogroups occur more frequently in some populations, and are absent in others [40,41]. In order to obtain good estimations, it is very important to obtain mtDNA data from different populations. In **Chapter 5.2**, I provided a detailed description of assessing the mitochondrial control region variation for maternal lineage identification in the population of the Netherlands. I generated mtDNA population data from 680 individuals sampled at 54 sites across the Netherlands, which served as a start of a reference mtDNA database for applications in forensic and missing person casework in the Netherlands. The generated mtDNA population data are incorporated in to the EMPOP

database. The efforts by EMPOP to collect as much possible mtDNA population from different population groups to create larger databases and measure the haplotype variation helps to achieve reliable interpretation of mtDNA variation. The inclusion of population data from Netherlands into the EMPOP is an attempt towards reliable interpretation of mtDNA variation in population of Netherlands. Similar attempts are made by the Y-STR Haplotype Reference Database (YHRD) for reliable interpretation of Y data variation. Whenever an mtDNA profile of a sample obtained from the crime-scene matches with a reference sample, they might be originating from the same individual. However, there are chances that the match could also be random. Hence it is important to calculate the random match probability, which is an estimation of the rarity of a certain profile in a population that is based on the frequency of the profile in a database. The frequency estimate of a profile is influenced by the size and composition of the database, together with the geographic representation and distribution of the samples [42,43]. The random-match probability was calculated to be 0.37% and the haplotype diversity was estimated to be very high at 99.63%. A total of 509 different haplotypes were identified and using Build 16 of PhyloTree [44] the samples were assigned to 57 different haplogroups/paragroups. The most common haplogroup was HV*. The mtDNA based F_{ST} values were compared among various world populations (selected from the literature to have overlapping sequence region with the mtDNA data presented in this chapter, i.e. entire control-region sequences). It can be seen, for example, that FST values for the Dutch populations range from 0.0021 (with Austria) to 0.51454 (with Angola). When compared to autosomal genome data, theoretically one could expect mtDNA-based F_{ST} values to be somewhat larger than autosomally-based F_{ST} values because of the lower Ne of mtDNA compared to autosomes. Lao et al., 2013 [45] have briefly described the mean FST value across all genome-wide autosomal SNPs used was only 0.003 (after setting negative values to zero) and the mean combined F_{ST} value between pairs of subpopulations was even smaller at 0.00038. From these results, it is seen that there is very small overall genetic differentiation among the 54 Dutch subpopulations sampled across the entire country. The genetic differentiation between geographic subpopulations from within the Netherlands was observed to be smaller than between geographic subpopulations from within other northern European countries studied thus far in a systematic fashion, such as Sweden [45]. A subtle population substructure was reported within the Netherlands by Lao et al. [45] and the Genome of the Netherlands (GoNL) consortium [46]. However, no population substructure

was detected in the Netherlands with the dataset used in this study. Hence the effects of population substructure need not be taken into account when interpreting forensic mtDNA evidence in the Netherlands as long as it comes from Dutch European sample donors.. Given the number of individuals and their geographic distribution across the country, this study provides a good start for constructing a reference database for the population of Netherlands, and further expansion of the data content by including more individuals of Dutch European ancestry as well as from the major immigration groups (left out from the current study) will make the database more representatives of the Netherlands

In **Chapters 5.1** and **5.2**, I discussed the biogeographical ancestry prediction by genotyping the mtDNA coding region SNPs, and the maternal lineage identification via sequencing of the entire mtDNA control region, respectively. Sanger sequencing of the noncoding control region of the mt genome has been the method of choice to delineate haplotypes that enable maternal lineage identification [47-51]. However, due to the high level of homoplasmy in the control region that can obscure phylogenetic signatures, Sanger sequencing of the control region may not always infer the maternal haplogroups accurately. Limiting the analysis of mtDNA to control region alone, in general, restricts the overall power of mtDNA testing [52]. Moreover, many haplogroup-defining variants can be found outside the control region as well as mentioned in **Chapter 5.1**. Maternal haplogroup information is greatly improved by genotyping of mtDNA coding-region SNPs [32-35]. However, due to the limitations in multiplexing capacity due to technical problems, it is not possible to genotype more than 40 SNPs in a single multiplex assay with previously used genotyping technology that allows coping with the low quality and low quantity of evidence DNA. Aiming to overcome limitations in the application of mtDNA caused by the relatively small numbers of SNPs previously targeted due to technological constraints, in **Chapter 5.3**, I described a short amplicon multiplex approach with 161 primer pairs (average 200 bp) to achieve sequencing of the entire mitochondrial genome through massively parallel sequencing on the Ion Torrent Personal Genome Machine. The introduction of MPS has made the sequencing of complete mtgenome cheaper and quicker [53]. With the MPS technology it has been possible to sequence thousands of reads with a high coverage. The tilling approach described in this chapter allows the analyses of challenged samples from anthropological to forensic significance. It also provides substantially more information with more accurate haplogroup assignment and greater diversity. This approach would make sequencing of degraded forensic

samples feasible and cost-effective as long as the DNA is not strongly degraded. Preliminary sensitivity testing revealed that it was possible to achieve detailed haplogroup inference with 100pg of genomic DNA input. The AmpliSeq-based MPS approach we developed here for complete human mt genome analysis is designed for and suited to non-degraded and mildly to considerably degraded DNA leading to DNA fragmentation down to a size range of around 200 bp and larger, as often confronted with in forensic, medical, and anthropological studies. Our approach is not suitable for more severely degraded DNA as often confronted with in ancient DNA studies. For such strongly fragmented DNA, hybridization capture methods are more suitable than AmpliSeq-based approaches and have been developed in the field of ancient DNA research [54–57]. One potential disadvantage of using a tiling approach based on small amplicons over a large fragment amplification approach is the potential detection of nuclear copies of mt sequences (NUMTs) with the tiling approach that likely are not detected with the long-range approach simply because NUMTs typically are much smaller. From the samples that were analysed in this study, there was no evidence that the tiling MPS approach can pick up NUMTs, which needs to be confirmed with analysing more samples. The most established currently available MPS protocols for whole mtDNA genome sequencing are based on PCR amplification of large mtDNA fragments, typically several kilobases in size [52,58], which fail when applied to degraded DNA as encountered in many mtDNA applications. A recent publication [59] described a midi-sized amplicon approach for whole mtDNA MPS analysis using 62 PCR amplicons of 300–500 bp (average about 380 bp) in two multiplex assays, which proved useful for human hair analysis. However, many DNA samples encountered in mtDNA testing, especially for forensic and anthropological purposes are more severely degraded resulting in smaller-sized fragments. The NGS-based tool described in this chapter is specifically designed for simultaneous whole human mtDNA genome analysis in DNA samples, where degradation has produced DNA fragments, smaller than those that can be successfully analysed with the previously proposed midi-sized amplicon approach [59], but large enough for PCR amplification without a hybridization capture.

MtDNA has advantages as a forensic typing locus owing to its characteristics: high copy number within the cell, presence of highly mutating hypervariable regions, sequences variations and small size. Nevertheless, mtDNA has some disadvantages as well; being maternally inherited, any maternally related individual is expected to share the same mtDNA

sequence. Individuals within the same maternal lineage are therefore indistinguishable by mtDNA analysis. Hence, mtDNA cannot be considered as a unique identifier [60]. Further, two unrelated/ individuals might share an unknown, common maternal relative in their distant past, and in such a case a false positive match might be suggested. In such a scenario, multi-locus nuclear DNA analysis would have a higher discriminatory power between the two. Lack of recombination and the inheritance of the mtDNA genome as a single unit also limit the discriminatory power of the mtDNA by the size of the database. Another issue with mtDNA analysis is the resolution of mixtures and the occurrence of heteroplasmy. The interpretation of DNA profiles often becomes complicated when forensic casework samples contain DNA from multiple contributors. With careful interpretation and statistical evaluation, it has become possible to determine the major and minor component of an allele in conventional STR typing [61]. In mtDNA typing, it is not possible to reliably resolve mixtures using Sanger sequencing technology, therefore, the results of mtDNA mixtures are usually reported as inconclusive [39]. There is a high likelihood of overestimating the number of contributors due to occurrence of heteroplasmy (see below), or underestimating the number of contributors due to them sharing the same mtDNA haplogroup. The SWGDAM guidelines include instructions on the interpretation of mtDNA mixtures. As mentioned in Chapter 5.1, there is a possibility of detecting a mixture only if the mtDNA haplogroups of the different contributors are different. Deconvolution of mixtures is now possible with the clonal nature of the MPS technology. MPS analysis of the whole mitochondrial genome (mtgenome) is highly robust in determining the mixed DNA sequences and estimating the number of contributors to a mixed sample. However, the major challenge is in the software analyses of the whole mt genome sequences from the mixed samples [62].

MtDNA heteroplasmy is defined as the co-existence of different mtDNA types within the same individual [63]. High levels of heteroplasmy have been observed in the non-coding regions, while no variability in heteroplasmy was detected in the coding regions [64]. Levels of heteroplasmy seem to differ between individuals [65,66]. Heteroplasmy has also been reported to differ between tissues within individuals, with the highest frequencies seen in muscle tissue, and also to increase with age [67]. The presence of heteroplasmy can either complicate the data interpretation, or improve the match probability in DNA identification. Interpretation guidelines have been carefully drafted to provide recommendations on dealing with heteroplasmy based on the knowledge derived from Sanger sequencing [68,69].

The recent MPS technologies can further refine our knowledge on heteroplasmy [70], as the use of MPS strongly improves the degree of detection of authentic heteroplasmy because of its quantitative approach. As described in **Chapter 5.3**, it was possible to detect 22 point heteroplasmies with our MPS approach in a given sample set, of which only 7 were detected earlier using the Sanger sequencing approach [71]. With the MPS approach the individual mtDNA variants can be resolved and sequenced, with no cloning or additional technical steps needed. In the near future, this technique will allow for valid reporting of mtDNA heteroplasmy on a routine basis in crime laboratories across the countries.

Besides the PGM, there are other sequencing platforms like the Illumina's MiSeq, HiSeq and Pacific Biosciences sequencers in the market. In a comparison study conducted by Quail et al., [72], Ilumina's MiSeq had the lowest sequencing error rates compared to Ion Torrent's PGM and Pacific Biosciences Sequencer. Ion Torrent PGM's SNP calling was higher than that of Illumina. The PGM and MiSeq are closely matched in terms of utility and ease of workflow. The limited yield and high cost per base currently prohibit large scale sequencing on the Pacific Biosciences instrument [72]. The choice of using a specific equipment will depend on the availability of the resources, finances and existing infrastructure [72]. I have used the PGM for the work in this thesis as our laboratory had access to the PGM.

Conclusion and Outlook

Before any new method can be used for forensic casework it needs to be validated to demonstrate that it can cope with the particular needs of forensic analysis, which therefore also applied to previously and newly developed methods for Forensic DNA Phenotyping, biogeographic ancestry and maternal lineage inference from mtDNA. Moreover forensic samples are generally degraded, of poor quality, low DNA quantity and sometimes from multiple contributors. Hence, it is highly important that any method that has been developed for use upon forensic samples needs to be functional on degraded and samples with low DNA quantity. The method needs to be highly robust, accurate and easily implementable in other laboratories. The three multiplex systems – IrisPlex, HIrisPlex and HIrisPlex-S, as discussed in **Chapters 2-4** have been successfully validated to be implemented in forensic laboratories with varying levels of personnel experience. Interpretation of the results together with the statistical value strengthens their value for application on challenging forensic samples. The current DNA prediction of the eye, hair and skin colour can be further improved by moving

towards quantitative prediction to provide more detailed colour phenotypes. I also present in **Chapter 5.1** the successful forensic validation of a mtDNA multiplex system for inferring maternal biogeographic ancestry as prerequisite for casework application. In forensic practise however, mtDNA use for maternal ancestry inference shall be combined with Y-chromosome use for paternal ancestry inference and autosomal DNA use for inferring bi-parental ancestry especially useful when analysing admixed individuals. Moreover on mtDNA, I presented in **Chapter 5.2** a study that shall serve as start of a national mtDNA database for establishing haplotype frequencies needed to estimate match probabilities using highest quality criteria as accepted by the worldwide EMPOP database to which the data were contributed. Using state-of-the-art MPS technology I demonstrated in **Chapter 5.3** that the entire mtgenome can be analysed in a single test via short amplicons that allow the successful analysis of degraded (albeit not severely degraded) DNA. However, further research is needed to determine the level of detail for prediction of appearance and biogeographic ancestry.

In addition to prediction of eye, hair, skin colour and regional biogeographic ancestry, several markers associated with body height, male pattern baldness, hair structure and facial morphology [73–79], as well as ancestry-informative markers on the level of geographic sub-regions [80] have been discovered. The ability to estimate individual-specific appearance via DNA-predicted facial morphology is an important goal towards identifying unknown persons. Understanding the genes that can determine normal facial variation is at the beginning stage. It will take a long way and wait until enough predictive DNA markers are available for pratical FDP of the face [24]. If such prediction turns out to be possible, it is unlikely to be achievable with small sets of genetic markers. Therefore, new technologies are needed for the parallel genotyping of large numbers of SNPs, which can successfully deal with small amounts of degraded DNA [73].

With the progressive expansion of MPS technologies, more practical approach would be to consolidate the enlarged list of appearance predicting markers (eye, hair and skin colour, body height, male pattern baldness, hair structure, facial morphology markers) together with fine-scale ancestry-informative markers from mtDNA, non-recombining Y and autosomal SNPs within a single MPS-based test in the future. A recent study already demonstrated the combination of >430 Y-SNPs in a single MPS test [81], which did not investigate the maximal number of SNPs combinable via targeted MPS that is expected to be larger, how much larger remains to be identified in future studies. Such a large multiplexed-panel for

DNA intelligence use would provide high-throughput analysis at a reduced amount of time and reagents and would be a boon to the forensic field. However, much more fundamental genetic work is needed to identify and validate DNA predictors of comprehensive appearance and fine-level bio-geographic ancestry and to establish the respective prediction models as well as applied research on MPS to find out how many and which of such appearance and ancestry informative SNPs can be combined in a single test. Finally, recent progress in correlating biomarkers such as sjTREC [82], mRNA [83] and DNA methylation [83] with age provides promises to add age prediction to Forensic DNA Phenotyping as third component in addition to appearance and ancestry, which needs to be explored further including possibilities to combine age informative markers together with appearance and ancestry informative DNA markers via MPS to develop a multi-component DNA intelligence tool for future forensic applications in all cases where the standard DNA profile leads to no identification of the sample donor, because the person is completely unknown to the investigators.

References

[1]. Mengel-From J, Wong TH, Morling N, Rees JL, Jackson IJ. Genetic determinants of hair and eye colours in the Scottish and Danish populations. BMC Genet. 2009 Dec 30;10:88.

[2].Spichenok O, Budimlija ZM, Mitchell AA, 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 Nov;5(5):472-8.

[3]. Ruiz Y, Phillips C, Gomez-Tato A, Alvarez-Dios J, Casares de Cal M, Cruz R, Maroñas O, Söchtig J, Fondevila M, Rodriguez-Cid MJ, Carracedo A, Lareu MV. Further development of forensic eye color predictive tests. Forensic Sci Int Genet. 2013 Jan;7(1):28-40.

[4] Walsh S, Liu F, Ballantyne KN, 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 Jun;5(3):170-80.

[5]. Walsh S, Lindenbergh A, Zuniga SB, Sijen T, de Knijff P, Kayser M, Ballantyne KN. Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence. Forensic Sci Int Genet. 2011 Nov;5(5):464-71.

[6] Walsh S, Wollstein A, Liu F, Chakravarthy U, Rahu M, Seland JH, Soubrane G, Tomazzoli L, Topouzis F, Vingerling JR, Vioque J, Fletcher AE, Ballantyne KN, Kayser M. DNA-based eye colour prediction across Europe with the IrisPlex system. Forensic Sci Int Genet. 2012 May;6(3):330-40.

[7] Liu F, van Duijn K, Vingerling JR, Hofman A, Uitterlinden AG, Janssens AC, Kayser M. Eye color and the prediction of complex phenotypes from genotypes. Curr Biol. 2009 Mar 10;19(5):R192-3.

[8] Prestes PR, Mitchell RJ, Daniel R, Ballantyne KN, van Oorschot RAH. Evaluation of the IrisPlex system in admixed individuals. Forensic Sci Int Genet. 2011 Dec; 3(1): e283-4.

[9] Dembinski GM, Picard CJ. Evaluation of the IrisPlex DNA-based eye color prediction assay in a United States population. Forensic Sci Int Genet. 2014 Mar; 9:111-17.

[10] 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 Jul;7(4):453-60.

[11] Grimes EA, Noake PJ, 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.

[12] Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, Manolescu A, Karason A, Palsson A, Thorleifsson G, Jakobsdottir M, Steinberg S, Pálsson S, Jonasson F, Sigurgeirsson B, Thorisdottir K, Ragnarsson R, Benediktsdottir KR, Aben KK, Kiemeney LA, Olafsson JH, Gulcher J, Kong A, Thorsteinsdottir U, Stefansson K. Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet. 2007 Dec;39(12):1443-52.

[13] Spichenok RK, Henderson MS, Walsh MH, Garrison NA, Kelch JT, Cohen-Barak O, Erickson DT, John Meaney F, Bruce Walsh J, Cheng KC, Ito S, Wakamatsu K, Frudakis T, Thomas M, Brilliant MH. Predicting phenotype from genotype: normal pigmentation. J Forensic Sci. 2010 Mar 1;55(2):315-22.

[14] Maroñas O, Phillips C, Söchtig J, Gomez-Tato A, Cruz R, Alvarez-Dios J, de Cal MC, Ruiz Y, Fondevila M, Carracedo Á, Lareu MV. Development of a forensic skin colour predictive test. Forensic Sci Int Genet. 2014 Nov;13:34-44.

[15] 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 Jan;7(1):98-115.

[16] Tucker VC, Hopwood AJ, Sprecher CJ, McLaren RS, Rabbach DR, Ensenberger MG, Thompson JM, Storts DR. Developmental validation of the PowerPlex(®) ESI 16 and PowerPlex(®) ESI 17 Systems: STR multiplexes for the new European standard. Forensic Sci Int Genet. 2011 Nov;5(5):436-48.

[17] Freire-Aradas A, Fondevila M, Kriegel AK, Phillips C, Gill P, Prieto L, Schneider PM, Carracedo A, Lareu MV. A new SNP assay for identification of highly degraded human DNA. Forensic Sci Int Genet. 2012 May;6(3):341-9.

[18] Mulero JJ, Chang CW, Lagacé RE, Wang DY, Bas JL, McMahon TP, Hennessy LK. Development and validation of the AmpFlSTR MiniFiler PCR Amplification Kit: a MiniSTR multiplex for the analysis of degraded and/or PCR inhibited DNA. J Forensic Sci. 2008 Jul;53(4):838-52.

[19] Sobrino B, Brión M, Carracedo A. SNPs in forensic genetics: a review on SNP typing methodologies. Forensic Sci Int. 2005 Nov 25;154(2-3):181-94. Review.

[20] Draus-Barini J, Walsh S, Pośpiech E, Kupiec T, Głąb H, Branicki W, Kayser M. Bona fide colour: DNA prediction of human eye and hair colour from ancient and contemporary skeletal remains. Investig Genet. 2013 Jan 14;4(1):3.

[21] King TE, Fortes GG, Balaresque P, Thomas MG, Balding D, Maisano Delser P, Neumann R, Parson W, Knapp M, Walsh S, Tonasso L, Holt J, Kayser M, Appleby J, Forster

P, Ekserdjian D, Hofreiter M, Schürer K. Identification of the remains of King Richard III. Nat Commun. 2014 Dec 2;5:5631.

[22] Rees JL. Genetics of hair and skin color. Annu Rev Genet. 2003;37:67-90. Review.

[23] Adhikari K, Fontanil T, Cal S, Mendoza-Revilla J, Fuentes-Guajardo M, Chacón-Duque JC, Al-Saadi F, Johansson JA, Quinto-Sanchez M, Acuña-Alonzo V, Jaramillo C, Arias W, Barquera Lozano R, Macín Pérez G, Gómez-Valdés J, Villamil-Ramírez H, Hunemeier T, Ramallo V, Silva de Cerqueira CC, Hurtado M, Villegas V, Granja V, Gallo C, Poletti G, Schuler-Faccini L, Salzano FM, Bortolini MC, Canizales-Quinteros S, Rothhammer F, Bedoya G, Gonzalez-José R, Headon D, López-Otín C, Tobin DJ, Balding D, Ruiz-Linares A. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. Nat Commun. 2016 Mar 1;7:10815.

[24] Kayser M. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes.Forensic Sci Int Genet. 2015 Sep;18:33-48.

[25] Liu F, Wollstein A, Hysi PG, Ankra-Badu GA, Spector TD, Park D, Zhu G, Larsson M, Duffy DL, Montgomery GW, Mackey DA, Walsh S, Lao O, Hofman A, Rivadeneira F, Vingerling JR, Uitterlinden AG, Martin NG, Hammond CJ, Kayser M. Digital quantification of human eye color highlights genetic association of three new loci. PLoS Genet. 2010 May 6;6(5)

[26] Andersen JD, Johansen P, Harder S, Christoffersen SR, Delgado MC, Henriksen ST, Nielsen MM, Sørensen E, Ullum H, Hansen T, Dahl AL, Paulsen RR, Børsting C, Morling N.Genetic analyses of the human eye colours using a novel objective method for eye colour classification. Forensic Sci Int Genet. 2013 Sep;7(5):508-15.

[27] Frudakis T, Venkateswarlu K, Thomas MJ, Gaskin Z, Ginjupalli S, Gunturi S, Ponnuswamy V, Natarajan S, Nachimuthu PK. A classifier for the SNP-based inference of ancestry. J Forensic Sci. 2003 Jul;48(4):771-82. Erratum in: J Forensic Sci. 2004 Sep;49(5):1145-6.

[28] Shriver MD, Mei R, Parra EJ, Sonpar V, Halder I, Tishkoff SA, Schurr TG, Zhadanov SI, Osipova LP, Brutsaert TD, Friedlaender J, Jorde LB, Watkins WS, Bamshad MJ, Gutierrez G, Loi H, Matsuzaki H, Kittles RA, Argyropoulos G, Fernandez JR, Akey JM, Jones KW.Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. Hum Genomics. 2005 Jun;2(2):81-9.

[29] Seielstad MT, Minch E, Cavalli-Sforza LL. Genetic evidence for a higher female migration rate in humans.Nat Genet. 1998 Nov;20(3):278-80.

30. Brandstätter A, Parsons TJ, Parson W. Rapid screening of mtDNA coding region SNPs for the identification of west European Caucasian haplogroups. Int J Legal Med. 2003 Oct;117(5):291-8.

[31] Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ.Harvesting the fruit of the human mtDNA tree. Trends Genet. 2006 Jun;22(6):339-45. Epub 2006 May 4. Review.

[32] Schlebusch CM, Naidoo T, Soodyall H. SNaPshot minisequencing to resolve mitochondrial macro-haplogroups found in Africa. Electrophoresis. 2009 Nov;30(21):3657-64.

[33] van Oven M, Vermeulen M, Kayser M. Multiplex genotyping system for efficient inference of matrilineal genetic ancestry with continental resolution. Investig Genet. 2011 Mar 23;2:6.

[34] Ballantyne KN, van Oven M, Ralf A, Stoneking M, Mitchell RJ, van Oorschot RA, Kayser M. MtDNA SNP multiplexes for efficient inference of matrilineal genetic ancestry within Oceania. Forensic Sci Int Genet. 2012 Jul;6(4):425-36.

[35] Vallone PM, Just RS, Coble MD, Butler JM, Parsons TJ.A multiplex allele-specific primer extension assay for forensically informative SNPs distributed throughout the mitochondrial genome. Int J Legal Med. 2004 Jun;118(3):147-57.

[36] Sanchez JJ, Phillips C, Børsting C, Balogh K, Bogus M, Fondevila M, Harrison CD, Musgrave-Brown E, Salas A, Syndercombe-Court D, Schneider PM, Carracedo A, Morling N. A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis. 2006 May;27(9):1713-24.

[37] Quintáns B, Alvarez-Iglesias V, Salas A, Phillips C, Lareu MV, Carracedo A. Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing. Forensic Sci Int. 2004 Mar 10;140(2-3):251-7.

[38] Green RL, Lagacé RE, Oldroyd NJ, Hennessy LK, Mulero JJ. Developmental validation of the AmpFℓSTR® NGM SElect™ PCR Amplification Kit: A next-generation STR multiplex with the SE33 locus. Forensic Sci Int Genet. 2013 Jan;7(1):41-51.

[39] Melton T, Holland C, Holland M. Forensic Mitochondrial DNA Analysis: Current Practice and Future Potential.Forensic Sci Rev. 2012 Jul;24(2):101-22. Review.

[40] Torroni A, Schurr TG, Cabell MF, et al. Asian affinities and continental radiation of the four founding Native American mtDNAs. American Journal of Human Genetics. 1993;53(3):563-590.

[41] Quintana-Murci L, Semino O, Bandelt HJ, Passarino G, McElreavey K, Santachiara-Benerecetti AS. Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nat Genet. 1999 Dec;23(4):437-41.

[42] Budowle B, Wilson MR, DiZinno JA, Stauffer C, Fasano MA, Holland MM, Monson KL. Mitochondrial DNA regions HVI and HVII population data. Forensic Sci Int. 1999 Jul 12;103(1):23-35

[43] Röhl, A., Brinkmann, B., Forster, L. et al. An annotated mtDNA database. Int J Leg Med 2001 Aug 115(29).

[44] van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat. 2009 Feb;30(2):E386-94.

[45] O. Lao, E. Altena, C. Becker, S. Brauer, T. Kraaijenbrink, M. van Oven, P. Nürnberg, P. de Knijff, M. Kayser, Clinal distribution of human genomic diversity across the Netherlands despite archaeological evidence for genetic discontinuities in Dutch population history, Invest. Genet. 4 (1) (2013) 9.

[46] Genome of the Netherlands Consortium, Whole-genome sequence variation, population structure and demographic history of the Dutch population, Nat. Genet. 46 (8) (2014) 818–825.

[47] Gill P, Ivanov PL, Kimpton C, Piercy R, Benson N, Tully G, Evett I, Hagelberg E, Sullivan K. Identification of the remains of the Romanov family by DNA analysis. Nat Genet. 1994 Feb;6(2):130-5.

[48] Wilson MR, DiZinno JA, Polanskey D, Replogle J, Budowle B. Validation of mitochondrial DNA sequencing for forensic casework analysis. Int J Legal Med. 1995;108(2):68-74.

[49] Holland MM, Parsons TJ. Mitochondrial DNA Sequence Analysis - Validation and Use for Forensic Casework.Forensic Sci Rev. 1999 Jun;11(1):21-50. Review.

[50] Brandstätter A, Niederstätter H, Pavlic M, Grubwieser P, Parson W. Generating population data for the EMPOP database - an overview of the mtDNA sequencing and data evaluation processes considering 273 Austrian control region sequences as example. Forensic Sci Int. 2007 Mar 2;166(2-3):164-75.

[51] Palo JU, Hedman M, Söderholm N, Sajantila A.Repatriation and identification of the Finnish World War II soldiers. Croat Med J. 2007 Aug;48(4):528-35.

[52] Parson W, Huber G, Moreno L, Madel MB, Brandhagen MD, Nagl S, Xavier C, Eduardoff M, Callaghan TC, Irwin JA. Massively parallel sequencing of complete mitochondrial genomes from hair shaft samples.Forensic Sci Int Genet. 2015 Mar;15:8-15.

[53] Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46.

[54] Noonan JP,Hofreiter M, SmithD, Priest JR, RohlandN, Rabeder G, Krause J, Detter JC, Paabo S, Rubin EM. 2005. Genomic sequencing of Pleistocene cave bears. Science 309:597–600.

[55] Anderung C, Persson P, Bouwman A, Elburg R, Gotherstrom A. 2008. Fishing for ancient DNA. Forensic Sci Int Genet 2:104–107.

[56] Briggs AW, Good JM, Green RE, Krause J,Maricic T, Stenzel U, Lalueza-Fox C, Rudan, P, Brajkovi´c D, Kucan ˇ Z, Guˇsi´c I, Schmitz R, et al. 2009. Targeted retrieval and analysis of five neandertal mtDNA genomes. Science 325:318–321.

[57] Templeton JE, Brotherton PM, Llamas B, Soubrier J,Haak W,Cooper A,Austin JJ. 2013. DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification. Investig Genet 4:26.

[58] Sosa MX, Sivakumar IK, Maragh S, Veeramachaneni V, Hariharan R, Parulekar M, Fredrikson KM, Harkins TT, Lin J, Feldman AB, Tata P, Ehret GB, Chakravarti A. Next-generation sequencing of human mitochondrial reference genomes uncovers high heteroplasmy frequency. PLoS Comput Biol. 2012;8(10):e1002737.

[59] Parson W, Strobl C, Huber G, Zimmermann B, Gomes SM, Souto L, Fendt L, Delport R, Langit R, Wootton S, Lagacé R, Irwin J. Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). Forensic Sci Int Genet. 2013 Sep;7(5):543-9.

[60] Melton T. Mitochondrial DNA Heteroplasmy. Forensic Sci Rev. 2004 Jan;16(1):1-20. Review.

[61] Ladd C, Lee HC, Yang N, Bieber FR. Interpretation of complex forensic DNA mixtures. Croat Med J. 2001 Jun;42(3):244-6. Review.

[62] Kim H, Erlich HA, Calloway CD. Analysis of mixtures using next generation sequencing of mitochondrial DNA hypervariable regions. Croat Med J. 2015 Jun;56(3):208-17.

[63] Salas A, Lareu MV, Carracedo A. Heteroplasmy in mtDNA and the weight of evidence in forensic mtDNA analysis: a case report.Int J Legal Med. 2001;114(3):186-90.

[64] Jazin EE, Cavelier L, Eriksson I, Oreland L, Gyllensten U. Human brain contains high levels of heteroplasmy in the noncoding regions of mitochondrial DNA. Proc Natl Acad Sci U S A. 1996 Oct 29;93(22):12382-7.

[65] Bendall KE, Macaulay VA, Baker JR, Sykes BC. Heteroplasmic point mutations in the human mtDNA control region. Am J Hum Genet. 1996 Dec;59(6):1276-87.

[66] Gocke CD, Benko FA, Rogan PK. Transmission of mitochondrial DNA heteroplasmy in normal pedigrees. Hum Genet. 1998 Feb;102(2):182-6.

[67]Calloway CD, Reynolds RL, Herrin GL Jr, Anderson WW. The frequency of heteroplasmy in the HVII region of mtDNA differs across tissue types and increases with age. Am J Hum Genet. 2000 Apr;66(4):1384-97.

[68] Tully G, Lareu M; European DNA Profiling (EDNAP) Group. Controversies over heteroplasmy. Electrophoresis. 2001 Jan;22(1):180-2.

[69] Parson W, Gusmão L, Hares DR, Irwin JA, Mayr WR, Morling N, Pokorak E, Prinz M, Salas A, Schneider PM, Parsons TJ; DNA Commission of the International Society for Forensic Genetics. DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing. Forensic Sci Int Genet. 2014 Nov;13:134-42.

[70] Just RS, Irwin JA, Parson W. Mitochondrial DNA heteroplasmy in the emerging field of massively parallel sequencing. Forensic Sci Int Genet. 2015 Sep;18:131-9.

[71] Hartmann A, Thieme M, Nanduri LK, Stempfl T, Moehle C, Kivisild T, Oefner PJ. Validation of microarray-based resequencing of 93 worldwide mitochondrial genomes. Hum Mutat. 2009 Jan;30(1):115-22.

[73] Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012 Jul 24;13:341.

[73] Kayser M, de Knijff P. Improving human forensics through advances in genetics, genomics and molecular biology. Nat Rev Genet. 2011 Mar;12(3):179-92.

[74] Phillips C. Forensic genetic analysis of bio-geographical ancestry. Forensic Sci Int Genet. 2015 Sep;18:49-65.

[75] Lango Allen H, Estrada K, Lettre G, Berndt SI et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010 Oct 14;467(7317):832-8.

[76] Liu F, Hamer MA, Heilmann S, Herold C, Moebus S, Hofman A, Uitterlinden AG, Nöthen MM, van Duijn CM, Nijsten TE, Kayser M. Prediction of male-pattern baldness from genotypes. Eur J Hum Genet. 2016 Jun;24(6):895-902.

[77] Pośpiech E, Karłowska-Pik J, Marcińska M, Abidi S, Andersen JD, van den Berge M, Carracedo Á, Eduardoff M, Freire-Aradas A, Morling N, Sijen T, Skowron M, Söchtig J, Syndercombe-Court D, Weiler N, Schneider PM, Ballard D, Børsting C, Parson W, Phillips C, Branicki W; EUROFORGEN-NoE Consortium. Evaluation of the predictive capacity of DNA variants associated with straight hair in Europeans. Forensic Sci Int Genet. 2015 Nov;19:280-8.

[78] Liu F, van der Lijn F, Schurmann C, et al. A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans. Gibson G, .PLoS Genetics. 2012;8(9):e1002932.

[79] Paternoster L, Zhurov AI, Toma AM, Kemp JP, St Pourcain B, Timpson NJ, McMahon G, McArdle W, Ring SM, Smith GD, Richmond S, Evans DM. Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. Am J Hum Genet. 2012 Mar 9;90(3):478-85.

[80] Vongpaisarnsin K, Listman JB, Malison RT, Gelernter J. Ancestry informative markers for distinguishing between Thai populations based on genome-wide association datasets. Leg Med (Tokyo). 2015 Jul;17(4):245-50.

[81] Ralf A, van Oven M, Zhong K, Kayser M. Simultaneous analysis of hundreds of Y-chromosomal SNPs for high-resolution paternal lineage classification using targeted semiconductor sequencing. Hum Mutat. 2015 Jan;36(1):151-9.Zubakov D, Liu F, van Zelm MC, Vermeulen J, Oostra BA, van Duijn CM, Driessen GJ, van Dongen JJ, Kayser M, Langerak AW.Estimating human age from T-cell DNA rearrangements. Curr Biol. 2010 Nov 23;20(22):R970-1

[82] Lech K, Liu F, Ackermann K, Revell VL, Lao O, Skene DJ, Kayser M. Evaluation of mRNA markers for estimating blood deposition time: Towards alibi testing from human forensic stains with rhythmic biomarkers. Forensic Sci Int Genet. 2016 Mar;21:119-25.

[83] C.I. Weidner, Q. Lin, C.M. Koch, L. Eisele, F. Beier, P. Ziegler, et al., Aging of blood can be tracked by DNA methylation changes at just three CpG sites, Genome Biol. 15 (2014) R24.

Summary
Nederlandse Samenvatting

Currently, genetic identification of human individuals is based on comparative grounds that involves the comparison of DNA profiles derived from short tandem repeat (STR) markers obtained from evidence and reference samples. Although the success of forensic STR profiling is conspicuous, the DNA donor remains unidentified, when an evidential STR profile does not match with the known suspect or the profiles from criminal offender database. In such circumstances, alternative DNA analysis methods, termed as 'DNA intelligence', which have been developed over the recent years, become relevant, if legally allowed. DNA intelligence involves the prediction of externally visible characteristics (EVCs) from DNA, such as the eye, hair and skin color, as well as DNA inference of biogeographical ancestry (BGA) of unknown DNA sample donors. The DNA-based prediction of EVCs and BGA can thus aid in investigations by reducing the number of possible suspects or other individuals and helping to find unknown perpetrators, unidentifiable with STR profiling. DNA intelligence is also relevant in missing person and disaster victim identification, particularly in the absence of ante-mortem samples and known relatives which – if available – are used for STR-based identification. In this thesis, I present a suite of scientific studies that introduce and/or validate genetic approaches for determining human appearance and ancestry for future forensic applications in DNA intelligence.

In **Chapter 1**, I introduced the overall topic of my thesis, starting with a brief summary about single nucleotide polymorphisms (SNPs) as they are employed in predicting EVCs and BGA. I described concisely the knowledge available about the particular SNPs involved in eye hair and skin color; and summarized the IrisPlex and HIrisPlex systems, previously developed for the prediction of eye and hair color from DNA. I continue by discussing in short the genetic knowledge available about BGA and the potentiality of the SNP markers, especially from mitochondrial DNA (mtDNA), for ancestry inference from DNA. I also summarized about the SNP genotyping techniques, such as SNaPshot, Sanger Sequencing and next generation sequencing, as those are used in the studies described in the subsequent chapters.

In **Chapter 2**, I illustrated an inter-laboratory exercise, as part of a collaborative study, to further test the reliability and consistency of the IrisPlex system, a previously developed DNA test for human eye color, across different forensic laboratories with varying levels of experience. The study, organized by the European DNA Profiling (EDNAP) Group of the International Society of Forensic Genetics (ISFG), involved 21 laboratories and three tasks.

In Task 1, the laboratories assessed 10 blood and saliva samples, and reported the genotypes with the predicted eye colors. 99.4% of the genotypes and 99% of the eye colour phenotypes were correctly reported. In Task 2, the laboratories assessed 5 artificially degraded samples; 98.7% of the genotypes 96.2% of eye colour phenotypes were correctly determined. In the voluntary Task 3, 96% of the eye colour phenotypes were inferred correctly by the 19 labs that volunteered. This study demonstrated that the IrisPlex system is easy to implement across forensic laboratories around the world with varying pre-existing experiences, and provides high reproducibility and robustness.

In **Chapter 3**, I focused on the HIrisPlex system, which is DNA-based hair colour prediction tool. In **Chapter 3.1**, the forensic developmental validation of the HIrisPlex system, a DNA test for human eye and hair color based on 24 SNPs (6 for eye color and 22 for hair color) recently developed in our department, was demonstrated by vigorous testing to meet the strict guidelines implemented by the Scientific Working Group on DNA Analysis Methods (SWGDAM). This study demonstrated that the HIrisPlex genotyping assay, which simultaneously analyses all 24 SNPs in a single reaction, met all the requirements in terms of sensitivity, specificity, reproducibility, precision and accuracy, mixture studies, casework samples and population studies. With the HIrisPlex assay, it was possible to obtain a complete 24-SNP profile from a minimum DNA input of 63 pg. Accurate genotyping results were obtained from simulated casework samples that include blood, semen, saliva, hair, and trace DNA samples in 88% of cases. It also produced consistent results when concordance testing was performed between five independent forensic laboratories with varying pre-existing experiences. Furthermore, this study delivered enhanced eye and hair color prediction models based on enlarged underlying databases of genotypes and phenotypes (eye color: N=9188 and hair color: N=1601) relative to the previously presented initial IrisPlex and HIrisPlex models based on much less data. Additionally, an online web-based system was introduced that allows the users of the HIrisPlex system to obtain eye and hair color probabilities from full and partial HIrisPlex DNA profiles in their forensic casework applications.

In **Chapter 3.2**, I presented the application of the validated (**Chapter 3.1**) HIrisPlex system to missing person/disaster victim identification in the absence of ante-mortem samples and known relatives as normally used for STR-based identification. To exemplify its suitability, we analysed the HIrisPlex system in 49 DNA samples obtained from bones or teeth of World War II victims excavated at six sites, mostly mass graves, in Slovenia. Full

15-loci autosomal STR (plus amelogenin) profiles were obtained in 44 samples(89.8%) and complete HIrisPlex profiles were obtained for all 49 samples except for the *MC1R* DNA marker N29insA, which dropped out in 41 (83.7%) samples. Eye and hair color phenotypes were obtained from the HIrisPlex genotypes, now available for use in the search for the putative relatives. Two skeletal samples originated from two brothers as confirmed by STR genotyping when analysed together with a living sister's reference sample. The living sister of the two brothers whose skeletal remains were analysed by STR-profiling, unveiling the family relationship of the three, confirmed their HIrisPlex-predicted eye and hair colors. The promising results from this study corroborate the suitability of the HIrisPlex system in providing investigative leads in missing person and disaster victim identification.

In **Chapter 4**, I demonstrated the development of the HIrisPlex-S system, an extension of the HIrisPlex system to predict skin color in addition to eye and hair color from DNA. A second multiplex with 17 SNPs targeting the skin color specific SNPs was added to the existing HIrisPlex multiplex, which also includes some skin color informative SNPs. The additionally developed statistical prediction model provided prevalence-adjusted overall prediction accuracies (expressed by AUC) ranging from 0.76 to 1 for five categories of skin colour; very light, light, medium, medium to dark, and dark skin color. Furthermore, the HIrisPlex-S assay underwent rigorous validation testing in terms of sensitivity, reproducibility, precision and accuracy, mixture studies, casework samples and population studies in preparation of its future forensic casework application.

In **Chapter 5**, I moved to the second aspect of DNA intelligence, biogeographic matrilineal ancestry inference. In **Chapter 5.1**, I described the developmental validation study of a hierarchical system of five multiplex assays targeting all the 62 ancestry-informative mitochondrial SNPs for inferring matrilineal biogeographic ancestry origin on the continental level. As shown, the assays were highly sensitive and produced full profiles with as little as 1pg of DNA input. Concordance testing across three laboratories produced reproducible and consistent results. In addition, the assays were shown to be highly robust and efficient in providing information from degraded samples and from simulated casework samples of different substrates such as blood, semen, hair, saliva and trace DNA samples. With this study, I demonstrated that the assays met the requirements of the SWGDAM (as in **Chapter 3.1** and **4**) and demonstrate their suitability for application to forensic casework.

In **Chapter 5.2** I turned to another forensic application of human mtDNA, which is in maternal lineage identification as relevant in cases where the evidence material does not contain sufficient genomic DNA for successful STR profiling but still contains enough mtDNA for mtDNA profiling. Instead of mtDNA SNPs, as used in **Chapter 5.1** for maternal ancestry inference, sequence data from the mtDNA control region are used for maternal lineage identification. However, reliable databases are crucial for population frequency estimations of mtDNA sequences that provide a match between evidence and reference material in forensic casework. Aiming the start of a Dutch reference database for forensic use, we generated complete mtDNA control-region sequences under ultra-high quality standards in a total of 680 randomly selected Dutch individuals sampled at 54 sites across the Netherlands representing ten geographic sub-regions. Haplotype diversity of the entire sample set was very high at 99.63% and, accordingly, the random-match probability was 0.37%. No population substructure within the Netherlands was detected with our dataset.

In **Chapter 5.3**, I moved from mtDNA control region analysis via classical Sanger sequencing (**Chapter 5.2**), to whole mitochondrial genome analysis via more recently established massively parallel sequencing (MPS) or next-generation sequencing (NGS). MPS provides a solution to simultaneous whole mtDNA genome analysis allowing high-throughput analysis at reduced per-sample costs. In **Chapter 5.3**, I established a MPS tiling approach for simultaneous whole human mitochondrial genome sequencing using 161 short overlapping amplicons (average 200 bp) with the Ion Torrent Personal Genome Machine. It was feasible to infer detailed haplogroups with 100 pg genomic input DNA. Further, I showed that it was possible to resolve and infer maternal genetic ancestry at complete resolution from degraded DNA (about 220bp).

Finally, in **Chapter 6**, I provided a comprehensive discussion of all the results described in **Chapters 2-5** of this thesis. I discussed the overall findings and their relevance in forensics for routine casework applications. Additionally, I addressed the challenges currently faced in predicting EVCs and BGA and their potential future perspectives.

Genetische identificatie van mensen is tegenwoordig gebaseerd op het vergelijken van DNA profielen afgeleid van 'short tandem repeat' (STR) markers welke worden verkregen uit bewijsmateriaal of referentie samples. Alhoewel het succes van forensische STR profilering opmerkelijk is, wordt de DNA donor niet geïdentificeerd als het STR profiel afkomstig uit bewijsmateriaal niet overeenkomt met een het profiel van een bekende verdachte of met een profiel uit de DNA database. In zulke gevallen zijn alternatieve DNA analsyse methodes meer van toepassing, mits bij wet toegestaan. Deze zogenaamde 'DNA intelligentie' methodes zijn de afgelopen jaren ontwikkeld, en betreffen de voorspelling van externe zichtbare eigenschappen ('externally visible characteristics' (EVCs)) vanuit DNA, zoals oog, haar en huidskleur, alsook het afleiden van de biogeografische afkomst (BGA) van onbekende DNA sample donoren. De DNA-gebaseerde voorspelling van EVCs en BGA kan daarom bijdragen aan opsporing door middel van het reduceren van het aantal mogelijke verdachten of andere individuen, en aan het vinden van onbekende daders die niet geïdentificeerd kunnen worden aan de hand van STR profielen. DNA intelligentie is ook relevant bij vermiste personen en slachtoffer-identificatie bij rampen, voornamelijk in de afwezigheid van ante-mortem samples of in de afwezigheid van bekende verwanten welke – mits beschikbaar – gebruikt worden voor STR-gebaseerde identificatie. In dit proefschrift presenteer ik een set van wetenschappelijke studies waarin de genetische aanpak geïntroduceerd dan wel gevalideerd wordt voor het bepalen van het menselijke voorkomen en de afkomst, voor toekomstige forensische toepassingen in DNA intelligentie.

In **Hoofdstuk 1** introduceer ik de hoofdlijnen van mijn proefschrift, startend met een korte samenvatting van single nucleotide polymorfismes (SNPs) omdat deze gebruikt worden bij het voorspellen van EVCs en BGA. Ik beschrijf summier de beschikbare kennis van SNPs welke betrokken zijn bij oog, haar en huidskleur, en ik geef een samenvatting van de IrisPlex en HIrisPlex systemen, welke eerder zijn ontwikkeld voor het voorspellen van oog en haarkleur vanuit DNA. Vervolgens bediscussieer ik kort de genetische kennis die beschikbaar is over BGA en de potentie van SNP markers, met name van mitochondriaal DNA (mtDNA), in het afleiden van afkomst vanuit DNA. Ik geef ook een samenvatting van SNP genotyperingstechnieken, zoals SNaPshot, Sanger Sequencing en 'next generation sequencing', omdat deze gebruikt worden in de studies zoals ze beschrijven worden in de volgende hoofdstukken.

In **Hoofdstuk 2** geef ik een inter-laboratoire oefening weer als onderdeel van een samenwerkingsstudie, waarin door meerdere forensische laboratoria met verschillende ervaringsniveaus, de betrouwbaarheid en de houdbaarheid werd onderzocht van het IrisPlex systeem, een DNA test welke eerder werd ontwikkeld voor het voorspellen van oogkleur. Deze studie werd georganiseerd door de Europese DNA Profilering (EDNAP) Groep van de Internationale Sociëteit van Forensische Genetica (ISFG), er waren 21 laboratoria bij betrokken met 3 taken. Voor taak 1 onderzochten de laboratoria 10 bloed en speeksel samples, en rapporteerde de genotypes met de voorspelde oogkleuren. 99.4% van de genotypes en 99% van de oogkleur fenotypes werden correct gerapporteerd. Voor taak 2 onderzochten de laboratoria 5 kunstmatig gedegradeerde samples; 98.7% van de genotypes en 96.2% van de oogkleur fenotypes werden correct bepaald. Taak 3 was vrijwillig, hierin werd 96% van de oogkleur fenotypes correct bepaald door de 19 laboratoria die vrijwillig deelnamen. Deze studie liet zien dat het IrisPlex systeem sterk reproduceerbaar en robuust is, en dat het makkelijk te implementeren is in forensische laboratoria over de hele wereld met variërende vooraf-bestaande ervaring.

In **Hoofdstuk 3** focus ik op het HIrisPlex systeem, dit is een DNA-gebaseerde haar en oogkleur predictie test welke recentelijk ontwikkeld is door onze afdeling en gebruik maakt van 24 SNPs (6 voor oogkleur en 22 voor haarkleur). In **Hoofdstuk 3.1** wordt de forensische validatie van het HIrisPlex systeem beschreven, waarbij deze doortastend getest is om te voldoen aan de strenge richtlijnen van het SWGDAM (Scientific Working Group on DNA Analysis Methods). Deze studie demonstreerde dat het HIrisPlex genotyperingsassay, welke simultaan 24 SNPs in een enkele reactie analyseert, voldeed aan alle eisen inzake sensitiviteit, specificiteit, reproduceerbaarheid, precisie en accuraatheid, mengsel-studies, 'casework' samples en populatiestudies. Met het HIrisPlex assay was het mogelijk om complete 24-SNP profielen te verkrijgen met een minimale hoeveelheid DNA van 63 pg. Accurate resultaten van genotypering werden verkregen uit 88% van de gesimuleerde 'casework' samples waaronder bloed, sperma, speeksel, haar en samples met zeer lage hoeveelheden DNA. Het produceerde ook consistente resultaten wanneer er getest werd op overeenkomstigheid tussen 5 onafhankelijke forensische laboratoria met variërende vooraf-bestaande ervaring. Verder versterkte deze studie de bestaande predictiemodellen voor oog en haarkleur, gebaseerd op grotere onderliggende databases van genotypes en fenotypes (oogkleur: N= 9188 en haarkleur: N=1601) in vergelijking met eerder gepresenteerde initiele IrisPlex en

HIrisplex modellen welke gebaseerd waren op kleinere onderliggende databases. Daarnaast werd een online web-gebaseerd systeem geintroduceerd dat gebruikers van het HIrisPlex systeem de mogelijkheid biedt om oog en haarkleur waarschijnlijkheids-scores te verkrijgen van volledige en partiele HIrisPlex DNA profielen toepasbaar voor hun forensische casework.

In **Hoofdstuk 3.2** presenteer ik de toepassing van het gevalideerde (**Hoofdstuk 3.1**) HIrisPlex systeem bij het identificeren van vermist personen / slachtoffers bij rampen in de afwezigheid van ante-mortem samples en bekende verwanten zoals deze normaliter gebruikt worden voor STR-gebaseerde identificatie. Om de toepasbaarheid te toe te lichten hebben we het HIrisPlex systeem geanalyseerd in 49 DNA samples verkregen uit botten en tanden van slachtoffers van de Tweede Wereldoorlog, welke zijn opgegraven op 6 verschillende locaties, voornamelijk massagraven, in Slovenië. In 44 samples (89.9%) werden volledige 15-loci autosomaal STR (plus amelogenine) profielen bereikt, en in alle 49 samples werden complete HIrisPlex profielen verkregen, met als uitzondering de *MC1R* DNA marker N29insA, deze viel af in 41 (83.7%) samples. Oog en haarkleur fenotypes werden afgeleid met behulp va de HIrisPlex genotypes, deze fenotypes zijn nu te gebruiken in de zoektocht naar mogelijke familieleden. Twee skelet samples waren afkomstig van twee broers, dit werd bevestigd door de STR genotypes te analyseren in combinatie met dat van een referentie sample afkomstig van een nog-in-leven-zijnde zus. De levende zus van de twee broers van wie de overblijfselen geanalyseerd zijn met behulp van STR-profilering, waarbij de familiebanden onthuld werden, bevestigde de oog en haarkleur van de broers zoals deze door HIrisPlex correct waren voorspeld. De veelbelovende resultaten van deze studie onderbouwen de toepasbaarheid van het HIrisPlex systeem in het verstrekken van aanwijzingen in onderzoek naar vermiste personen en slachtoffer identificatie.

In **Hoofdstuk 4.1** demonstreer ik de ontwikkeling van het HIrisPlex-S systeem, een uitbreiding van het HIrisPlex systeem om naast oog en haarkleur, ook huidskleur te voorspellen vanuit DNA. Een tweede multiplex met 17 SNPs gericht op huidskleur-specifieke SNPs werd toegevoegd aan de bestaande HIrisPlex multiplex, welke ook sommige huidskleur-informatieve SNPs bevat. Een toevoeging op het statistische voorspellingsmodel werd toegevoegd, deze leverde voorkeurs-aangepaste algehele voorspellingsnauwkeurigheden (uitgedrukt door AUC) van tussen de 0.76 en 1 voor de vijf categorien van huidskleur; erg licht, licht, medium, medium tot donker, donker. Bovendien onderging het HIrisPlex-S assay rigoureuze validatie testen inzake sensitiviteit, specificiteit, reproduceerbaarheid, precisie en

accuraatheid, mengsel-studies, 'casework' samples en populatiestudies, ter voorbereiding op de toepassing van het systeem in forensisch casework.

In **Hoofdstuk 5** richt ik mij op het tweede aspect van DNA intelligentie; het afleiden van de biogeografische matrilineale afkomst. In **Hoofdstuk 5.1** beschrijf ik de ontwikkelende validatie studie van een hiërarchisch systeem bestaand uit vijf multiplex assays gericht op alle 62 afkomst-informatieve mitochondriale SNPs voor het herleiden van de matrilineale biogeografische afkomst op continentaal niveau. Zoals ik laat zien, waren de assays erg sensitief en produceerden zij volledige profielen met een minimale inbreng van 1 pg DNA. Het testen op overeenkomstigheid tussen drie laboratoria leverde reproduceerbare en consistente resultaten. Bovendien lieten de assays zien dat ze erg robuust en efficiënt informatie konden verstrekken uit gedegradeerde samples en uit gesimuleerde casework samples van verschillende substraten zoals blood, sperma, haar, speeksel en samples met zeer lage hoeveelheden DNA. Met deze studie laat ik zien dat de assays voldoen aan de strenge richtlijnen van het SWGDAM (zoals in **Hoofdstuk 3.1** en **4**), en dat ze geschikt zijn om toe te passen in forensisch casework.

In **Hoofdstuk 5.2** wend ik mij tot identificatie van maternale afstamming, een andere forensische toepassing van humaan mtDNA die relevant is in zaken waarbij het bewijsmateriaal geen genoeg genomisch DNA bevat voor succesvolle STR profilering, maar nog wel genoeg mtDNA bevat voor mtDNA profilering. In plaats van mtDNA SNPs, zoals gebruikt worden in **Hoofdstuk 5.1** voor het herleiden van maternale afkomst, wordt sequentie data gebruikt van de mtDNA controle regio voor het identificeren van de maternale afkomst. Echter, betrouwbare databases zijn cruciaal voor het bepalen van populatie frequenties van mtDNA sequenties die overeenkomen tussen bewijs en referentie materiaal in forensisch casework. Met als doel het opstarten van een Nederlandse referentie database voor forensische toepassing, hebben we complete mtDNA control-regio sequenties gegenereerd onder ultra-hoge kwaliteitstandaarden in 680 willekeurig geselecteerde Nederlandse individuen van wie samples zijn verzameld op 54 locaties in heel Nederland waarbij alle geografische sub-regio's zijn vertegenwoordigd. Haplotype diversiteit van de gehele sample set was zeer hoog (99.63%) en, derhalve was de willekeurige overeenkomst zeer laag (0.63%). Er werd geen populatie substructuur binnen Nederland gedetecteerd in onze dataset.

In **Hoofdstuk 5.3** verschuif ik van mtDNA controle-regio analyses via klassieke Sanger sequencing (**Hoofdstuk 5.2**) naar gehele mitochondriale genoom analyses met

behulp van meer recentelijk gevestigde massaal-parallel-sequencing (MPS) ofwel next-generatie sequecing (NGS) technieken. MPS levert een oplossing om simultaan het gehele mtDNA genoom te analyseren waarbij het mogelijk is om hoge-doorvoer analyses uit te voeren tegen gereduceerde kosten per sample. In **Hoofdstuk 5.3** ontwikkel ik een gelaagde MPS benadering om simultaan het gehele humane mitochondriale genoom te sequencen gebruikmakend van 161 korte overlappende amplicons (gemiddeld 200 bp lang) met de Ion Torrent Personal Genome Machine. Het was mogelijk om gedetailleerde haplogroepen te herleiden met een inbreng van 100 pg genomisch DNA. Verder liet ik zien dat het mogelijk was om op volledige resolutie de maternale genetische afkomst op te lossen en af te leiden uit gedegradeerd DNA (ongeveer 220 bp).

Als laatste geef ik in **Hoofdstuk 6** een allesomvattende discussie van alle resultaten zoals deze beschreven zijn in **Hoofdstuk 2-5** van dit proefschrift. Ik bediscussieer de algemene bevindingen en hun forensische relevantie in routinematige casework toepassingen. Bovendien bespreek ik de uitdagingen waarmee we momenteel geconfronteerd worden bij het voorspellen van EVCs en BGA en bespreek ik hun potentiele toekomstperspectieven.

Curriculum vitae and Portfolio

Curriculum Vitae

Personal details

Name	Lakshmi Chaitanya Reddy
Date/Place of birth	26-07-1985, in Bangalore, India
Nationality	Indian
Address	Ringspoor 10
	2908 BC Capelle aan den Ijssel, The Netherlands
Email	v.subramanyamreddy@erasmusmc.nl
Mobile	06 83571868

Education

2012 – 2016	**PhD Studies** Department of Genetic Identification (formerly: Forensic Molecular Biology), Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands PhD thesis: *Genetic Approaches to Appearance and Ancestry: Improving Forensic DNA Analysis* Promotor: Prof. Dr. Manfred Kayser
2007– 2010	**Master of Science in Biomedical Sciences, specialization in Forensic Genetics** University of North Texas Health Science Center (UNTHSC) at Fort Worth, Texas. Master Diploma Thesis: *Validation of Leica LMD 6000 for separation of sperm and epithelial cells.* Promotor: Dr. Arthur J Eisenberg
2003 – 2007	**Bachelor of Engineering in Biotechnology** New Horizon College of Engineering, VT University, Bangalore, India Bachelor Diploma Thesis: *Effect of heavy metals on buffalo sperm motility using Computer Assisted Semen Analyzer (CASA).* Promotor: Dr. Snehalatha Nadigar

Work experience

2012 – 2016	**PhD research** Dissertation title: *Genetic Approaches to Appearance and Ancestry: Improving Forensic DNA Analysis* Department of Genetic Identification (formerly: Forensic Molecular Biology), Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands
2009 (3 months)	**Research Intern** Research project title: *Validation of Leica LMD 6000 for separation of sperm and epithelial cells.* Southwest Louisiana Criminalistics Laboratory, Louisiana, USA
2007 (4 months)	**Research Intern** Research project title: *Effect of heavy metals on buffalo sperm motility using Computer Assisted Semen Analyzer (CASA).* National Institute of Animal Nutrition and Physiology, Bangalore, India

List of Publications

1. Walsh S, **Chaitanya L**, Clarisse L, Wirken L, Draus-Barini J, Kovatsi L, Maeda H, Ishikawa T, Sijen T, de Knijff P, Branicki W, Liu F, Kayser M. *Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage.* Forensic Sci Int Genet. 2014 Mar;9:150-61.

2. **Chaitanya L**, van Oven M, Weiler N, Harteveld J, Wirken L, Sijen T, de Knijff P, Kayser M. *Developmental validation of mitochondrial DNA genotyping assays for adept matrilineal inference of biogeographic ancestry at a continental level.* Forensic Sci Int Genet. 2014 Jul;11:39-51.

3. **Chaitanya L**, Walsh S, Andersen JD, Ansell R, Ballantyne K, Ballard D, Banemann R, Bauer CM, Bento AM, Brisighelli F, Capal T, Clarisse L, Gross TE, Haas C, Hoff-Olsen P, Hollard C, Keyser C, Kiesler KM, Kohler P, Kupiec T, Linacre A, Minawi A, Morling N, Nilsson H, Norén L, Ottens R, Palo JU, Parson W, Pascali VL, Phillips C, Porto MJ, Sajantila A, Schneider PM, Sijen T, Söchtig J, Syndercombe-Court D, Tillmar A, Turanska M, Vallone PM, Zatkalíková L, Zidkova A, Branicki W, Kayser M. *Collaborative EDNAP exercise on the IrisPlex system for DNA-based prediction of human eye colour.* Forensic Sci Int Genet. 2014 Jul;11:241-51.

4. Liu F, Visser M, Duffy DL, Hysi PG, Jacobs LC, Lao O, Zhong K, Walsh S, **Chaitanya L**, Wollstein A, Zhu G, Montgomery GW, Henders AK, Mangino M, Glass D, Bataille V, Sturm RA, Rivadeneira F, Hofman A, van IJcken WF, Uitterlinden AG, Palstra RJ, Spector TD, Martin NG, Nijsten TE, Kayser M. *Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up.* Hum Genet. 2015 Aug;134(8):823-35.

5. Santos C, Fondevila M, Ballard D, Banemann R, Bento AM, Børsting C, Branicki W, Brisighelli F, Burrington M, Capal T, **Chaitanya L**, Daniel R, Decroyer V, England R, Gettings KB, Gross TE, Haas C, Harteveld J, Hoff-Olsen P, Hoffmann A, Kayser M, Kohler P, Linacre A, Mayr-Eduardoff M, McGovern C, Morling N, O'Donnell G, Parson W, Pascali VL, Porto MJ, Roseth A, Schneider PM, Sijen T, Stenzl V, Court DS, Templeton JE, Turanska M, Vallone PM, van Oorschot RA, Zatkalikova L, Carracedo Á, Phillips C; EUROFORGEN-NoE Consortium. *Forensic ancestry analysis with two capillary electrophoresis ancestry informative marker (AIM) panels: Results of a collaborative EDNAP exercise.* Forensic Sci Int Genet. 2015 Nov;19:56-67.

6. **Chaitanya L**, Ralf A, van Oven M, Kupiec T, Chang J, Lagacé R, Kayser M. *Simultaneous Whole Mitochondrial Genome Sequencing with Short Overlapping Amplicons Suitable for Degraded DNA Using the Ion Torrent Personal Genome Machine.* Hum Mutat. 2015 Dec;36(12):1236-47.

7. **Chaitanya L**, van Oven M, Brauer S, Zimmermann B, Huber G, Xavier C, Parson W, de Knijff P, Kayser M. *High-quality mtDNA control region sequences from 680 individuals sampled across the Netherlands to establish a national forensic mtDNA reference database.* Forensic Sci Int Genet. 2016 Mar;21:158-67.

8. **Chaitanya L**, Zupanič-Pajnič I, Walsh S, Balažic J, Zupanc T, Kayser M. *Bringing colour back after 70 years: Predicting eye and hair colour from skeletal remains of World War II victims using the HIrisPlex system.* Forensic Sci Int Genet. 2016 Oct;26:48-57.

Mandatory Courses	**Year**	**Workload**
Biochemistry and Biophysics (BB) course	2013	5 days
Genetics (G) course	2014	5 days
Biostatistical Methods I: Basic Principles A (CC02a) course	2014	4 days
Safely Working in the Laboratory	2014	1 day
Specific courses		
Forensic Summer School	2012	3 days
Basic Introductory Course in SPSS Course	2013	3 days
R Course	2013	4 days
Adobe Photoshop & Illustrator	2014	1 day
Adobe InDesign	2014	1/2 day
Microsoft Access Basic Course	2014	1 day
Microsoft Excel Advanced Course	2014	1 day
Seminars and workshops		
Forensic Genomics Consortium Netherlands (FGNC) Yearly Seminars	2011-2013	3 days
MGC PhD Symposia	2012-2013	2 days
MGC PhD Workshop **Munster**	2014	4 days
CLHC - Forensic PhD Symposium	2015	1 day
Conferences		
DNA in Forensics **Innsbruck**	2012	3 days
DNA in Forensics **Brussels**	2014	3 days
Posters and Presentations		
Departmental presentation	2013	1 day
International Society for Forensic Genetics (ISFG) **Poster**	2013	1 day
MGC PhD Workshop **(Munster) Poster**	2014	4 days
DNA in Forensics **(Brussels) Poster**	2014	3 days
International Society for Forensic Genetics (ISFG) **Presentation**	2015	5 days
Teaching activities (Supervision of 7 students)	2012-2016	
Training Courses		
DNA Sanger Sequencing **Innsbruck**	2012	5 days
Ion PGM Training **Darmstadt**	2014	3 days

Acknowledgements

Finally the day is here. My PhD journey would not have been possible without the immense encouragement and support from my colleagues, friends and family. Without their contributions and guidance I could not have accomplished my PhD. My heartfelt gratitude to all those who made this fruitful journey memorable.

Foremost, I would like to thank my honorific supervisor Prof. Manfred Kayser for giving me the opportunity to pursue a PhD. Dear Manfred, thank you so much for your constant support, advice and guidance. Your motivation and guidance have helped me grow as a research scientist. Your deep insights helped me at various stages of my research. Thank you for all the time and effort you took to go through and improvise my manuscripts and this thesis.

My sincere thanks to my reading committee, Prof. Andre Uitterlinden, Prof. Peter de Knijff and Prof. Ronny Decorte. Thank you for taking the time to read my thesis and for all your comments and suggestions to improve my thesis further. I would also like to extend my thanks to Prof. Walther Parson, Prof. Ate Kloosterman and Prof. Robert Hofstra for taking the time to come to Rotterdam and being part of the defence committee, despite your busy schedules.

Special thanks to my wonderful paranymphs Karolina and Noorie. Karolina, I cannot thank you enough for being so helpful and supportive. You are such a great colleague. I can never forget our wonderful chats about work, cooking and delicious cakes and polish delicacies. Noorie, thank you for all the motivation and being an inspiration. You are so energetic, fun-filled and always helpful. More than just friends, both of you are family.

My colleagues: Thank you to the fantastic team I have been fortunate to work with over the years. Susan, though you are in the US, you seem just a click away. Thank you for clarifying my doubts and helping me out whenever needed. Mijke, you are such an inspiration. Thank you for being such a great colleague and a special thanks for the samenvatting. Mannis, you were always very helpful. I have learnt a lot from you. Thank you so much for sharing your knowledge and guiding me. Thank you Arwin for all the help and support in the lab. Thank you Fan for helping me with all the statistics. Dmitry, Iris, Silke, Oscar, Kaiyin, Nefeli, Athina and Celia, it is such a pleasure to know you all and work with you all.

I am incredibly grateful to all my friends in India and Netherlands who stood by to support me along the way. Thank you for helping me during my PhD journey.

My family is my backbone. Words cannot express the feelings I have for my parents' unconditional support. Mummy and Daddy, I would not be here if it is not for you. Your firm and kind-hearted personality has affected me to be steadfast and never bend to difficulty. You always let me know that you are proud of me, which motivates me to work harder and do my best.

To my best half Vivek. You are a constant source of strength and inspiration. I can honestly say that it was only your determination and constant encouragement (and sometimes a kick on my backside when I needed one) that ultimately made it possible for me to see this journey through to the end. I am truly grateful to all my family members for all the love, support, and patience. Without my family, I would not have been able to thrive in my doctoral program or balance my research with everything else. Thanks for joining me in this scholarly adventure – I could not accomplish this feat without you all by my side.