Aus dem Institut für Rechtsmedizin
der Universität zu Köln
Direktor: Universitätsprofessor Dr. med. M. A. Rothschild

Binary Polymorphisms as Ancestry Informative Markers

Inaugural-Dissertation zur Erlangung der Würde
eines doctor rerum medicinalium
der Hohen Medizinischen Fakultät
der Universität zu Köln

vorgelegt von
Daniel Zaumsegel
aus Troisdorf

promoviert am 18.12.2013

Gedruckt mit Genehmigung der Medizinischen Fakultät
der Universität zu Köln
im Dezember 2013

Hundt Druck GmbH
Zülpicher Straße 220
D-50937 Köln

Dekan: Universitätsprofessor Dr. med. Dr. h.c. Th. Krieg

1. Berichterstatter: Universitätsprofessor Dr. rer. nat. P. M. Schneider

2. Berichterstatter: Universitätsprofessor Dr. med. B. Wollnik

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Dissertationsschrift ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskriptes habe ich Unterstützungsleistungen von folgenden Personen erhalten:
Universitätsprofessor Dr. rer. nat. P. M. Schneider
Frau Gabi Förster
Frau Kerstin Schöbel

Weitere Personen waren an der geistigen Herstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die Hilfe einer Promotionsberaterin/eines Promotionsberaters in Anspruch genommen.
Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertationsschrift stehen.

Die Dissertationsschrift wurde von mir bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Köln, 11. Mai 2013,

Die dieser Arbeit zugrunde liegenden Experimente sind nach entsprechender Anleitung durch Univ.-Prof. Dr. P. M. Schneider von mir selbst durchgeführt worden.

Acknowledgements

I would like to extend my thanks to my supervisor, Prof. Dr. Peter Schneider for providing me with the opportunity for some highly interesting research as well as for his continuous support throughout the years. I would also like to thank "my girls", Dr. Conny Schmitt, Dr. Magdalena Bogus, Dr. Iva Gomes, Gabi Förster, Kerstin Schöbel and Karin Schuster as well as the "following generation", Miriam Sirker and Theresa Gross for the pleasant work athmosphere in and around the lab, amazing birthday breakfasts and providing help and assistance whenever needed.

Zuguterletzt danke ich meiner Familie, die mich immer unterstützt hat, insbesondere meinen Eltern, die immer für mich da waren.
Ohne Euch hätte ich es nicht geschafft!

Meiner Familie

Table of contents

Glossary

AIM ancestry informative markers

AMOVA analysis of molecular variance

bp base pair(s)

CE capillary electrophoresis

DVI disaster victim identification

EVC externally visible characteristics

FDP forensic DNA phenotyping

InDel insertion/deletion polymorphism

kbp kilobasepairs

mb megabasepairs

MCMC Markov Chain Monte Carlo

mt-DNA mitochondrial DNA

PCR polymerase chain reaction

RFLP restriction fragment length polymorphism

SBE single base extension

SNP single nucleotide polymorphism

STR short tandem repeat

VNTR variable number of tandem repeats

1 Introduction

1.1 Historical Perspective on Forensic Genetics

The scientific field of forensic genetics is a relatively young and rapidly developing discipline. With the discovery of the ABO blood group system by Landsteiner in 1900 and the publication of a genetic theory for the inheritance of these blood groups by von Dungern and Hirschfeld in 1910, the field of forensic haematogenetics as a predecessor to modern forensic genetics was founded. In the following years, a large variety of blood group, red cell enzyme and soluble serum protein markers were discovered (reviewed in [40]) and highly discriminatory profiles could be obtained by analysing those markers in combination [46]. Since the markers used at that time were all protein based, large amounts of relatively fresh biological material were necessary for successful typing as proteins are prone to rapid degradation when exposed to the environment.

In criminal cases, where traces found at a scene of crime had to be compared with a reference sample of a suspect, these markers showed considerable limitations: Random match probabilities, i.e. the probability of two unrelated individuals sharing the same marker profile ranged between 0.01 and 0.001 when a set of 8 blood group systems could be analysed from a blood stain [54]. Other body fluids encountered at crime scenes (saliva, semen etc.) do not express all the available markers [42] and therefore random match probabilities in these cases were much greater [54]. With random match probabilities in these ranges, the evidential value of a positive match based on blood group systems remained relatively weak and needed to be backed by other evidence. An exclusion of a suspect based on these methods on the other hand was considered a fact [42]. Further problems could arise from the mixture of different tissues in criminal evidence, such as a mixture of semen and vaginal epithelial cells in a vaginal swab from a rape victim. Due to the generally much more abundant material of the victim and proteolytic enzymes in seminal fluid degrading the available protein, the material of the perpetrator was usually masked and therefore not detected [42, 54]. The described limitations were of no concern

in cases of paternity testing, since in these cases sufficient amounts of blood from all individuals in question could be readily acquired allowing the analysis of a maximum set of markers on a regular basis.

Developments in the 1960s and 1970s like restriction enzymes to cleave DNA at specific recognition sites [5], Sanger sequencing [104] and Southern blotting [113] made it possible to study DNA in more detail. Southern Blotting was first used to detect DNA polymorphisms (so called restriction fragment length polymorphisms, RFLP) in 1978 [56], but it took until 1985 for Sir Alec Jeffreys to realise the forensic potential of the minisatellite or variable number of tandem repeat (VNTR) loci he had been working on [50–52]. The method termed "DNA fingerprinting" by Jeffreys et al. used a common core motif as a "multilocus" probe in Southern Blot experiments, detecting multiple VNTR markers simultaneously resulting in highly variable multi-band electropherograms. While this method used a lot less biological material than the previously mentioned protein based methods and can be applied on any body fluid or tissue containing DNA with equal discriminatory power, it still requires DNA amounts in the range of micrograms [52]. However, due to the fact that multiple polymorphic sites were analysed with the same probe simultaneously, the discriminatory power of this method was already extremely high compared to the previously analysed protein markers with a random match probability of 3×10^{-11} for a single probe and 5×10^{-19} when two probes were analysed together [52].

The VNTR polymorphisms analysed are repetitive sequences with repeat lengths of between ten and 100 base pairs repeating between 3 and 100 times per allele resulting in total fragment lengths in the range of kilobases [51]. This limits the use of VNTR polymorphism analysis in forensic genetic case work since a "genetic fingerprint" can only be generated from good quality, high molecular weight DNA. The quality of DNA obtained from a crime scene strongly depends on the environmental conditions at the scene as well as the methods of sampling and storage prior to analysis. While it is generally no problem to successfully analyse fresh samples, success rate can be considerably reduced in older samples up to the point where profile generation becomes impossible due to DNA degradation [42]. Due to the fact that multiple polymorphisms are analysed simultaneously with the same probe, resolution of mixed stains is not possible with this method.

While the multilocus probes remained in use for some years for paternity testing cases due to their superior discriminatory power, they were rapidly

replaced by specific probes cloned for the detection of individual highly polymorphic VNTR loci (so called single locus probes) making interpretation of the resulting electropherograms much easier [54]. Instead of a single multilocus probe detecting several VNTR polymorphisms simultaneously, single locus probe were applied to the Southern blot sequentially, typically four probes per blot, resulting in clear band patterns of two fragments per individual per probe and therefore even allowing the detection and possible resolution of mixtures [55].

The largest limitation of the forensic genetic methods up to that point, namely the large amount of material needed, could finally be overcome after the development of the polymerase chain reaction (PCR). The PCR process, first described by Kary Mullis [100, 101], made it possible to amplify specific regions of DNA *in vitro*, increasing the sensitivity of the DNA analysis up to a point where DNA almost from single cells could be reliably analysed. Because PCR-based detection methods were not only much more sensitive but also far less time consuming, allowing the generation of a DNA profile in the time frame of several hours instead of days, the classic RFLP-typing methods using single locus probes and Southern Blotting were rapidly replaced in forensic genetic routine. Due to their high discriminatory power however, these markers remained in use for some time longer in the field of paternity testing and kinship analysis.

The first reported PCR based application in forensic genetics was the analysis of sequence variation in the mitochondrial DNA used for the identification of skeletal remains by Stoneking et al. in 1991[117]. Shortly after, the analysis of short tandem repeat (STR) or microsatellite markers emerged as the method of choice for forensic DNA analysis [20, 48, 53]. This marker class had been recently discovered [71, 133] and was structurally similar to the formerly used minisatellite markers. In contrast to minisatellites however, the repeat motifs of these STR markers consisted of only two to six base pairs repeated usually less than 30 times per allele resulting in total amplicon lengths of less than 350 bp [18, 21, 32, 46, 71, 133]. The short fragment lengths of STR markers allowed the generation of DNA profiles from degraded DNA found for example in skeletal remains exposed to the environment for extended time periods [48, 53]. DNA mixtures can readily be detected and even assessed in a semi-quantitative way based on signal intensities [114].

1.2 State of the Art

Modern day forensic genetics comprises of three related major fields of interest: analysis of biological traces found at a crime scene and identification of the possible donor of such stains, paternity testing and kinship analysis, and the identification of victims of mass disasters. Biological trace analysis generally deals with the analysis of limited biological material of often questionable quality. Therefore, the techniques employed need to be highly specific and sensitive in order to provide useful results. Identification of stain donors is achieved by directly comparing DNA profiles from crime scene samples with reference profiles obtained from suspects, victims, witnesses or other individuals possibly relevant to the crime scene. In cases without suspects, DNA profiles of crime scene samples are routinely compared to national DNA databases set up in many nations throughout the world. Such databases usually contain reference DNA profiles of convicted criminals or suspects in criminal investigation as well as DNA profiles of crime scene stains with unknown donor [109].

In the area of paternity and kinship analysis, the amount of available DNA and its quality is usually of no concern. DNA profiles of possibly related individuals are evaluated based on Mendel's laws of inheritance and assessed for the compatibility with the relationships in question. While the identification of stain donors is based on a direct, one to one comparison of two DNA profiles, paternity and kinship testing evaluates the portions of a DNA profile which are inherited from one generation to the next. Therefore, only parts of the available genetic information of each DNA profile can be used for the comparison. In addition, the possibility of mutations and other genetic effects have to be taken into account. Identification of victims of mass disaster (termed DVI - disaster victim identification) usually requires the identification of bodily remains by DNA analysis which are impossible to identify by other means. Methods employed combine techniques of direct identification similar to those used in criminal investigations with kinship analysis methods. DNA profiles obtained from the remains can be compared to DNA profiles obtained from personal belongings of missing persons or by kinship analysis involving living relatives of the potential victims.

1.2.1 Introduction to Forensic Biostatistics

In order to assess the weight of the evidence provided by DNA analysis, biostatistical calculations are employed. Identification of a stain donor in criminal casework is achieved by direct comparison of the DNA profile of a crime scene sample with the DNA profile of an individual in question. In cases where the DNA profiles do not match, the individual in question is excluded as the donor of the analysed stain with absolute certainty. If the DNA profiles do match, the so called match probability, i.e. the probability of a random, unrelated individual possessing the same DNA-profile is calculated. This calculation is based on population genetic principles of the Hardy-Weinberg Law providing a mathematic representation of the relationship between genotype and allele frequencies in a given population [49, 116]. The Hardy-Weinberg Law states that, within an infinitely large, randomly mating population, the genotype frequencies at any given locus remain constant and are dependent on the allele frequencies of the available alleles. The relationship of the genotype frequency and the allele frequencies of the given alleles can be described by a simple mathematical equation:

Given a locus with two alleles A and B with respective allele frequencies p and q, the frequencies of the three possible genotypes AA, AB and BB can be calculated according to the simple binomic formula $p^2 + 2pq + q^2 = 1$ with the genoptype frequency of genotype $AA = p^2$, the frequency of genotype $AB = 2pq$ and the frequency of genotype $BB = q^2$.

The requirements of the Hardy-Weinberg Law, namely infinite population size, random mating, absence of migration, natural selection or mutations, are obviously not met by modern-day human populations. Still, the law can be considered to be applicable for the estimation of genotype frequencies in forensic genetics. Finite population size leads to a phenomenon of random genetic drift slightly varying the allele frequencies between one generation and the next. Genetic drift is more pronounced the smaller the population in question, but even in small isolated populations it has been shown that genetic drift will not lead to the loss of alleles with a frequency $> 1\%$ [28, 84]. Most populations relevant for forensic genetic analyses are sufficiently large for genetic drift to have no significant effect. The prerequisite of a randomly mating population can be considered as fulfilled, although humans do not mate completely randomly. Since the genetic markers used for forensic genetic analyses are generally non-coding and therefore do not influence phenotypes potentially leading

Table 1.1: Advantages and disadvantages of forensic genetic marker classes

Marker Class	Advantages	Disadvantages
Autosomal STR	high variability, easy detection, deconvolution of mixtures	genetic instability, susceptible to DNA degradation
Y-STR	lineage specificity, male/female mixture resolution, reconstruction of pedigrees	only present in males, no discrimination within paternal lineages, linkage between markers
X-STR	advantages in solving certain deficiency cases in kinship analysis	linkage between markers
SNP	small fragment size, genetic stability, high potential for automation	low variability, more difficult and time consuming detection
Indel	small fragment size, genetic stability, easy detection, high potential for automation	low variability
mt-DNA	lineage specificity, resistance to degradation, multiple copies per cell	no discrimintation within maternal lineages

to mating decisions, human mating pattern can be considered random at least for the markers used. The non-coding characteristic of the used markers also ensures the absence of natural selection.

In order to determine the frequency of a complete DNA profile, the genotype frequencies of each marker are simply multiplied. This "product rule" of probability theory only applies for independent events. Therefore, it can only applied if all the markers analysed for a given DNA profile are segregating independently [23], essentially proven to be fulfilled for the standard set of STR markers widely used in forensic genetic practice [91].

1.2.2 Forensic Genetic Markers

An array of different genetic markers are used throughout the field with varying applicability depending on the task at hand and the individual properties of the markers. Advantages and disadvantages of the commonly used marker classes are summarised in Tab. 1.1 and described in more detail below.

Short Tandem Repeat Markers

Short tandem repeat markers are repetitive sequences with a core motif of about 2 - 6 basepairs (bp) in length and generally have alleles with lengths below 350 bp [46]. They are found throughout the human genome including the 22 autosomes and the X and Y gonosomes with an estimated abundance of about 1 locus in 20 kb for tri- and tetrameric STRs [32]. Computational methods analysing the available human genome sequence data have identified more than 100.000 STRs of which more than 20.000 are of the tetrameric type most commonly used in forensic genetics [21]. STR markers combine a variety of favourable features for their application in forensic genetics.

Their overall short amplicon length makes them suitable for easy amplification via PCR allowing successful typing from very low amounts of even low quality degraded DNA [114]. Although STRs are not as polymorphic as the historically used VNTR polymorphism, the ability to analyse a large number of loci in parallel by multiplex-PCR results in comparably discriminatory profiles while being technically more robust and more easily interpretable as well as less time consuming [19]. Modern multiplex PCR systems utilising fluorochrome labelled PCR primers [64] and capillary electrophoresis allow for the analysis of markers with similar fragment lengths in the same reaction by using different fluorescent dyes and multicolour-detection.

Autosomal STRs Currently a set of around 25 core markers is used throughout the forensic community in slightly varying combinations [18, 43, 44]. Most of these markers have been chosen either because they have been used in forensic casework prior to the creation of national DNA databases or because they have already been part of commercially available STR typing kits at the time of the creation of such databases [18]. The most recent addition to this standard set of markers are the five markers chosen by the European DNA Profiling Group (EDNAP) to improve the discriminatory power when sharing DNA data across multiple countries [43, 44] and later adopted by an EU Council Recommendation [126]. Availability of commercial kits for typing of the selected markers as well as the combined effort of the worldwide forensic genetics community have produced a large amount of allele frequency data for a large variety of populations, thus further increasing the usefulness of the core marker set [17, 19]. Due to differences in the allele frequency distribution between different populations, it has been suggested that STR profiles could

be used to infer the biogeographic ancestry of an unknown stain donor [72] However, the usefulness of autosomal STRs for this purpose is limited due to their relatively high mutation rate of approx. 0.1 to 0.5% per meiosis [18].

The STRs selected as core markers are spread out over all 22 autosomes with only very few markers sharing the same chromosome. Even the markers located on the same chromosome are spaced at a distance of tens of million base pairs, so independent segregation of most of the used markers can be assumed [6, 18]. Notable exceptions are the markers D5S818 and CSF1PO used for example in the United States national DNA database, which are located on chromosome 5 at a distance of approx. 26 Megabasepairs (mb), and the markers D12S391 and vWA used throughout the world in national DNA databases, which are both located on chromosome 12 at a distance of only 6.3 mb. While the first pair, D5S818 and CSF1PO have been thoroughly evaluated and found to be inherited independently [7], some dispute has recently arisen over the possibly linked inheritance of the marker pair D12S391 and vWA [83], but the effect has been shown to be negligible in identity and simple kinship cases while methods are available to account for this effect in more complex kinship cases encompassing more than two meioses [45].

Gonosomal STRs Polymorphic STR markers are not only found on the autosomes, but also on the two remaining chromosomes, the X and the Y gonosomes. However, due to the special features of the sex chromosomes, these markers are significantly different from the autosomal STRs. There are also considerable differences between markers located on the X chromosome and those on the Y chromosome. While the Y-chromosome is only present in males, and then only in a single copy as opposed to the two copies available for each autosome, the X-chromosome is present in one copy in males and in two copies in females. These peculiarities, while on the one side requiring special consideration when using these markers in combination with autosomal STRs, on the other hand make the gonosomal STRs highly interesting for certain scenarios such as reconstruction of family trees in DVI cases or the deconvolution of mixed stains in cases of sexual assault.

Y-chromosomal markers As stated above, the Y-chromosome is only present in males as a single copy accompanied by a single copy of the X-chromosome. Because of its small size, no recombination during meiosis occurs in the largest

part of the Y-chromosome thus leading to all currently used markers on the Y-chromosome being inherited as a haplotype. Currently a set of 17 Y-STRs is in use throughout the forensic genetic community after thorough evaluation and standardisation of the methods involved in analysing haplotypic markers [60, 65, 78]. Because it is impossible to extrapolate haplotype frequencies from allele frequencies at each locus by multiplication due to complete linkage between the loci, these frequencies can only be estimated based on whole genomes. This can be accomplished based on large databases of properly curated haplotype profiles, such as the "Y Chromosome Reference Database" (YHRD) [96]. Y-chromosomal STRs are a tool to gain additional information not available through analysis of autosomal STRs in certain special scenarios:

In cases of stains with male/female mixtures, a clear Y-STR profile of the male component can be obtained at ratios of over 1:1000 [93]. In cases of sexual assault, where usually the female component in a given sample outweighs the male component, Y-STR analysis regularly allows a full profile of the male component to be obtained where autosomal STRs show a complete absence of a mixture.

Because the Y-chromosome is only present in males and because no recombination occurs between the Y-STR markers, the Y-chromosome is inherited in a patrilinear fashion. While this is a drawback in the identification of individuals since all male members of a family descendant from one father will exhibit the same Y-STR profile, this can be considered a positive feature in other circumstances.

The lineage-specificity of Y-STR profiles allows for the reconstruction of pedigrees in deficiency cases e.g. in the identification of mass disaster victims. Reconstruction of family pedigrees is possible over many generations along the paternal line using Y-STRs making them useful for genealogical studies as well. Another useful feature of the lineage-specificity is the possibility to infer the more recent biogeographical ancestry of a stain donor from the analysis of Y-STR profiles [27], however ancestry estimates based on the Y-chromosome alone might be misleading.

X-chromosomal markers While the X-chromosome exists as a homologous pair in females and can, with some limitations described below, be considered in a way similar to autosomes including the occurrence of meiotic recombination, it is only hemizygous in healthy males, thus comparing more closely to

the Y-chromosome. A set of 30 STRs spread out over the entire X-chromosome has been evaluated for forensic use up to date [120, 121]. Due to the close proximity of the used markers to each other, linkage disequilibrium and pairwise linkage have to be considered when dealing with markers situated on the same chromosome. The X-STR markers used routinely in forensic genetic applications have been shown to be clustered in four linkage groups with very little recombination occurring within each linkage group while recombination rates between the linkage groups is considerably higher [82].

Further difficulty arises from the fact that several rare chromosomal aberrations involving the X-chromosome do exist and are compatible with life. The discovery of such aberration or the diagnosis of testicular feminisation (XY genotype in a phenotypically female person due to androgen insensitivity[135]) rules out the use of X-chromosomal markers for analysis [120]. On the other hand, X-chromosomal analyses can offer invaluable information in a variety of forensic cases.

While X-STR analysis for the identification of individuals is at its best of equal power as the analysis of autosomal STRs in cases of female individuals and of considerably lower discriminative power in males, it can help greatly in the detection of female trace material in a mixture with overwhelmingly male background [120]. The main forensic genetic application of X-chromosomal markers however, is the benefit this marker class gives in deficiency cases of kinship testing.

X-STRs provide advantages in cases where either father/daughter or mother/son relationships are questioned while they perform comparably to autosomal STRs in cases of mother/daughter relationships. X-STRs do not offer any information in father/son cases, since the son inherits his only X-chromosome from his mother. Because the X-chromosome found in male individuals has to be inherited from the mother, X-STR analysis can also help solving deficiency cases where the alleged father is not available for analysis if other relatives, such as a sister or daughter are available for testing, and the child in question is female [120].

Mitochondrial DNA

Mitochondrial DNA (mtDNA) is a small, circular molecule of DNA independent of the genomic DNA. MtDNA was discovered as a distinct molecule in the 1960s [81, 107] and fully sequenced by the beginning of the 1980s by Anderson

et al.. This first sequence, later corrected by Andrews et al., termed the "revised Cambridge Reference Sequence" (rCRS), revealed the mtDNA molecule to be of 16.569 bp in length containing 37 genes. Because of the large amount of genes on the rather short mtDNA molecule, variation within the sequence is quite limited, most of the variation occurring in a short, gene-free stretch termed the mitochondrial control region or "D-loop". The variation detected in the mitochondrial DNA consists almost entirely of single nucleotide polymorphisms. Therefore, sequencing of parts or the whole D-loop of mtDNA has become the method of choice for the analysis of mtDNA variation [85].

In order to efficiently communicate observed mitochondrial variation, reporting the differences observed when comparing to the rCRS has been adopted as straightforward method throughout the sciences and nomenclature rules have been set to standardise reporting where this method is ambiguous [8]. MtDNA is an interesting target for forensic genetic research because it is present in each cell in multiple copies and the mitochondria are exclusively inherited in a matrilinear way [12, 13] but in contrast to the Y-chromosome transmitted equally to male and female offspring. The multicopy presence of mtDNA per cell (depending on cell type in a ratio of up to 1000:1, oocytes up to 200.000, spermatozoa only 50-100:1) makes it possible to still obtain enough DNA for mtDNA analysis from very low amounts of biological material. Also, due to its small size, the mtDNA is considerably more resistant to degradation than human genomic DNA allowing successful typing of mtDNA in old samples or difficult material such as telogenic hair shafts [85].

The matrilinear inheritance pattern of mtDNA has much the same implications as the patrilinear inheritance of the Y-chromosome, mainly the possibility to estimate the matrilinear biogeographic ancestry of a sample at the cost of not being able to discriminate individuals from the same matrilinear line.

Similarly to the Y-chromosomal analysis, haplotype frequencies for mtDNA can only be estimated on the basis of large databases. The most comprehensive mtDNA database to date, EMPOP, has been set up as a collaborative project initiated on a suggestion by the European DNA Profiling Group (EDNAP) in 1999 and has been made publicly available in 2006 [86].

Single Nucleotide Polymorphisms

In contrast to STR systems, single nucleotide polymorphisms are sequence variations involving the substitution of a single base in the DNA. To distin-

guish SNPs as a true polymorphism from rare genetic variants or individual differences, the allele frequency of the least abundant allele in a population has to be $> 1\%$ [15].

SNPs are highly abundant in the human genome at with about one polymorphic site per 1000 bp [24, 75, 99, 122]. SNPs found in the human genome are almost completely di-allelic (i.e. displaying only two distinct alleles) with tri- or tetra-allelic SNPs being extremely rare [15].

The fact that there are only two alleles per locus is the main disadvantage of SNPs in comparison with STRs, since in order to obtain a similar discriminatory power as the 10-15 STRs currently used for forensic identification purposes, a set of around 50 SNPs is required [2, 41]. This also limits the usefulness of SNPs in mixture resolution.

On the other hand, SNPs have several distinct advantages for use as forensic markers: Because the polymorphism consists of only a single base, SNPs can be amplified in extremely short fragments of less than 100 bp in length, making them especially suitable for the analysis of highly degraded DNA [36, 37]. Also, the mutation rate is considerably lower than that of STRs [89], making them especially suitable for kinship cases where mutations of STR loci have to be considered [108]. Because of their low allele count, accurate estimates for allele frequencies can be acquired by the analysis of a smaller number of samples than with STRs, making validation of SNPs for forensic purposes easier [103]. It is also possible to efficiently amplify the short fragments needed for SNP analysis in single-tube multiplex PCR reactions allowing the analysis of a large number of individual markers from very low trace amounts of DNA. Although several SNP typing panels achieving a sufficient power of discrimination using several different high-throughput genotyping techniques (e.g. [29, 89, 103]) have been developed, SNP typing has not become a routinely used method in forensic genetics.

The main reason for not using SNPs in routine forensic genetic casework is the fact that these markers are not part of the national DNA databases being used to identify donors of unknown stains in criminal investigations. Therefore SNP typing for the purpose of identification can only be used as a supplementary method in cases where reference material is available for parallel analysis. Another downside to SNP analysis in routine forensic laboratories is the fact that due to SNPs being sequence polymorphisms instead of length polymorphisms, typing methods require different technical prerequisites as well as specialised knowledge not available in every laboratory [90].

Insertion/Deletion Polymorphisms

Insertion/Deletion polymorphisms (indels) are the most recent addition to the forensic genetic portfolio of markers having emerged during the past 10 years. Indels are able to bridge the gap between the two most common forensic markers, STRs and SNPs by combining the advantages from both worlds.

The fact that indels stem from a single mutation event, namely the spontaneous insertion or deletion of a DNA fragment during DNA replication, occurring with a low frequency makes them genetically quite stable [79]. Because they are length polymorphisms, they can be analysed by the same flourescently labelled PCR and electrophoresis techniques as STRs. They are highly abundant throughout the genome with one indel every 1.5 kb according to the landmark study of Mills et al. (2006) revised in 2011 [76, 77].

The majority of indels identified in these studies being shorter than 100 bp in length offers the possibility for large scale multiplexing and makes these markers ideally suited for the analysis of degraded DNA as shown by Pereira et al. (2009). Because of the desirable features of indels combined with their ease of incorporation into the workflow of the routine forensic genetic laboratory, a first commercial indel typing kit has become available and has already been validated for routine use [68]. This first commercially available indel typing kit is marketed by the Qiagen company (Hilden, Germany) as Investigator DIPplex® kit containing 30 indel markers as well as the amelogenin locus for sex determination. The markers are distributed over 18 autosomes and can be analysed in a single tube PCR assay and detected by capillary electrophoresis using five flourochromes.

The large number of available indels makes it possible to select panels for a variety of uses in addition to the identification of individuals, including the analysis of genetic structure in human populations [10, 97, 98], inferring biogeographic ancestry [88, 136] and assessing individual admixture [106]. That said, indels are up to now only used as a supplementary method in forensic genetics due to mainly the same reasons as SNPs - missing support in national DNA databases and limited power for the resolution of DNA mixtures.

1.3 New Approaches in Forensic Genetics

The methods currently employed in the forensic genetic routine laboratory make use of statistical methods to assess the weight of the evidence based

on a comparison of genetic profiles obtained from a crime scene sample with the genetic profiles obtained from a reference sample or database of DNA profiles. In cases, where intelligence is scarce, no witnesses or suspects are available and a database search with a crime scene DNA profile does not yield any result, a criminal investigation might hit a dead end despite the fact that good quality DNA traces have been recovered from the crime scene. Recent forensic genetic research is focussed on obtaining further intelligence from crime scene DNA samples in order to guide criminal investigations, such as information about externally visible physical characteristics of the stain donor or his biogeographic ancestry [59].

1.3.1 Physical Traits

Current forensic genetic routine analysis using the commercially available STR typing kits in many cases already includes the gender of the stain donor as such a physical trait utilising the analysis of a length difference of the Amelogenin gene between the X- and the Y-chromosomal version of the gene [73]. While this test is known not to be fail-safe [14, 105], it is still the most accurately predictable externally visible trait to date [59]. The gender alone is however only moderately useful in reducing the number of suspects in crime cases without other leads. More useful would be the possibility to deduce information about the suspects externally visible characteristics (EVCs) from the biological stains left at the crime scene.

While human facial features are understood to be individual-specific except for monozygotic twins, recent advances in genetics like high-density microarray-based genotyping technologies have made the discovery of genetic markers for group-specific complex traits like some broader externally visible characteristics a possibility [25, 95]. The most promising advances in this area have been achieved in finding genetic markers corresponding to the eye- and hair-colour of a stain donor.

Starting with the discovery of several SNPs in the melanocyte-stimulating hormone receptor (MC1R) gene explaining a red hair and fair skin phenotype in northern Europeans in 1995 [127], several other genes involved in melanogenesis have been reported to be involved in the definition of skin- and hair-colour and associated markers (reviewed in e.g. [58, 124]).

Further analysis of the genes relevant for melanogenesis revealed that iris-colour is another physical trait influenced by much the same set of genes

[61, 119]. The data about pigmentation related markers accumulated to date suggest, as is to be expected, that extreme phenotypes such as red hair, fair skin, blue or brown eyes, are most accurately predictable while intermediate phenotypes are more difficult to discriminate [33, 59, 118].

Considering that many of the markers associated with pigmentation, especially the markers associated with iris colour, used in the studies described are non-coding and therefore not causative for the variation in question causes further difficulty in the analysis.

Variation of iris colour is an almost exclusively European phenomenon caused most likely by a "founder effect" mutation leading to preferential choice of attractive eye colours in mating decisions [38] while elsewhere only very little variation of eye colour is found. Non-causative markers with association to eye-colour in European populations can however still be found in non-European populations without such association [59].

Therefore, in order to correctly predict such externally visible characteristics based on associated but not causative markers, accurate knowledge of the biogeographic origin of the stain donor is required. In a forensic context, this results in the need to accurately predict the biogeographic ancestry from the DNA in combination with the prediction of externally visible characteristics. In a reciprocal approach however, this regional specificity of certain physical traits and associated markers may allow these markers to be used as ancestry informative in the prediction of biogeographic ancestry.

1.3.2 Biogeographic Ancestry

Assessing the geographic origin of an individual based on a DNA sample is possible by several strategies. Based on the analysis of lineage-specific markers such as Y-chromosomal or mitochondrial DNA analysis, it is possible to establish the possible origin of a stain donor's paternal (Y-chromosome) or maternal (mt-DNA) lineage by phylogenetic approaches [27, 57, 128].

Both the mitochondrial DNA as well as the Y-chromosome show genetic features highly advantageous for this kind of analysis, namely uni-parental inheritance, lack of recombination and comparably small effective population size [55].

Because of these special population genetic features, these markers are especially susceptible to genetic drift leading to accelerated differentiation between haplogroups allowing the establishment of phylogenies accurately describing

the migration history of the human species [125] and therefore allow the prediction of biogeographic ancestry of a DNA sample. On the other hand, assessing admixture proportions or recent demographic changes are more difficult to perform using these kinds of markers.

When dealing with admixed populations, as is the case with all modern day populations to a certain extent, lineage-specific markers will be useful in identifying the population(s) of origin based on the paternal and/or maternal lineages. At the same time, the results based on lineage markers may be divergent from autosomal markers and thus give misleading results regarding the actual population a sample donor belongs to. This population stratification and admixture can be used as an advantage in large scale genome wide association studies using a case/control set-up due to the linkage disequilibrium introduced [22, 111]. On the other hand, these effects might also lead to false-positive associations due to population substructure not accounted for by the available demographic information of the tested subjects [25]. Accounting for those effects in genome wide association studies can be as easy as using the available data of unselected markers to estimate genetic admixture of the typed samples. In smaller scale follow-up studies, the use of a specially selected panel of ancestry informative markers (AIM) in order to account for these effects may be necessary [123]. Markers suitable for the inclusion in such ancestry informative panels ideally present with large allele frequency differences between different ancestral or geographically distant populations.

While autosomal STR markers routinely used in forensic genetics have been used to this means with some success in the past [19, 72, 110], these markers are less suited for this task due to their mutational instability and resulting high intra-population variability compared to relatively low inter-population variability [89]. More recently, single nucleotide polymorphisms (SNPs) have emerged as the marker class of choice for this task [39, 63, 90, 102] due to the availability of high-density SNP-typing technologies. Based on their previously described technical advantages in comparison to SNPs, indel markers have also received recent attention as ancestry informative markers [10].

2 Publications

2.1 SNPs for the analysis of human pigmentation genes - A comparative study

Zusammenfassung: Die Bestimmung charakteristischer phänotypischer Merkmale eines möglichen Spurenverursachers aus einer am Tatort gesicherten Spuren-DNA ist in unterschiedlichen forensischen Fällen von Interesse. Kandidaten für solche Marker finden sich in Assoziation mit Genen mit Einfluss auf die Pigmentierung und damit auf Haut-, Augen- und Haarfarbe. Die hier vorgestellte Studie zielt auf die Etablierung eines Marker-Panels zur Untersuchung der Pigmentierung eines unbekannten Spurenlegers anhand seiner DNA-Probe.

Citation: Zaumsegel, D., Rothschild, M. A., Schneider, P. M.,: SNPs for the analysis of human pigmentation genes - A comparative study. Forensic Science International: Genetics Supplement Series 1 (2008), 544–546 [137]
Reproduced with friendly permission of Elsevier Ltd., United Kingdom

Available online at www.sciencedirect.com
ScienceDirect
Forensic Science International: Genetics Supplement Series 1 (2008) 544–546
www.elsevier.com/locate/FSIGSS

Research article

SNPs for the analysis of human pigmentation genes—A comparative study

Daniel Zaumsegel *, Markus A. Rothschild, Peter M. Schneider

Institute of Legal Medicine, University of Cologne, Melatenguertel 60-62, 50823 Cologne, Germany

Received 9 September 2007; received in revised form 18 October 2007; accepted 7 November 2007

Abstract

In specific forensic cases, a genetic marker set allowing to deduce information about phenotypic features of the individual in question may be helpful to investigators. Candidates for such markers include pigmentation genes relevant for eye, hair and skin colour.

The project presented here aims at the development of a marker set which may be able to derive phenotypic information regarding the pigmentation pattern of an individual from DNA.
© 2008 Elsevier Ireland Ltd. All rights reserved.

Keywords: SNP; Pigmentation genes; AIM

1. Introduction

A set of 11 single nucleotide polymorphisms (SNPs) for which linkage to such phenotypic information has been published in the recent literature [1–4] have been selected to develop an initial assay based on the SNaPshot technology. A multiplex-PCR to amplify these SNPs has been developed and appropriate single-base extension primers have been designed (Table 1).

After optimisation of reaction conditions to allow efficient amplification and extension of all products in a simple two-step procedure (multiplex-PCR and multiplex-SBE), followed by capillary electrophoresis and automated allele calling using the GeneMapper 4.0 software, a first study has been performed to gather information about allele distribution in the general population. In a comparative study, two populations with strongly varying phenotypes (i.e. northern Europeans vs. sub-Saharan Africans) are currently being tested for significant differences regarding the selected SNPs.

2. Samples and methods

Two groups with 71 individuals of sub-Saharan African origin and of 87 individuals of Northern European origin (mainly German) have been selected on the basis of their origin (place of birth) from routine paternity cases. Whenever possible, unrelated individuals have been chosen.

A panel of 11 SNPs (refer to Table 1) has been selected for the development of a multiplex assay based on the SNaPshot technology (Applied Biosystems). Both, PCR- and SBE-primers have been designed to allow all reactions to be run in multiplex assays. For the PCR-primers, design criteria were:

- final product length between 200 and 300 bp; T_m around 60 °C; no cross reactions between primers and other PCR products.

Criteria for the design of the SBE-primers were as follows:

- Fragment length of the SBE products between 20 and 60 bp; no mis-priming on any of the other PCR products; T_m of the specific part of each primer around 60 °C; distance of at least 5 bp between SBE fragments detecting in the same dye channel; fragments detecting in different dye channels may overlap in their length.

Both PCR- and SBE-primers have been designed using the Primer3 software via the web interface [5] and tested for cross reactions and mis-priming using a locally installed BLAST programme [6].

After optimisation of the PCR- and SBE-protocols for multiplex reactions with the complete set of SNPs in one reaction, a first genotyping study has been performed on two populations of routine samples available in the laboratory. Fragment separation was carried out using an AB 3130 capillary sequencer with POP-4 polymer, followed by GeneMapper 4.0 analysis. Genotype results were exported into a database, where statistical analyses were performed.

* Corresponding author. Tel.: +49 221 478 88222.
E-mail address: daniel.zaumsegel@uk-koeln.de (D. Zaumsegel).

1875-1768/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.fsigss.2007.11.016

Table 1
Data on the selected SNPs, corresponding dbSNP resources, SBE-primers, and detection strategy

SNP	dbSNP Reference	Location	Strand	SNP	T_m	Fragment length	Detection bases	Detection dyes	Published in
A	rs1800401	OCA2	–	C/T	54.7	25	C/T	Red/black	[1]
B	rs7170989	OCA2	+	C/T	60.6	25	G/A	Blue/green	[1]
C	rs4778138	OCA2	+	A/G	56.3	30	T/C	Red/black	[1]
D	rs1426654	SLC24A5	–	A/G	59.2	30	A/G	Blue/green	[4]
E	rs16891982	SLC45A2	+	C/G	60.3	35	G/C	Blue/black	[4]
F	rs26722	SLC45A2	–	C/T	58.8	40	C/T	Red/black	[3,4]
G	rs7495174	OCA2	–	A/G	59.2	40	A/G	Blue/green	[1]
H	rs3733808	SLC45A2	+	C/G	60.9	45	G/C	Blue/black	[2]
I	rs1375164	OCA2	–	C/T	60.3	50	C/T	Red/black	[1]
K	rs4778231	OCA2	+	C/T	60	50	G/A	Blue/green	[1]
M	rs1800407	OCA2	–	A/G	60.3	55	A/G	Blue/green	[1]

3. Results and conclusions

After optimising primer concentrations reaction conditions, multiplex-SBE typing of the 11 SNPs could be carried out reliably. The automatic allele calling of the GeneMapper 4.0 software was routinely used (Fig. 1). Although the bins overlap for different detection colours in the graphical display (Fig. 1a), this does not affect the analysis procedure.

In a first study the newly established multiplex SNP-typing method has been used to type the two sample collections described above. Genotype frequencies have been calculated and compared to published data obtained from dbSNP and

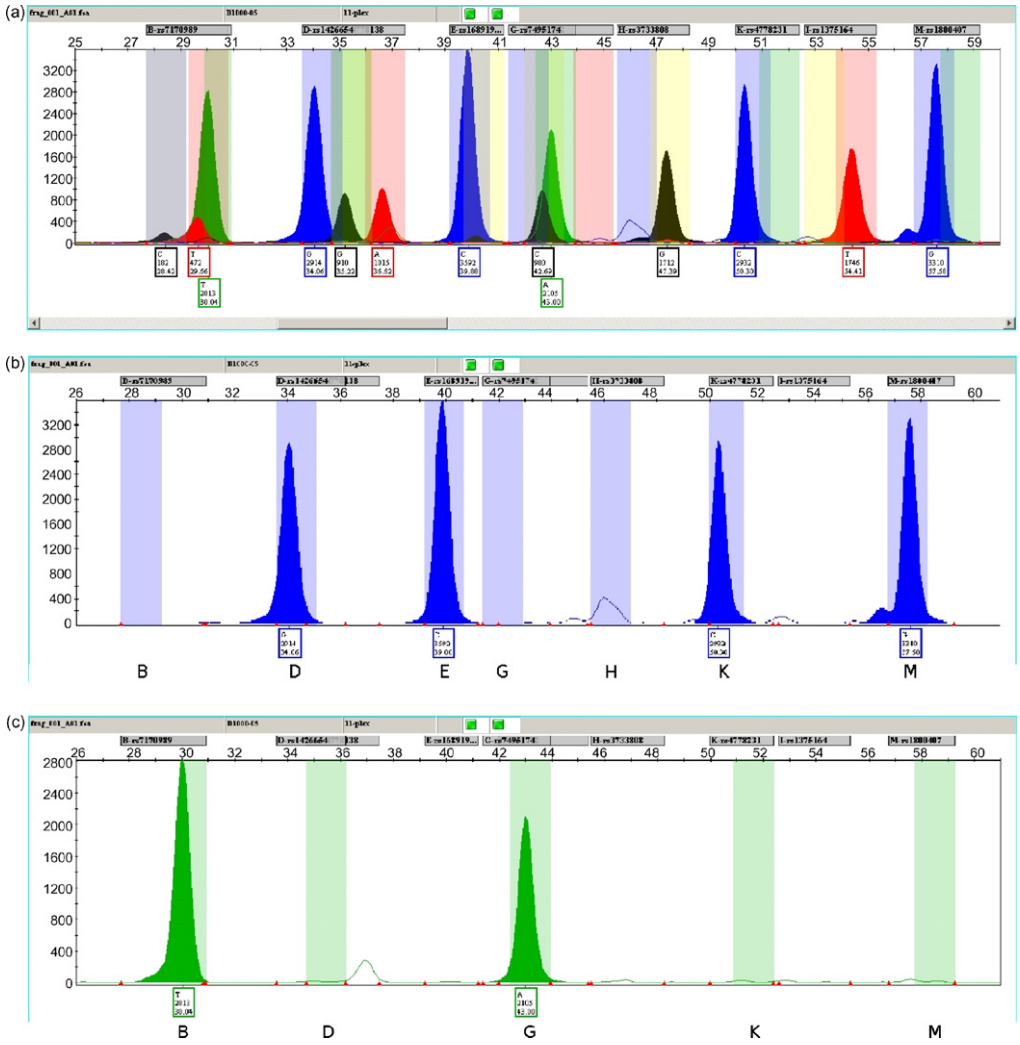


Fig. 1. Typical SNaPshot genotyping result for one sample, GeneMapper 4.0 diagram view. (a) Combined view; (b–e) single colour view with bins. Bins are labelled at the bottom according to the SNP lettering from Table 1. Allele calling algorithm in the GeneMapper 4.0 software are stringent enough to recognize artefact peaks inside a bin (see (b)).

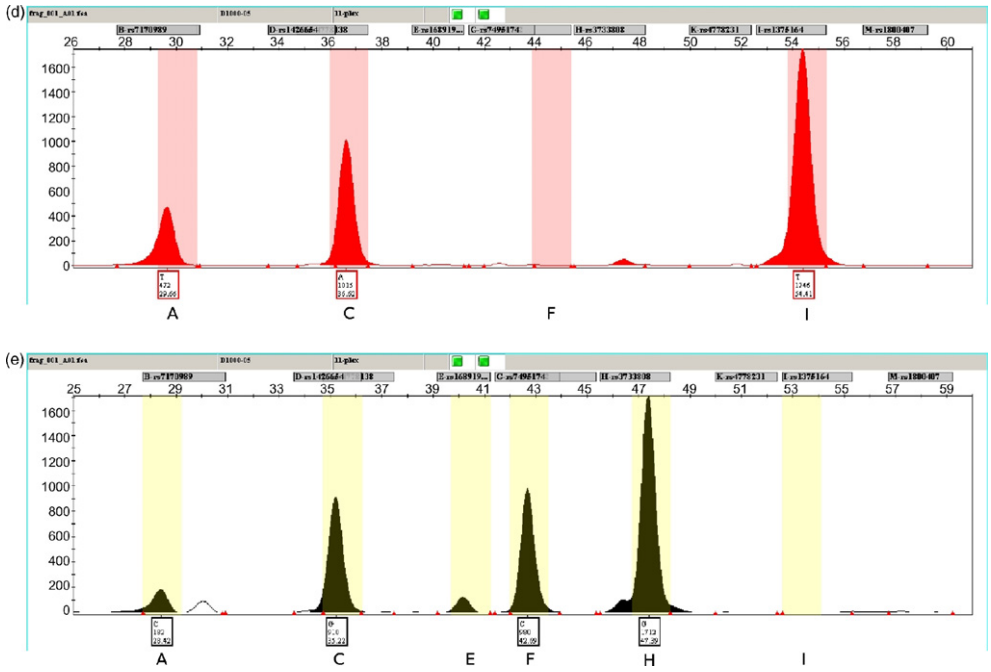


Fig. 1. (*Continued*).

HapMap and were in accordance. Allele frequencies for the studied populations are displayed in Fig. 2.

The SNPs were selected based on their relevance for human pigmentation. They were taken from genes associated with oculocutaneous albinism II (OCA2), and genes coding for solute carrier family 45, member 2, associated with oculocutaneous albinism type 4 (SLC45A2, MATP), and SLC24A5, another intracellular carrier associated with pigmentation. The two sample populations were chosen to have clear differences in skin pigmentation. However, only 5 of the 11 SNPs analysed were informative based on this admittedly very broad criterion: B, C, I (OCA2), as well as D (SLC24A5), and E (SLC45A2).

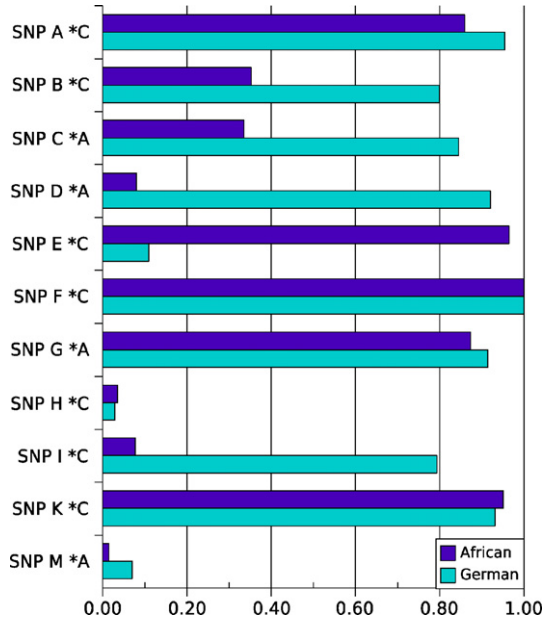


Fig. 2. Observed allele frequencies for the selected 11 SNPs in sub-Saharan Africans and Northern Europeans.

Thus these five SNPs can also be considered as ancestry informative, whereas the remaining six SNPs may have other effects which cannot be detected in the two groups studied here. Our study will be revised and extended to include more specific and informative markers, as well as probands with more specific pigmentation features.

Funding source

The research was funded in part by the Helmholz Stiftung. The funding source had no involvement in the development of the paper or decisions related to the paper.

Conflict of interest

None.

References

[1] D.L. Duffy, G.W. Montgomery, W. Chen, Z.Z. Zhao, L. Le, M.R. James, N.K. Hayward, N.G. Martin, R.A. Sturm, A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation, Am. J. Hum. Genet. 80 (2) (2007) 241–252.

[2] J. Graf, R. Hodgson, A. van Daal, Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation, Hum. Mutat. 25 (3) (2005) 278–284.

[3] K. Nakayama, S. Fukamachi, H. Kimura, Y. Koda, A. Soemantri, T. Ishida, Distinctive distribution of AIM1 polymorphism among major human populations with different skin color, J. Hum. Genet. 47 (2) (2002) 92–94.

[4] M. Soejima, Y. Koda, Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2, Int. J. Legal Med. 121 (1) (2007) 36–39.

[5] S. Rozen, H. Skaletsky, Primer3 on the WWW for general users and for biologist programmers, Methods Mol Biol. 132 (2000) 365–386.

[6] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (17) (1997) 3389–3402.

2.2 A 21 marker insertion deletion polymorphism panel to study biogeographic ancestry

Zusammenfassung: Insertions-/Deletions-Polymorphismen haben in letzter Zeit erhöhte Aufmerksamkeit im Bereich der Forensischen Genetik erhalten. Diese Marker-Klasse verbindet die vorteilhaften genetischen Eigenschaften der Einzelnukleotidpolymorphismen (d. h. niedrige Mutationrate, genetische Stabilität und kurze Amplicon-Länge) mit den technischen Vorteilen der Short Tandem Repeat Marker (einfache Detektion mittels fluoreszenz-markierter PCR und Kapillarelektrophorese). InDel Marker eignen sich für die Analyse der biogeographischen Herkunft, da signifikante Unterschiede in den Allelfrequenzen zwischen den großen Bevölkerungsgruppen für eine große Anzahl dieser Marker bekannt sind. Wir haben einen Multiplex-PCR-Assay zur Bestimmung der biogeographischen Herkunft von forensischen DNA-Proben basierend auf Insertions-/Deletions-Polymorphismen entwickelt. Ein Panel von 21 kurzen InDel-Polymorphismen mit bekannten Allelfrequenz-Unterschieden zwischen den drei großen kontinentalen Bevölkerungsgruppen (Europäer, Afrikaner und Asiaten) wurde in einer Multiplex-PCR-Reaktion zusammengefasst. Der Assay ist hochsensitiv und benötigt weniger als 0,5 ng genomische DNA für eine erfolgreiche Typisierung. Aufgrund der kurzen Fragmentlängen der PCR-Produkte von weniger als 200 Basenpaaren eignet sich der Assay hervorragend für die Analyse von schwierigen forensischen Proben. Eine Populationsgenetische Studie zeigt die Leistung des Assays in Bezug auf die Bestimmung der biogeographischen Herkunft von Personen. Die ausgewählten 21 Marker sind ausreichend, um zwischen den drei großen kontinentalen Bevölkerungsgruppen zu unterscheiden.

Citation: Zaumsegel, D., Rothschild, M. A., Schneider, P. M.,: A 21 marker insertion deletion polymorphism panel to study biogeographic ancestry. Forensic Science International: Genetics 7 (2013) 305–312 [138]
Reproduced with friendly permission of Elsevier Ltd., United Kingdom.

Forensic Science International: Genetics 7 (2013) 305–312

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

A 21 marker insertion deletion polymorphism panel to study biogeographic ancestry

Daniel Zaumsegel *, Markus A. Rothschild, Peter M. Schneider

Institute of Legal Medicine, Faculty of Medicine, University of Cologne, Melatengürtel 60/62, 50823 Köln, Germany

ARTICLE INFO

Article history:
Received 28 July 2012
Received in revised form 21 November 2012
Accepted 26 December 2012

Keywords:
DIP
InDel
Insertion–deletion polymorphism
AIM
Ancestry informative marker
Forensic genetics

ABSTRACT

Insertion/deletion polymorphisms have recently received increased interest in the forensic genetics community. This class of markers combines the advantageous genetic properties of single nucleotide polymorphisms (i.e., low mutation rate, genetic stability, and short amplicon size) with the technical advantage of short tandem repeat markers (simple detection by fluorescence-labelled PCR and capillary electrophoresis). For a large number of indel markers significant differences in allele frequencies between the major populations have been reported, making this class of markers suitable for the analysis of biogeographic ancestry. We have developed a multiplex PCR assay designed to establish the biogeographic ancestry of forensic DNA samples based on insertion/deletion polymorphisms. A panel of 21 short indels with allele frequency differences between three major population groups (European, African and Asian) was selected to be incorporated into a single-tube multiplex PCR assay. The assay is highly sensitive, requiring less than 0.5 ng of genomic DNA for successful typing. Due to the short fragment lengths below 200 bp, the assay is ideally suited for the typing of challenging forensic genetic case work samples. A population genetic study has been performed proving the performance of the assay in inferring the ancestral population of individuals. The chosen 21 markers are sufficient to distinguish between three major global population groups. Furthermore, the assay design leaves room for an extension in order to cover additional population groups.

© 2013 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Forensic genetic research has largely focused on the use of genetic polymorphisms for the identification of individuals in criminal casework. The method of choice is typing of short tandem repeat polymorphisms (STRs), due to their high power of discrimination and the availability of allele frequency data for a large variety of populations [1,2].

Recently, different forensic applications have moved into focus, such as the prediction of the biogeographic ancestry of an unknown stain donor. For this application, however, STRs are less useful due to their mutational instability and resulting high intra-population variability compared to relatively low inter-population variability [3]. Single nucleotide polymorphisms (SNPs) have emerged as the marker class of choice for this task [4,5].

SNPs combine a variety of characteristics required for the use as ancestry informative markers (AIMs), such as low mutation rate, high density of distribution throughout the genome and a full range of allele frequency patterns across populations [3], as well as robustness for typing highly degraded DNA samples [6,7]. However, being sequence polymorphisms in contrast to fragment length polymorphisms like the commonly used forensic STR markers, the use of SNPs poses some technical difficulties for the routine forensic laboratory due to the lack of adequate SNP typing equipment [5].

Bridging the gap between these established forensic methods, insertion/deletion polymorphisms (indels) have received increased attention during the last ten years. Indels are abundant in the genome with at least one indel every 7.2 kb according to the landmark study of Mills et al. [8]. Since indels derive from a single mutation event occurring with a low frequency, they are genetically quite stable [9] and may show significantly different allele frequency distributions between distant populations making them ideal candidates for ancestry informative markers [10,11].

The discovery of a large number of short indels by Mills et al., further improved by the follow-up study from 2011 [12], makes these markers especially interesting for forensic applications, since short indels can be analysed in the routine forensic laboratory employing standard techniques such as PCR with fluorochrome-labelled primers and capillary electrophoresis, and offer considerable multiplexing capabilities as well as potential for incorporation into automated high-throughput genotyping systems. In addition, short indels can greatly improve amplification success even with

* Corresponding author. Tel.: +49 221 478 88335/88222; fax: +49 221 478 88370.
E-mail address: daniel.zaumsegel@uk-koeln.de (D. Zaumsegel).

1872-4973/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2012.12.007

Table 1
Markers selected for inclusion in the AIM indel assay, including chromosomal position, location in the genome, reported alleles and fluorochrome label.

rs number	Chromosome	Position (bp)	Alleles	Expected amplicon length	Fluorochrome
rs4646006	1	15,717,609	–/CTCA	61–65	6-FAM
rs140864	1	36,391,662	–/TTC	91–94	HEX
rs140858[a]	1	96,836,326	–/CT	86–88	HEX
rs2308026	4	119,404,855	–/CA	90–92	6-FAM
rs33948716	4	123,994,263	–/CCT	158–161	6-FAM
rs1610963	5	112,274,982	–/ATAACTAA	163–171	6-FAM
rs3834371	8	130,940,151	–/GAGT	108–112	HEX
rs140847	9	12,617,325	–/GCTT	152–156	HEX
rs35906376	11	36,007,532	–/AGGACT	114–120	HEX
rs2307666	11	64,486,500	–/GTTAC	97–102	6-FAM
rs33972805	11	126,288,872	–/CT	126–128	6-FAM
rs2308171	13	43,778,155	–/TCTG	132–136	HEX
rs2308036	15	65,207,011	–/CC	98–100	HEX
rs3069460	16	88,362,823	–/AGTACTG	70–77	6-FAM
rs16711	17	20,023,011	–/TTTCTTCCTA	164–174	HEX
rs5828358	19	53,833,222	–/CAGA	67–71	HEX
rs11471448	20	17,363,020	–/GCA	129–132	6-FAM
rs34785121	20	57,744,778	–/TGGA	136–140	6-FAM
rs6481	22	34,031,900	–/GTGGA	147–152	6-FAM
rs34123598	22	35,599,490	–/ATCT	116–120	6-FAM
rs3218285	22	35,866,670	–/CAACCAT	80–87	6-FAM
rs4253631	22	44,933,760	–/TTT	144–147	HEX

[a] Removed from final marker panel.

highly degraded DNA samples often encountered in forensic casework. Indel markers have already been used in a variety of studies ranging from the analysis of genetic structure in human populations [13–15], inferring biogeographic ancestry [11,16], assessing individual admixture [17] and identification of individuals [18].

The study described here aims at the development of a robust multiplexed PCR assay designed to predict the biogeographic ancestry of forensic casework DNA samples based on insertion/deletion polymorphisms. The PCR-based multiplex typing assay contains 21 short indels with allele length variations between 2 and 10 bp and sufficient allele frequency differences between three major population groups predominantly relevant for forensic casework in Central Europe (European, sub-Saharan African and Asian). Assay design was keyed towards flexibility by only using two fluorescent labels, leaving room for further extension of the assay. Short fragment lengths below 200 bp for each marker improve amplification success in degraded samples, and the design as a single-tube reaction with high sensitivity allows for the successful analysis of routine forensic samples with low DNA content.

2. Material and methods

2.1. Marker selection

An initial set of candidate markers was selected from the available online database of the United States National Center for Biotechnology Information, dbSNP [19] and the Marshfield Clinic diallelic insertion/deletion database [20] based on the following criteria: (i) biallelic autosomal indels, (ii) non-coding, but in close proximity to genes to take advantage of selection effects, (iii) allele length variation 2–10 bp, and (iv) a large allele frequency difference in one of the major population groups compared to the other two with the main focus on Europe, Africa, and Asia. Where markers are located on the same chromosome, care was taken to select these as far apart from each other as possible to avoid loss of information resulting from linkage between these markers. Flanking sequences of the selected indels were checked for sequence variants and repeat structures likely to interfere with primer design or to disrupt analysis. In total, 22 indels with allele length variations between 2 and 10 bp were selected from an initial candidate list of 60 markers for incorporation into the assay (detailed information in Table 1). However, one of the selected markers (rs140858) posed considerable technical difficulties in typing as observed during the data validation process, and was subsequently removed from the final panel.

2.2. Primer and assay design

Primer design was performed using the Primer3 Plus Web Interface [21,22] aiming for an amplicon size between 50 and 200 bp, an optimum T_m of 60 ± 2°C and an optimum GC content of ≈50%.

During primer design, the markers were assigned for labelling with two fluorochromes (6-FAM and HEX) according to amplicon size with at least 4 bp gaps between neighbouring amplicons. For primer pairs presenting with strong −1 bp artefacts caused by incomplete adenylation of the PCR product, the reverse primer was extended by a 5′ tail of GTTTCTT to promote full adenylation [23]. Other primer pairs were extended with unspecific tails as closely related to the GTTTCTT tag sequence as possible where necessary for incorporation in the multiplex assay to ensure even spacing of the amplicons.

All primer pairs obtained were checked for unspecific binding using the Basic Local Alignment Search Tool of the NCBI [24,25] against the whole human genome. Primer pairs were also checked for hairpin and primer dimer formation using the AutoDimer software [26]. If not stated otherwise, all PCR amplifications were performed using the Qiagen Multiplex PCR kit (Qiagen, Hilden, Germany) in a total volume of 10 μl containing 5 μl of 2× Multiplex Master Mix, 1.5 μl of primer mix (details regarding the primer sequences and concentrations in the primer mix are available in Supplementary table 1) and 0.25–0.5 ng genomic DNA filled up to 10 μl with deionised water.

Thermocycling conditions were as follows: initial denaturation at 95 °C for 15 min, 28 cycles of 30 s at 95 °C, 90 s at 63 °C and 90 s at 72 °C followed by a final extension step of 60 min at 68 °C. Amplification products were purified by gel filtration using Sephadex™ G-50 (GE Healthcare, Munich, Germany) and subse-

quently prepared for capillary electrophoresis using the AB3130 Genetic Analyzer (Applied Biosystems, Darmstadt, Germany) by adding 1 μl of purified amplification product to 10 μl of a 100:1 mixture of HiDi Formamide and GeneScan 550 ROX internal size standard (Applied Biosystems). Amplification success was initially tested in singleplex reactions for all primer pairs with several commercially available control DNA samples. The multiplex assay was optimised to generate uniform peak heights within the dye channels as well as similar peak heights between the dye channels.

To facilitate automated allele calling with the GeneMapper ID version 3.2 software (Applied Biosystems), an allelic ladder containing both alleles for all 21 markers was constructed. Heterozygous samples for each marker were amplified in singleplex reactions, mixed to obtain uniform peak heights and the mixture was reamplified in a multiplex reaction to obtain a sufficient amount of a reasonably balanced allelic ladder.

Genotyping success for all markers was verified by direct sequencing of both alleles from homozygous samples using the BigDye Terminator v.1.1 Cycle Sequencing Kit (Applied Biosystems, Germany, data not shown).

2.3. Sensitivity study

In order to assess the performance of the assay in respect to reproducibility and sensitivity, two commercially available control DNA samples (9947a and 9948) were analysed in triplicate at total DNA amounts between 2 and 0.01 ng of DNA per PCR. Genotypes of the control DNA samples for all markers were confirmed in singleplex reactions. Correctly typed genotypes as well as missing alleles and allelic drop-ins for each sample and DNA quantity were recorded and compared over all analyses.

2.4. DNA degradation study

The performance of the assay in cases of low quality DNA was assessed by analysing artificially degraded DNA samples and comparing the observed indel profiles with the profiles obtained using a routine forensic STR kit. Commercially available control DNA (Quantifiler Human DNA Standard, Applied Biosystems) was degraded using an adopted protocol for digestion with micrococcal nuclease described by Freire-Aradas et al. [27]. 10 μg of DNA were incubated at 37 °C with 0.75 U of micrococcal nuclease (MNase, Fermentas GmbH, St. Leon-Rot, Germany) in a total volume of 180 μl, using the reaction buffer from [27]. Aliquots of 10 μl were taken at 5, 10, 15 and 30 min and the enzyme was inactivated by the addition of 6 μl 50 mM EDTA and incubation at 85 °C for 15 min. Degradation success was verified by electrophoresis in a 2% agarose gel. The degraded DNA was quantified spectrophotometrically using the NanoDrop 2000 Spectrophotometer (PEQLAB Biotechnologie GmbH, Erlangen, Germany). The degraded DNA was analysed with the indel 21-plex and the AmpFℓSTR SEfiler Plus STR kit (Applied Biosystems) according to standard protocols.

2.5. Population genetic study

In order to assess the informativeness of the final 21-plex AIM assay for predicting the biogeographic ancestry of DNA samples, an extensive population genetic study was conducted.

2.5.1. Sample selection

Samples were selected from previous population studies as well as from recent routine paternity cases. All samples were collected with informed consent of the donors. The samples were anonymised, except for the population of origin and, where applicable, the genetic relationship to other samples (e.g., sibling and parent–child). DNA of the samples had been previously extracted using different methods and had been stored at −20 °C. A total of 379 unrelated samples from three major continental population groups (Europe (n = 70) Sub-Saharan Africa (n = 69), and East and South East Asia (n = 164)) as well as some intermediate groups (Middle East (n = 52), Indo-Pakistan (n = 12) and Afghanistan (n = 12)) have been selected and genotyped for all 21 markers.

2.5.2. Statistical analysis

Allele frequencies, expected heterozygosities and F_{st} pairwise genetic distances for all pairs of populations were calculated, and analysis of molecular variance (AMOVA) and exact tests for Hardy–Weinberg Equilibrium as well as for linkage disequilibrium between all pairs of markers were performed on all unrelated samples genotyped for the full set of 21 markers using the ARLEQUIN 3.5 software [28]. All results were corrected for multiple testing. The same samples were assessed for group membership with the computer program STRUCTURE (version 2.3.3) [29–31], with a burn-in of 100,000 steps and a run time of an additional 500,000 steps in the Markov Chain. Populations were considered to be potentially admixed and allele frequencies of the markers were considered to be potentially correlated between populations. K values were chosen from 2 to 6 and each run was performed in five replicates. Replicates of the STRUCTURE runs were aligned with the program CLUMPP (version 1.1.2) to remove effects of label switching and to show potential multimodality in the data [32]. For visualisation of the obtained results the software DISTRUCT (version 1.1) was used [33]. The data obtained by the STRUCTURE analysis was plotted using STRUCTURE HARVESTER version 0.6.92 [34] to detect the true number of clusters in the analysed data. In order to assess the usefulness of the assay for a quick prediction of the population of origin, the SNIPPER App suite [35] developed by the University of Santiago de Compostela [5,36] was employed. In order to apply the software originally developed for the analysis of SNP data, the indel profiles obtained were coded as A/C SNPs with the insertion allele being designated "A" and the deletion allele being designated "C". A thorough analysis of the population data including cross-validation of all samples against the population panel as provided on the SNIPPER website [35] was performed.

3. Results and discussion

In this study we developed a robust and sensitive ancestry informative marker panel based of 21 indels, which can be successfully amplified in a single 21-plex PCR and analysed by standard capillary electrophoresis techniques.

3.1. Marker selection and assay design

As shown in Table 1, an initial set of 22 indels with allele length differences of at maximum 10 bp had been chosen for incorporation into the assay. The markers are spread over a large portion of the human genome, although some chromosomes (chromosomes, short: chr. 1, 11 and 22) are slightly overrepresented in the selection (see Table 1). Still, no significant linkage disequilibrium between any of the marker pairs was detected (data not shown). However, one of the 22 initially selected markers (rs140858) showed significant deviation from Hardy–Weinberg equilibrium (HWE) in the Asian population samples as well as strongly varying PCR performance between samples, while no deviation was discovered in the other populations. Further investigation of the marker rs140858 revealed a possible SNP site (rs1684829) 31 bp upstream of rs140858 which lies inside the primer binding site of the forward PCR primer for that marker. Direct sequencing of this potential SNP site to confirm the cause for HWE deviation was disrupted due to a further indel polymorphism (rs56207212) in the

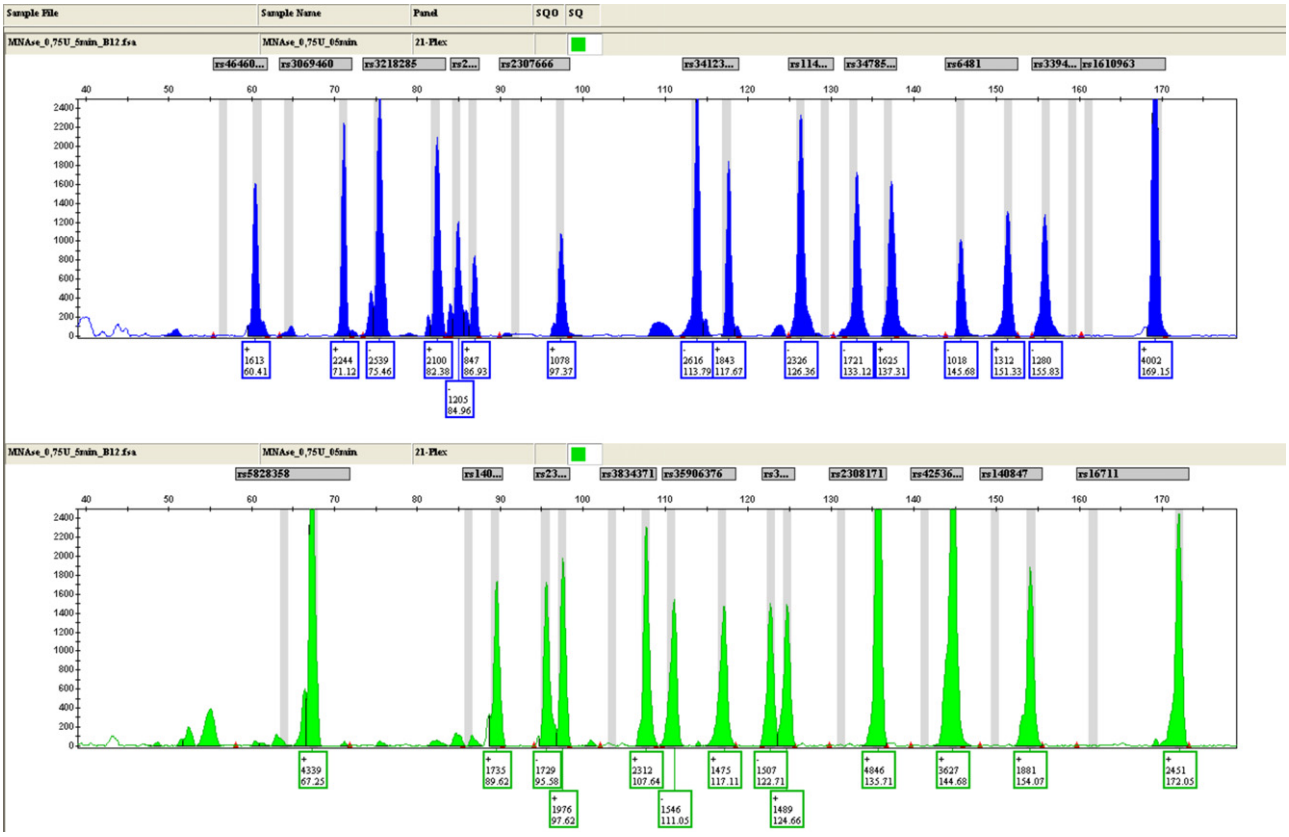


Fig. 1. Sample electropherogram of the 21-plex indel assay run with high quality DNA (Quantifiler Human Control DNA, Applied Biosystems) according to standard protocol.

direct proximity of the primer binding site. Due to these difficulties, the marker was removed from the panel prior to final analysis. No deviation from HWE was detected for the remaining markers after correction for multiple testing. During initial multiplex tests, several fragments presented with strong −1 bp peak artefacts as a result of incomplete adenylation of the PCR product. These artefacts could be largely prevented by the addition of an unspecific heptameric tail of GTTTCTT to the 5′ end of the reverse primer of the respective primer pairs as described in [23]. Since the addition of this tail did not pose any detrimental effects to PCR, the motif or part of the motif was also used for extension of the fragment length in cases where the initially selected primer pair resulted in a fragment length not suitable for incorporation into the multiplex due to overlap with neighbouring markers. In order to obtain a multiplex with balanced signal intensities as presented in Fig. 1, extensive tests were performed with varying primer concentrations and annealing temperatures leading to the primer mix concentrations presented in Supplementary table 1 and resulting in the current thermocycling conditions described above.

Since the shortest fragments in the multiplex run close to the primer peak in electrophoresis, several methods of PCR product cleanup were tried to reduce the primer peak and remove most of the PCR artefacts produced in the reaction (Qiagen MinElute columns, ExoI/SAP digestion, Sephadex G-50 filtration, data not shown). The Sephadex filtration has proven to be the most reliable cleanup method in this study. The allelic ladder contained both the insertion and deletion allele of every marker in the panel. It was not possible to achieve a complete balance between all markers, as some imbalances consistently reappeared after reamplification of the ladder master mix. The remaining imbalances however had no impact on the consistency of the automated allele calling performed by the GeneMapper ID Software (Aplied Biosystems, Darmstadt, Germany). The current assay design only using two dye channels leaves room for further extension of the multiplex. To allow for standardisation, the genotypes of all commercially available control DNA samples used throughout this study can be found in Supplementary table 2.

3.2. Sensitivity study

Analysis of a dilution series of two commercially available DNA samples of good quality was performed in triplicate to assess reproducibility and sensitivity of the assay. Best results were achieved with 0.25 ng of total DNA per reaction, although it was possible to obtain full profiles in the range between 0.1 and 0.5 ng of DNA (Fig. 2).

With lower amounts of DNA, it was still possible to amplify all markers, although results became increasingly inconsistent. Allelic imbalances between the markers as well as between the alleles of heterozygous markers and allelic dropout required the analysis of replicates and formulation of consensus profiles to recover lost data of individual PCR results. These artefacts are a consequence of stochastic effects in the PCR with a low number of template molecules [37,38]. With DNA amounts of 1 ng and above, strong signals leading to off-scale peaks and abnormally shaped or extra peaks due to pull-up of signals in other dye channels were observed as a consequence of inefficient matrix correction. Thus, the optimal amount of DNA is between 0.25 and 0.5 ng per assay. Due to its good sensitivity, the assay is ideally suited for the analysis of trace samples often encountered in forensic routine casework.

3.3. DNA degradation study

In order to assess the performance of the assay in cases of poor quality DNA, a degradation study was performed. Incubation of purified genomic DNA with micrococcal nuclease at 37 °C resulted

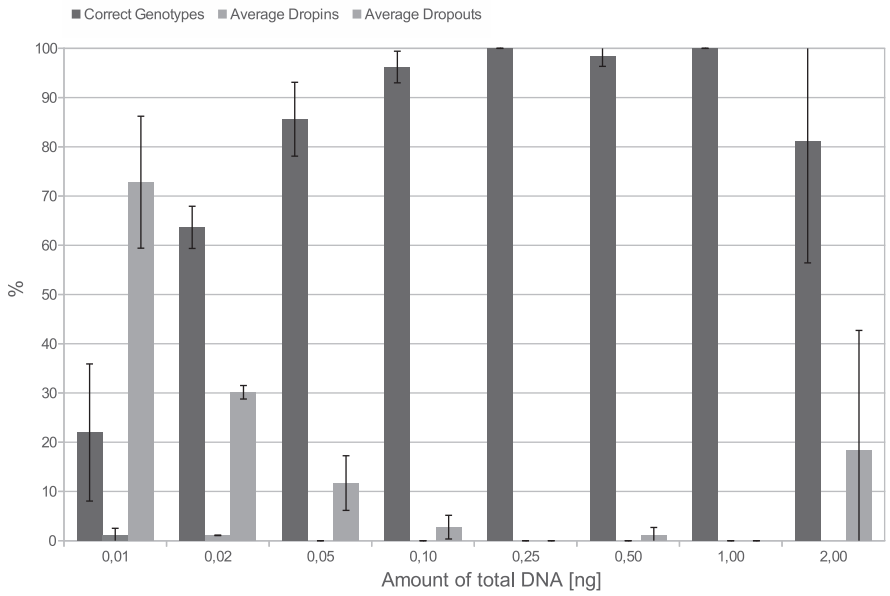


Fig. 2. Diagram of genotyping success: correctly typed genotypes as well as drop-ins and drop-outs are scored as average over two DNA samples, each analysed in triplicate.

in complete degradation of the genomic DNA into fragments <300 bp already after 5 min. Longer incubation times resulted in further degradation and smaller average fragments. After 30 min of incubation, the average fragment length visible on agarose gel was <200 bp (data not shown). The degraded DNA sampled at 5, 10, 15 and 30 min of degradation were analysed with the 21-plex indel assay and the SEfiler Plus STR kit (Applied Biosystems) as reference. With increasing degradation, the signal intensity of

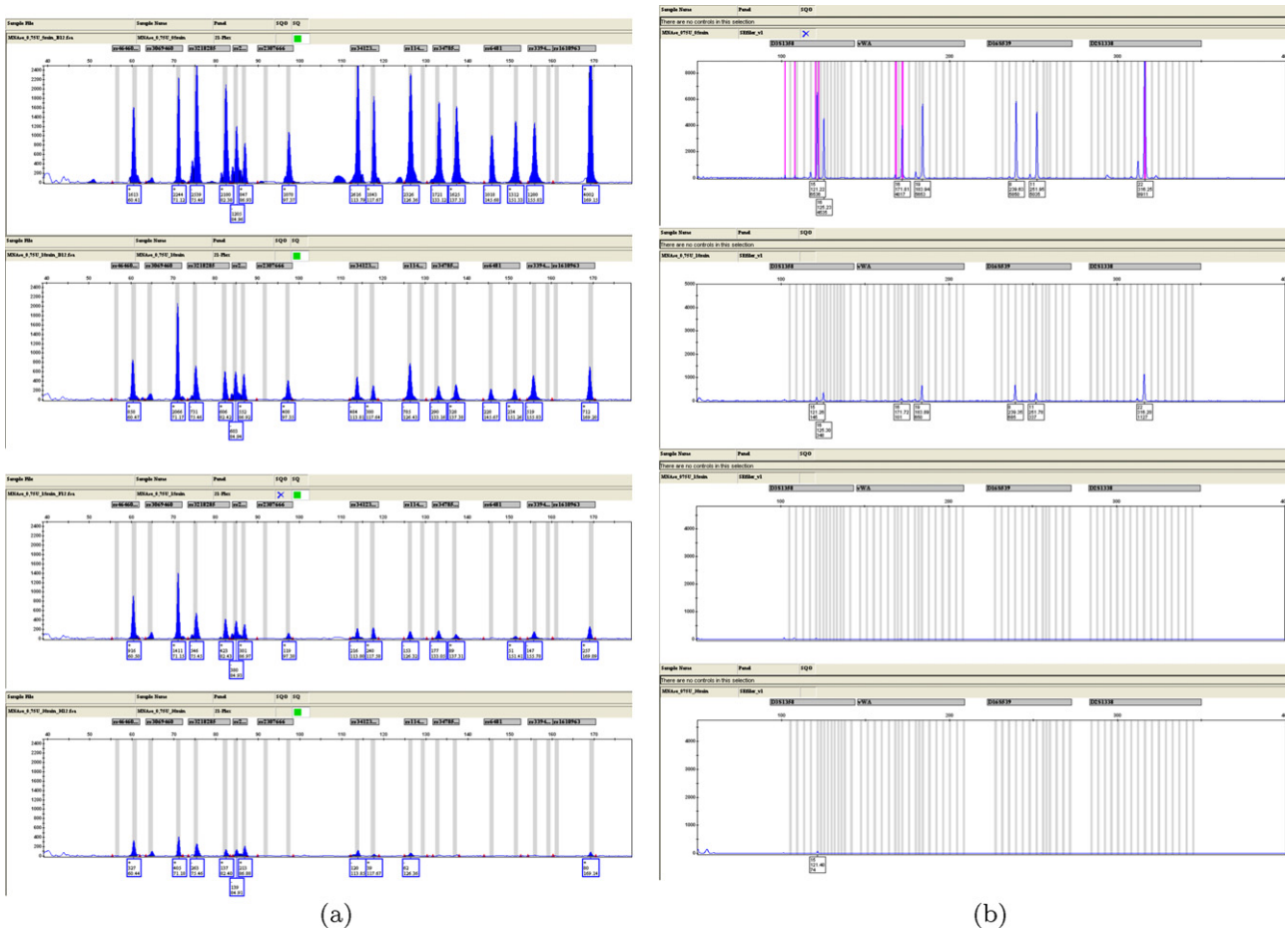


Fig. 3. Sample electropherograms obtained from the analysis of degraded DNA. (a) Depicts the blue dye channel of the 21-plex indel assay, (b) shows the corresponding SEfiler Plus profiles. Degradation times from top to bottom: 5, 10, 15 and 30 min.

Table 2
Pairwise population F_{st} estimates between the studied populations. Below diagonal: F_{st} values; above diagonal: corresponding p values (significance level: $p < 0.0024$ after correction for multiple testing) for 10,100 permutations.

	Africa	Europe	Middle East	Afghanistan	Indo-Pakistan	East Asia	South East Asia
Africa	–	$\leq 10^{-5}$	$\leq 10^{-5}$	$\leq 10^{-5}$	$\leq 10^{-5}$	$\leq 10^{-5}$	$\leq 10^{-5}$
Europe	0.28925	–	0.00822	0.00307	0.00218	$\leq 10^{-5}$	$\leq 10^{-5}$
Middle East	0.26634	0.00843	–	0.43768	0.57301	$\leq 10^{-5}$	$\leq 10^{-5}$
Afghanistan	0.29837	0.02861	0.00111	–	0.94713	$\leq 10^{-5}$	$\leq 10^{-5}$
Indo-Pakistan	0.28696	0.03136	−0.00144	−0.01804	–	$\leq 10^{-5}$	$\leq 10^{-5}$
East Asia	0.36943	0.23044	0.20196	0.18033	0.17475	–	0.00287
South East Asia	0.33029	0.19982	0.16981	0.13316	0.13204	0.00940	–

the 21-plex indel panel decreased equally over the full size range with only slightly stronger decrease for the longer amplicon lengths as shown in Fig. 3a.

Full profiles were detected for degradation times of up to 10 min, with only one of 29 alleles dropping out after 15 min of incubation. Even after 30 min, it was still possible to detect 19 of the 29 possible alleles, albeit at very low signal intensities. For the SEfiler Plus, however, profile quality rapidly deteriorated, already showing strong imbalances and PCR artefacts after 10 min of degradation (see Fig. 3b). The decrease in signal intensity is more pronounced for the long amplicon sizes as expected [6,7]. After 30 min of degradation, only fragments shorter than 200 bp can be detected at very low signal intensities, consistent with the average fragment size of the degraded DNA.

3.4. Population genetic analyses

A total of 379 individuals initially grouped in seven population groups (Africa, Europe, Middle East, Indo-Pakistan, Afghanistan, South-East Asia and East Asia) have successfully been typed for all 21 markers. The allele frequencies obtained for each marker for these regional groups are presented in Supplementary table 3. Analysis of pairwise genetic differences between these population groups revealed highly significant differences (significance level at $p < 0.0024$ after correction for multiple testing) between the three major population groups (African, European and Asian) consistent with previously published data (e.g., Yang et al. [11]). The difference between the two Asian subpopulations (South East Asian and East Asian) proved to be too small to be relevant ($F_{st} < 0.01$, see Table 2). The South East Asian and East Asian subpopulations were considered as one Asian population in further analyses. The pairwise comparisons involving the intermediate population groups from Afghanistan, Indo-Pakistan and the Middle East show very low F_{st} values, suggesting strong similarities between these populations. However, the small sample size of the Indo-Pakistani and the Afghan population samples may have prevented the observation of more informative differences.

AMOVA analysis at the population level already indicates about 20% of the genetic diversity found in the sample is represented by the variation between populations. Consistent with the above

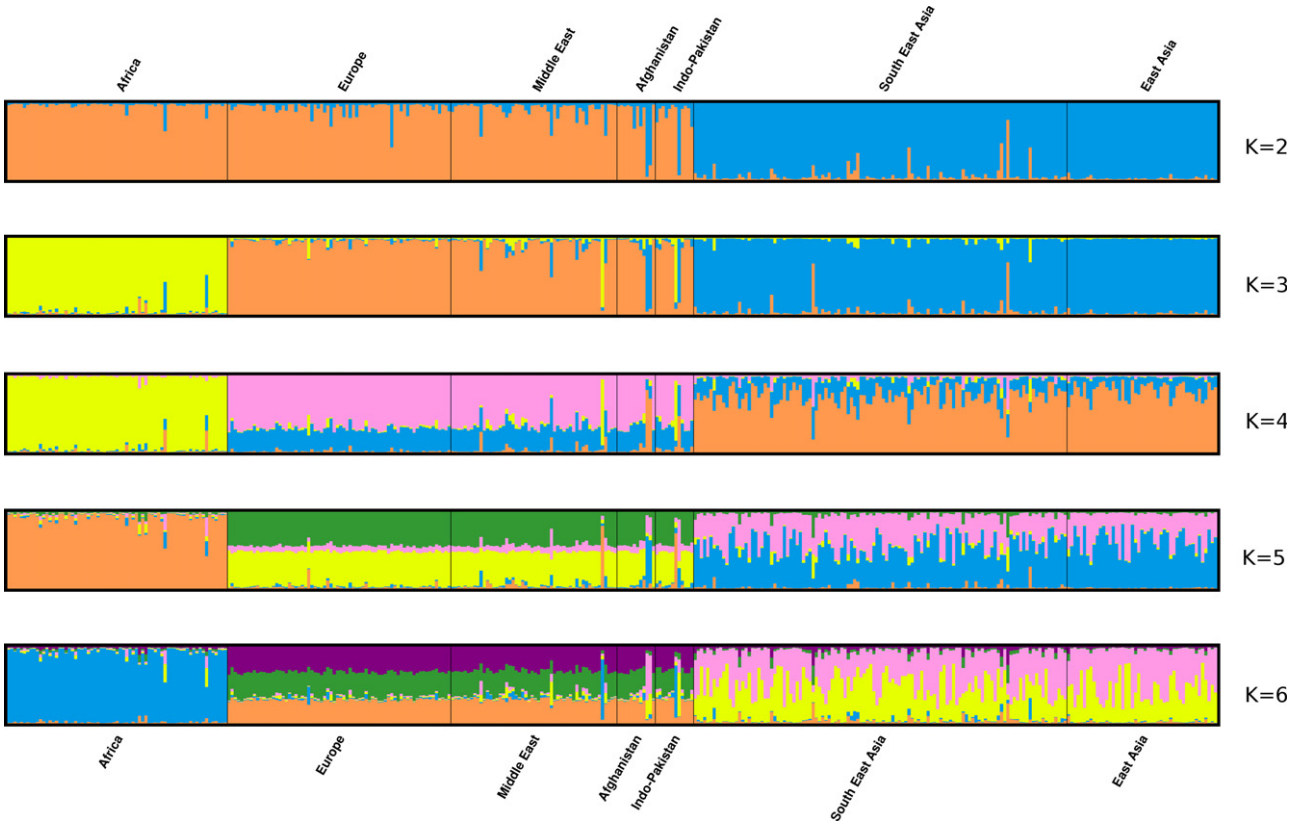


Fig. 4. Ancestral membership proportions of the studied populations based on five independent STRUCTURE runs treated with CLUMPP and plotted with DISTRUCT. Each vertical bar represents one individual and the colours represent the individual admixture proportions based on K assumed clusters (parental populations).

observations of F_{st} pairwise genetic difference measures, this fraction is maximised when considering a genetic structure with the populations grouped into three large geographic regions (Africa, West Eurasia consisting of Europe, Middle East, Indo-Pakistan and Afghanistan, and East Eurasia with East and South East Asia) resulting in a genetic variation between regions of approximately 24%. Only about 1% of the total variation is explained by the variation between populations within each group in the latter model.

In order to assess the predictive value of the marker set, cluster analysis was performed using the STRUCTURE program version 2.3.3 [29–31]. The algorithm used for the analysis was selected to account for possible admixture in the populations by considering possible linkage disequilibrium due to admixture [39]. Since several of the studied populations were found to be closely related (e.g., East Asian and South East Asian), a model considering allele frequencies to be correlated between populations was chosen to detect subtle population subdivisions [30]. Fig. 4 shows the STRUCTURE bar plots for $K = 2-6$ plotted with DISTRUCT [33] after aligning the five replicate analyses with the CLUMPP program [32].

Consistent with the results from the pairwise genetic distance analysis, the STRUCTURE algorithm established a clear distinction between the Asian population (blue ancestry proportion at $K = 2$ and $K = 3$ in Fig. 4) and all other population groups (orange ancestry proportion at $K = 2$ and $K = 3$ in Fig. 4) already at $K = 2$. Further separation into the three main continental population groups occurs at $K = 3$, with the African samples producing an additional distinct cluster (yellow) while the intermediate populations (Middle East, Indo-Pakistan and Afghanistan) remain clustered together with the European samples. No further substructure is detected for higher values of K, consistent with the results of the AMOVA analysis. The estimated ln probability of data ($-\ln P(D)$) used as an *ad hoc* predictor for the most probable number of clusters present in the data [29], as well as the ΔK statistic according to Evanno et al. [40] maximises at $K = 3$ (data not shown). Both methods for inferring the correct number of clusters suggested $K = 3$ as the most probable number of clusters present in the data. The clustering of the populations studied is overall consistent with the place of birth of the individuals in each population. Some individuals seem to be misclassified however. Especially the Afghan and Indo-Pakistani populations were found to be less homogenous than the other populations. While the majority of these populations appear to be part of the Western Eurasian cluster, several individuals more likely belong to the Eastern Eurasian cluster. This may be explained by the considerable level of admixture present in the region (e.g., [41]), and by possible errors in assigning individuals to populations since the only available criterion was the place of birth of each sample donor. Although the classification of individuals with the STRUCTURE algorithm is possible, it cannot be recommended for the classification of individual unknown DNA samples as it is very time consuming. Therefore, a different approach has been implemented by Phillips et al. [5] with the likelihood-based approach used in the SNIPPER App suite [35]. With this tool, a single, unknown genetic profile can be compared to a set of reference populations, the "training set". The software calculates individual maximum likelihood estimates for the inclusion of the unknown sample into each reference population. A cross-validation has been performed using the option "Perform a verbose cross-validation analysis of my population data with the best 21 SNPs." in the "Thorough analysis of population data with an Excel file of populations" program on the SNIPPER website. Each sample was tested in turn as unknown sample against the training set containing all remaining samples. For this analysis, the small population samples from Afghanistan and Indo-Pakistan were omitted. The obtained results were in agreement with the clustering obtained using the STRUCTURE classification algorithm. As presented in Table 3, over 97% of the samples with African origin were classified as African by the algorithm, with only about 3% of the samples having been misclassified as Eastern Asian. Similar success rates were obtained for the East Asian samples with only 3% of these samples having been misclassified as belonging to the Middle East.

Table 3
Estimated classification success for all tested individuals during cross-validation with the SNIPPER App suite. The percentage of all samples of a population of origin being classified as belonging to the population in each column.

	Africa	Middle East	East Asia	Europe
African origin	97.10%	0.00%	2.90%	0.00%
Middle Eastern origin	2.53%	77.22%	5.06%	15.19%
East Asian origin	0.00%	3.01%	96.99%	0.00%
European origin	0.00%	58.33%	0.00%	41.67%

The samples with European and Middle Eastern origin however were more difficult to classify. Consistent with the ancestry proportions identified in these groups in the STRUCTURE analysis, as well as with the results from pairwise F_{st} estimates (Table 2), a considerable proportion of these samples was misclassified by the algorithm. The Middle Eastern individuals showed strong similarities to the European individuals in the STRUCTURE analysis and consequently about 15% of these were classified as belonging to the European population by the SNIPPER algorithm. This similarity between the Middle East and Europe manifested even more in the European samples, which were classified as belonging to the Middle East in over 58% of the cases. Taken together, these data clearly demonstrate that a reliable prediction of European vs. Middle Eastern ancestry cannot be achieved with this marker set.

Overall, the population assignments obtained with both the STRUCTURE algorithm and the SNIPPER program were in good correlation with each other as well as with worldwide population structure as described in previous publications (e.g., [5,16,42]) using different marker sets. In contrast to these studies, informative prediction of biogeographic ancestry was achieved for three major populations using only 21 indel markers.

4. Conclusion

We have presented a robust and easy to use multiplex PCR assay to assess biogeographic ancestry of challenging samples based on 21 short insertion/deletion polymorphisms. Markers have been selected to be amplified in a single-tube multiplex reaction and analysed by standard capillary electrophoresis and fluorescent detection. The assay was proven to be highly sensitive, needing less than 0.5 ng of DNA for successful typing of all markers. The short amplicon length of less than 200 bp makes the assay suitable for the analysis of degraded DNA even at degradation levels where conventional STR assays do not provide any information anymore. We could demonstrate that a set of only 21 carefully selected biallelic indel markers is sufficient to characterise three major population groups on a global level (Africa, Europe and Asia). A larger marker set with specifically selected markers will be necessary to further distinguish intermediate populations such as the Middle East. Possible extensions to the assay might also include the addition of a set of markers selected for individual identification purposes in order to make the assay more versatile in the forensic genetics field.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2012.12.007.

References

[1] R. Chakraborty, D.N. Stivers, B. Su, Y. Zhong, B. Budowle, The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems, Electrophoresis 20 (8) (1999) 1682–1696.

[2] J.M. Butler, Forensic DNA Typing, Elsevier Academic Press, Oxford, UK, 2005.

[3] C. Phillips, R. Fang, D. Ballard, M. Fondevila, C. Harrison, F. Hyland, E. Musgrave-Brown, C. Proff, E. Ramos-Luis, B. Sobrino, A. Carracedo, M. Furtado, D.S. Court, P. Schneider, Evaluation of the genplex SNP typing system and a 49 plex forensic marker panel, Forensic Sci. Int. Genet. 1 (2) (2007) 180–185. http://www.sciencedirect.com/science/article/pii/S1872497307000610.

[4] T. Frudakis, K. Venkateswarlu, M.J. Thomas, Z. Gaskin, S. Ginjupalli, S. Gunturi, V. Ponnuswamy, S. Natarajan, P.K. Nachimuthu, A classifier for the SNP-based inference of ancestry, J. Forensic Sci. 48 (4) (2003) 771–782.

[5] C. Phillips, A. Salas, J.J. Sánchez, M. Fondevila, A. Gómez-Tato, J. Alvarez-Dios, M. Calaza, M.C. de Cal, D. Ballard, M.V. Lareu, A. Carracedo, SNPforID consortium, inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, Forensic Sci. Int. Genet. 1 (3/4) (2007) 273–280.

[6] M. Fondevila, C. Phillips, N. Naveran, L. Fernandez, M. Cerezo, A. Salas, A. Carracedo, M.V. Lareu, Case report: identification of skeletal remains using short-amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur, Forensic Sci. Int. Genet. 2 (3) (2008) 212–218.. http://dx.doi.org/10.1016/j.fsigen.2008.02.005.

[7] M. Fondevila, C. Phillips, N. Naveran, M. Cerezo, A. Rodriguez, R. Calvo, L.M. Fernandez, A. Carracedo, M.V. Lareu, Challenging DNA: assessment of a range of genotyping approaches for highly degraded forensic samples, Forensic Sci. Int. Genet. Suppl. Ser. 1 (1) (2008) 26–28.

[8] R.E. Mills, C.T. Luttig, C.E. Larkins, A. Beauchamp, C. Tsui, W.S. Pittard, S.E. Devine, An initial map of insertion and deletion (INDEL) variation in the human genome, Genome Res. 16 (9) (2006) 1182–1190. http://dx.doi.org/10.1101/gr.4565806.

[9] M.W. Nachman, S.L. Crowell, Estimate of the mutation rate per nucleotide in humans, Genetics 156 (1) (2000) 297–304. http://www.genetics.org/content/156/1/297.abstract.

[10] J.L. Weber, D. David, J. Heil, Y. Fan, C. Zhao, G. Marth, Human diallelic insertion/deletion polymorphisms, Am. J. Hum. Genet. 71 (4) (2002) 854–862. http://dx.doi.org/10.1086/342727.

[11] N. Yang, H. Li, L.A. Criswell, P.K. Gregersen, M.E. Alarcon-Riquelme, R. Kittles, R. Shigeta, G. Silva, P.I. Patel, J.W. Belmont, M.F. Seldin, Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine, Hum. Genet. 118 (3/4) (2005) 382–392. http://dx.doi.org/10.1007/s00439-005-0012-1.

[12] R.E. Mills, W.S. Pittard, J.M. Mullaney, U. Farooq, T.H. Creasy, A.A. Mahurkar, D.M. Kemeza, D.S. Strassler, C.P. Ponting, C. Webber, S.E. Devine, Natural genetic variation caused by small insertions and deletions in the human genome, Genome Res. 21 (6) (2011) 830–839. http://dx.doi.org/10.1101/gr.115907.110.

[13] N.A. Rosenberg, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, M.W. Feldman, Genetic structure of human populations, Science 298 (5602) (2002) 2381–2385. http://dx.doi.org/10.1126/science.1078311.

[14] N.A. Rosenberg, S. Mahajan, S. Ramachandran, C. Zhao, J.K. Pritchard, M.W. Feldman, Clines, clusters, and the effect of study design on the inference of human population structure, PLoS Genet. 1 (6) (2005) e70.

[15] L. Bastos-Rodrigues, J.R. Pimenta, S.D.J. Pena, The genetic structure of human populations studied through short insertion–deletion polymorphisms, Ann. Hum. Genet. 70 (Pt 5) (2006) 658–665. http://dx.doi.org/10.1111/j.1469-1809.2006.00287.x.

[16] R. Pereira, C. Phillips, N. Pinto, C. Santos, S.E.B. dos Santos, A. Amorim, Carracedo Ángel, L. Gusmão, Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing, PLoS ONE 7 (1) (2012) e29684 http://dx.doi.org/10.1371/journal.pone.0029684.

[17] N.P.C. Santos, E.M. Ribeiro-Rodrigues, A.K.C. Ribeiro-Dos-Santos, R. Pereira, L. Gusmão, A. Amorim, J.F. Guerreiro, M.A. Zago, C. Matte, M.H. Hutz, S.E.B. Santos, Assessing individual interethnic admixture and population substructure using a 48-insertion–deletion (INSEL) ancestry-informative marker (AIM) panel, Hum. Mutat. 31 (2) (2010) 184–190. http://dx.doi.org/10.1002/humu.21159.

[18] R. Pereira, C. Phillips, C. Alves, A. Amorim, A. Carracedo, L. Gusmão, A new multiplex for human identification using insertion/deletion polymorphisms, Electrophoresis 30 (21) (2009) 3682–3690. http://dx.doi.org/10.1002/elps.200900274.

[19] dbSNP, Short Genetic Variations, http://www.ncbi.nlm.nih.gov/snp/ (last visited: 19.08.11).

[20] Diallelic Insertion/Deletion Polymorphism Database, http://www.marshfieldclinic.org/mgs/ (last visited 19.08.11).

[21] Primer3 Plus Web Interface, http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi (last visited 19.08.11).

[22] A. Untergasser, H. Nijveen, X. Rao, T. Bisseling, R. Geurts, J.A. Leunissen, Primer3-Plus, an enhanced web interface to Primer3, Nucleic Acids Res. 35 (Suppl. 2) (2007) W71–W74.

[23] M.J. Brownstein, J.D. Carpten, J.R. Smith, Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping, BioTechniques 20 (6) (1996) 1004-6–1008-10.

[24] Basic Local Alignment Search Tool, http://blast.ncbi.nlm.nih.gov/Blast.cgi (last visited: 19.08.2011).

[25] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (3) (1990) 403–410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2.

[26] P.M. Vallone, J.M. Butler, AutoDimer: a screening tool for primer–dimer and hairpin structures, BioTechniques 37 (2) (2004) 226–231.

[27] A. Freire-Aradas, M. Fondevila, A.-K. Kriegel, C. Phillips, P. Gill, L. Prieto, P.M. Schneider, A. Carracedo, M.V. Lareu, A new SNP assay for identification of highly degraded human DNA, Forensic Sci. Int. Genet. 6 (3) (2012) 341–349. http://dx.doi.org/10.1016/j.fsigen.2011.07.010.

[28] L. Excoffier, H.E.L. Lischer, 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. http://dx.doi.org/10.1111/j.1755-0998.2010.02847.x.

[29] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2) (2000) 945–959.

[30] D. Falush, M. Stephens, J.K. Pritchard, Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies, Genetics 164 (4) (2003) 1567–1587.

[31] D. Falush, M. Stephens, J.K. Pritchard, Inference of population structure using multilocus genotype data: dominant markers and null alleles, Mol. Ecol. Notes 7 (4) (2007) 574–578.. http://dx.doi.org/10.1111/j.1471-8286.2007.01758.x.

[32] M. Jakobsson, N.A. Rosenberg, CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure, Bioinformatics 23 (14) (2007) 1801–1806. http://dx.doi.org/10.1093/bioinformatics/btm233.

[33] N.A. Rosenberg, distruct: a program for the graphical display of population structure, Mol. Ecol. Notes 4 (1) (2004) 137–138. http://dx.doi.org/10.1046/j.1471-8286.2003.00566.x.

[34] D. Earl, B. vonHoldt, STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method, Conserv. Genet. Resour. 4 (2012) 359–361. http://dx.doi.org/10.1007/s12686-011-9548-7.

[35] The SNIPPER App Suite, http://mathgene.usc.es/snipper/.

[36] C. Phillips, L. Prieto, M. Fondevila, A. Salas, A. Gómez-Tato, J. Alvarez-Dios, A. Alonso, A. Blanco-Verea, M. Brión, M. Montesino, A. Carracedo, M.V. Lareu, Ancestry analysis in the 11-M Madrid bomb attack investigation, PLoS ONE 4 (8) (2009) e6583 http://dx.doi.org/10.1371/journal.pone.0006583.

[37] P. Gill, J. Whitaker, C. Flaxman, N. Brown, J. Buckleton, An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA, Forensic Sci. Int. 112 (1) (2000) 17–40.

[38] M.A. Jobling, P. Gill, Encoded evidence: DNA in forensic analysis, Nat. Rev. Genet. 5 (10) (2004) 739–751. http://dx.doi.org/10.1038/nrg1455.

[39] J.C. Stephens, D. Briscoe, S.J. O'Brien, Mapping by admixture linkage disequilibrium in human populations: limits and guidelines, Am. J. Hum. Genet. 55 (4) (1994) 809–824.

[40] G. Evanno, S. Regnaut, J. Goudet, Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study, Mol. Ecol. 14 (8) (2005) 2611–2620. http://dx.doi.org/10.1111/j.1365-294X.2005.02553.x.

[41] Indian Genome Variation Consortium, The Indian Genome Variation database (IGVdb): a project overview, Hum. Genet. 118 (1) (2005) 1–11.

[42] H.K. el Khil, K. Fadhlaoui-Zid, L. Cherni, C. Phillips, M. Fondevila, A. Carracedo, A.B. Ammar-Elgaaied, Genetic analysis of the SNPforID 34-plex ancestry informative SNP panel in Tunisian and Libyan populations, Forensic Sci. Int. Genet. 5 (3) (2011) e45–e47. http://dx.doi.org/10.1016/j.fsigen.2010.07.007.

Supplementary Table 1: Composition of the 21-plex Primer Mix including primer sequences. Unspecific tail sequences are underlined. Concentrations refer to each Primer in a pair in the multiplex primer mix.

rs number	Concentration [μM]	Primer Sequence
rs4646006	2.0	F: GCAAAGGCTGGTAAATGGCACACA R: GTTTCTTAAGTGGGCAGCAATGGAGCTGC
rs3069460	0.15	F: GGCCACACCAGACAGTCCCC R: GGCTGTGTGCCTCTCATGCCAA
rs3218285	0.5	F: GCTCCAAGGACACCCGGACC R: GTTTCTTCATCTTCACACTCTCCTGCCCCTC
rs5828358	0.25	F: CAGCGACCATGGGGGACACC R: GTTTCTTCTATACTGACGCCCCCTGCCGT
rs4253631	1.0	F: CATGTTGGCCAGGCTGGGTCT R: GTTTTTGTGTTAATTTCATCCCCAATGCACAGA
rs34123598	0.3	F: GGAGGTGAGGTTGCTTCAGGCATT R: GCAGGCTCCCCAACCTGCCT
rs11471448	0.15	F: ACGCTGAGCCTCTGTCTTGGC R: TCAGGGACCAGGCACCCGAC
rs6481	0.8	F: GAAGGCAGAAGCCGGGGTGG R: GTTTCTGCAGAGACCATCTGGTTCACGAAG
rs33948716	0.4	F: AGGAAGGAGGGAAAAGGGGAATCAGG R: GTTTCTGACCACGTGGGTATGGATGCGTAAC
rs1610963	1.8	F: ATGGAAGGTTGTCCCTTTCC R: AGGTGTGCATCATCACCAAA
rs2308026	1.0	F: CCAAAGGGACCTGGCAGCTGG R: GTTTCTGTCTAGGCTCCTTTCATGGCCCG
rs33972805	1.0	F: ACTGAGCTAGTCCAGAGCCATCTATCATTC R: TCCACCAGCTCCATGCTTAGTAAGAAGA
rs140864	0.4	F: CAAAATCTGCTCCATGTCCAATCTGC R: GTTTCTGCCAGCCAGCCCCATGCTTC
rs34785121	0.5	F: GAGGTGCAGTGGAACCTGGCAC R: GTTTCTTCTCCCCGTCTCTCCCTGCAGTT
rs3834371	0.75	F: AGGGACTGCCTCCAAGGGTCA R: GTTTCTGTGGGCATTTGGGGCCTTTAAAATAC
rs35906376	0.5	F: TGAGGGCTGTTAAAGGGAGCAGG R: GGGGGTCTGACAGCAACTCTGGA
rs2308036	1.0	F: GATGGCAGCATGGGGCTCCG R: GTTTCTTACACCTGTGGGCTCAGGGTCA
rs2308171	0.5	F: ACCACAGCCCCTGGAAAGAGGT R: CGTTTCTGGCCGCTTCTGGA
rs140847	1.0	F: ACCAGGATAGCATTCAACAGTTTGAGGG R: GTTTCTGGTAGGCATGTAGAAATACGGTTCCAAAC
rs16711	0.6	F: CATCCGGCCTGGGCCAATTCT R: GTTTCTTACTTGCATGCCACAGAAGCTGA
rs2307666	0.3	F: GTGGTCACCTAAAAATGCGTGGAGAGT R: GTTTCTTTGTGGCCATGGTGATATTACGTCCC

Supplementary Table 2: Reference Genotypes of commercially available control DNA samples typed with the 21-plex indel panel. HCD = Quantifiler Human Control DNA; I: homozygous insertion; D: homozygous deletion; H: heterozygous genotype

DIP	HCD	9947A	9948
rs4646006	I	H	H
rs3069460	I	H	H
rs3218285	H	H	H
yrs5828358	I	H	H
rs4253631	I	H	D
rs34123598	H	I	I
rs11471448	D	H	I
rs6481	H	H	I
rs33948716	D	I	I
rs1610963	I	D	H
rs2308026	H	H	I
rs33972805	H	H	I
rs140864	I	I	I
rs34785121	H	D	D
rs3834371	I	D	I
rs35906376	H	H	H
rs2308036	H	D	D
rs2308171	I	I	H
rs140847	I	I	I
rs16711	I	I	H
rs2307666	I	I	D

Table 2.1: Print version of electronic supplementary Table 3: Population frequencies of the insertion allele

(n=)	Africa 69	Europe 70	Afghanistan 12	Middle East 52	Indo-Pakistan 12	South East Asia 117	East Asia 47
rs11471448	0.768	0.707	0.625	0.702	0.583	0.440	0.457
rs140847	0.761	0.814	0.792	0.788	0.625	0.355	0.351
rs140864	0.877	0.993	0.917	0.962	0.917	0.380	0.245
rs1610963	0.935	0.400	0.250	0.375	0.458	0.346	0.287
rs16711	0.978	0.571	0.667	0.587	0.750	0.970	0.989
rs2307666	0.812	0.329	0.667	0.529	0.667	0.786	0.755
rs2308026	0.920	0.593	0.583	0.538	0.583	0.731	0.798
rs2308036	0.348	0.157	0.250	0.231	0.292	0.667	0.809
rs2308171	0.391	0.736	0.875	0.788	0.833	0.923	0.979
rs3069460	0.536	0.643	0.625	0.683	0.542	0.564	0.404
rs3218285	0.942	0.450	0.375	0.413	0.333	0.504	0.553
rs33948716	0.072	0.714	0.667	0.683	0.750	0.936	0.979
rs33972805	0.696	0.479	0.583	0.606	0.542	0.872	0.883
rs34123598	0.580	0.914	0.917	0.827	0.875	1.000	1.000
rs34785121	0.341	0.000	0.000	0.019	0.000	0.000	0.000
rs35906376	0.826	0.450	0.750	0.471	0.625	0.594	0.479
rs3834371	0.754	0.464	0.625	0.558	0.625	0.474	0.489
rs4253631	0.761	0.314	0.042	0.279	0.208	0.115	0.074
rs4646006	0.790	0.550	0.417	0.385	0.208	0.645	0.702
rs5828358	0.790	0.286	0.417	0.327	0.333	0.799	0.691
rs6481	0.080	0.693	0.625	0.558	0.583	0.620	0.543

3 Discussion

Currently applied forensic genetic typing strategies are generally based on the comparison of one DNA profile with another one, be it crime scene samples with reference profiles of suspects or remains of deceased persons with profiles obtained from personal belongings or living relatives of missing persons. If a match between the DNA profiles in question can be observed, the probability of identity or relationship can be calculated using established biostatistical methods. In paternity and kinship analysis the transmission probabilities of the investigated genetic loci from one generation to the next as well as the observed occurrence of their alleles in the reference population is used to make statements about the probability of the postulated relatedness of the individuals in question. In criminal practice however it is not uncommon that a suspect is not known and searches in DNA databases also do not provide a match. If other sources of evidence do not provide investigative leads, such cases can remain unsolved although good quality DNA evidence has been found on the scene. Extracting additional information from the DNA sample, which may enable criminal investigators to limit the number of suspects in such cases would be a valuable addition to the forensic genetic "toolbox" of methods.

Recent advances in genetics including high-throughput genotyping techniques and methods for genome wide association studies have led to the identification of an ever increasing panel of genetic markers linked to complex genetic traits. In the forensic context, complex phenotypic traits which would allow the investigator to limit potential suspects to a small sub-group of the general population are of primary interest. Promising advances have been made in the DNA-based prediction of some broad group-specific externally visible characteristics (EVC) and the inference of the biogeographic ancestry of a stain donor. Predicting such externally visible characteristics (EVC) from DNA evidence (termed forensic DNA phenotyping - FDP) have become the focus of forensic genetic research in recent years. While human appearance is individual-specific with the exception of monozygotic twins, a clearly heritable component useful in a forensic scenario can be assumed for pigmentation-related traits like eye-colour, skin colour and hair colour. Early candidate gene studies examining

genes known to have an effect on skin-, hair- and eye-colour in mice (reviewed in [124]) revealed several SNPs in genes involved in melanogenesis which appeared to have an impact on human EVCs as well [31, 47, 80, 112, 127].

Based on these early works, the first publication (section 2.1) describes the results of a small scale pilot study in which several promising SNPs for the prediction of hair-, skin- and eye-colour have been combined into a small multiplex genotyping assay and tested on population samples of European and sub-Saharan African origin [137]. SNP typing was performed in a single-tube multiplex reaction incorporating 11 SNPs into one assay using the ABI PRISM®SNaPshot™kit (Life Technologies, Darmstadt, Germany).

The method is based on a two-step protocol consisting of a multiplex PCR amplifying short target sequences of the template DNA containing the SNP sites in question followed by a multiplex single base extension (SBE) reaction and capillary electrophoresis (CE). The process, generally following the same principle as Sanger sequencing [104], but only extending the primer sequence by exactly one base and thus detecting the SNP site in question, is outlined in Figure 3.1. The SNaPshot™protocol has become the standard method of SNP typing in routine forensic genetic applications due to its straightforward design using equipment readily available in the routine laboratory [103]. The enrichment of target sequences by PCR makes the approach useful for the analysis of limited quantities of trace DNA. Since the detection of the individual SNPs is performed in a separate second reaction, the originally amplified PCR products do not need to be separated by electrophoresis and therefore are allowed to overlap in their sizes [103]. This allows the design of PCR primers amplifying each SNP in the multiplex in a short fragment often < 150 bp [103, 131, 132] making the successful amplification even from severely degraded DNA possible.

While the SNaPshot™technique is sensitive and robust enough for the analysis of even challenging DNA samples in a forensic genetic scenario [103, 131, 132], analysis of the generated electropherograms is not a trivial task, especially for larger multiplexes. Automatic allele calling is generally possible using the manufacturer's GeneMapper 4.0 (Life Technologies, Darmstadt, Germany) software, but can be error prone due to imbalances in signal intensities arising as consequence of imbalances in the PCR or SBE reaction, or due to the difference in emission intensities of the fluorophores used in the SNaPshot™chemistry [103].

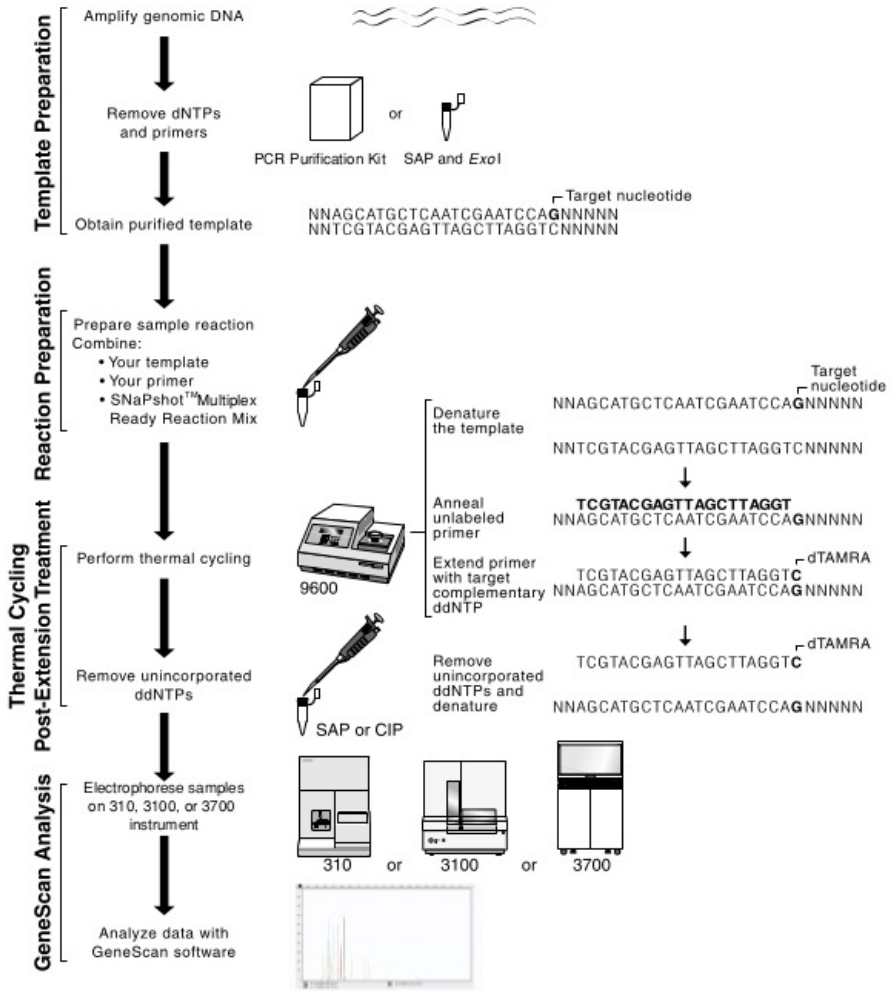


Figure 3.1: Outline of the ABI Prism®SNaPshot™technique for SNP detection; figure taken from the manufacturer's protocol [70]

The assay presented in the publication (section 2.1) was successfully applied to the analysis of two populations of 71 sub-Saharan African and 87 European individuals selected to have a high chance of exhibiting strongly divergent phenotypes (e.g. light skin and eye colour in Europeans vs. dark skin and eye colour in Africans). Figure 2 of the original publication shows significant allele frequency differences of $> 40\%$ for five out of the tested eleven markers, namely the SNPs rs7170989 (B), rs4778138 (C), rs16891982 (E) and rs1375164 (I) located in the OCA2 and SLC45A2 genes associated with iris-colour [31, 47, 80], as well as SNP rs1426654 (D) located in the gene SLC24A5 associated with skin colour [112].

While a study by Sturm et al. in the same year [118], mapping the region of the OCA2 locus with high resolution, found a single SNP in an intronic region of the HERC2 gene located upstream of OCA2 to be explanatory for 80% of brown eye phenotypes in the study population, the authors believe the effect to occur through regulatory effects of this HERC2 SNP on the expression of OCA2 [118]. The fact that the SNPs labelled B, C and I in our publication are located within < 300 kbp from each other (visualised in Fig. 3.2) inside a region of strong linkage disequilibrium [74] spanning the OCA2-HERC2 locus also containing the SNP rs12913832 identified by Sturm et al.[118] suggests that the detected differences in allele frequency might result from linkage to this highly eye-colour predictive marker. The SNP rs1800407 (designated M in our publication) located in the OCA2 gene has been confirmed to modify penetrance of the regulatory effect of the HERC2 SNP [118] with the C allele (detected as G on the opposite strand in the assay described in section 2.1) being associated with non-blue eye colour in populations of European descent [74, 118], especially when present together with the dark eye genotype of the highly predictive SNP rs12913832 in the HERC2 locus [131]. This is consistent with the slight difference in absolute allele frequencies ~ 0.06 between generally dark eyed Africans and potentially light eyed Europeans detected in our study.

The allele frequency difference observed in the remaining eye-colour associated SNP rs16891982 with the C allele almost exclusively found in the African population (allele frequency of 0.96 vs. 0.10 in Europeans, see section 2.1, Fig. 2 of the publication) suggests association of the C allele with brown eye colour, consistent with later observations by Walsh et al. finding the homozygote genotype GG (detected in our assay as C on the opposite strand) to be strongly associated with brown eye colour.

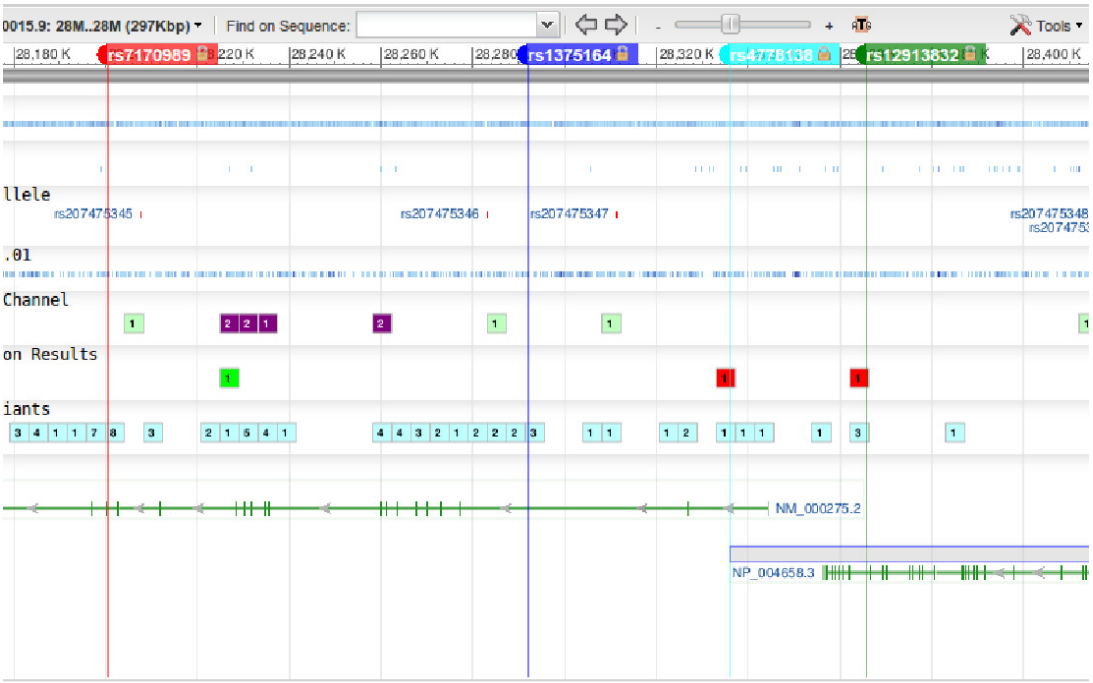


Figure 3.2: Location of the SNPs presenting with strong allele frequency differences between the studied populations relative to the highly eye-colour predictive SNP rs12913832 (green label, on the right side); distance between all markers: < 300 kbp. Figure generated using the GeneView tool on the dbSNP website [1]

Similarly, the marker found to be associated with skin colour, rs1426654 (designated D in our publication) presents with an allele frequency of > 0.9 for allele A in Europeans vs. < 0.1 in Africans, consistent with the observations of Spichenok et al. [115], who found the homozygous AA genotype of this marker to be highly predictive for light skin colour [115].

Summarising, the results of the small scale pilot study presented in section 2.1, clearly confirm the association of 6 out of the selected 11 SNPs with the admittedly broad phenotypic variation of dark eye- and skin colour vs. light eye- and skin colour consistent with the literature [74, 115, 131]. Due to missing individual phenotypic information in the two study populations, the power of prediction of the selected markers could not be assessed in more detail in the presented study. Additionally, we have been able to confirm the ancestry-informative character of pigmentation associated markers based on highly divergent allele frequencies between two continental Populations also postulated by other groups [26, 112, 131, 132]. This is not surprising given the fact that hair- and eye-colour variations are thought to be of European origin with the current phenotype frequency distribution caused by preferential partner selection [38]. Therefore, the notion has been raised that prediction of non-brown eye- and hair-colour could become more accurate when combined with the prediction of the biogeographic ancestry [59].

Recently, complete and comprehensive toolkits for the prediction of externally visible characteristics from forensic DNA samples have been presented with the IrisPlex for eye-colour prediction [131] and the HIrisPlex extending the IrisPlex for the simultaneous prediction of eye- and hair colour [132]. Those toolkits make use of the most predictive markers for EVCs known to date and combine highly sensitive SNaPshot^{TM}based SNP typing assays with user-friendly model based prediction algorithms for the prediction of EVCs from DNA evidence without the requirement for additional information about the biogeographic ancestry of the stain donor [131, 132], although the authors state that a worldwide study with available phenotypic information would be of interest to support this claim. Both assays have undergone extensive forensic validation and have been proven to be ready for routine use [30, 130].

While the prediction of biogeographic ancestry using pigmentation related markers appears to be possible at least on a population level [131], this could not be confirmed on individual level much more interesting in forensic genetics. For the inference of biogeographic ancestry on an individual level, using a large number of randomly selected markers [69] or the use of smaller pan-

els of specially selected ancestry informative marker (AIM) panels [62, 67] has been proven successful. Previously described technical characteristics in combination with an abundance of information as a result of the publication of the complete human genome sequence [24] resulted in single nucleotide markers becoming the class of choice for this task not only in forensic genetic research[39, 89, 90].

Intrigued by the technical advantages of the recently emerging short insertion/deletion markers [76, 77, 134] we decided to shift our research focus to the evaluation of this marker class for the prediction of biogeographic ancestry. Early work by Yang et al. and Bastos-Rodrigues et al. proved the suitability of indel markers as ancestry informative markers. It was shown that the power of discrimination directly depends on the number of markers [136], but consistent with the results of SNP-based studies, that with careful selection of the markers used, a differentiation of populations on a continental level is possible with a much smaller panel [10, 67, 90]. Similar results were obtained by Pereira et al., who published their 46 indel-AIM single-tube assay in 2012. With our work, we have been able to show that a comparable level of resolution can already be achieved with a set of only 21 markers amplified in one single-tube multiplex reaction [138] (see section 2.2 for original publication).

The assay presented has been proven to be comparable in sensitivity and specificity with previously published SNP-based assays, allowing successful genotyping of all markers with an input amount of between 0.1 and 1.0 ng of genomic DNA (see Fig. 2 of the original publication, section 2.2). The simple assay design consisting of a single-tube PCR reaction with optional PCR product cleanup followed by capillary electrophoresis, the steps required to produce a DNA profile are considerably reduced in comparison with the multi-step SNaPshot^{TM}procedure routinely used for SNP typing, thus further minimising the risk of contamination in the laboratory. By keeping the fragment length of all markers below 200 bp, the assay was specifically designed for the analysis of severely degraded DNA commonly found in forensic casework samples. Degradation studies with artificially degraded DNA have demonstrated amplification success for all markers at degradation levels where routinely used forensic STR typing kits are incapable of producing any result (section 2.2, Fig. 3 of the original publication), consistent with results obtained by Pereira et al. with their indel-based assay for forensic identification purposes [87]. Automated allele calling using the GeneMapper ID software (Life Technologies, Darmstadt, Germany) was possible based on an allelic ladder containing both the

short (insertion) and the long (deletion) allele at each locus. The allelic ladder was constructed by amplifying previously identified heterozygous DNA samples for each locus in singleplex reactions, mixing the obtained PCR products in order to obtain balanced peak heights across the markers and reamplifying the allelic ladder mix in a multiplex reaction. The efficiency of the assay in inferring biogeographic ancestry was assessed by conducting a population genetic study on 379 individuals from seven geographic regions. The predictive power of the marker set for the estimation of biogeographic ancestry was assessed using two different statistical approaches: model-based cluster analysis using the STRUCTURE software [34, 35, 94] and a Bayesian likelihood-ratio based approach implemented in the SNIPPER app suite [90] available online at `http://mathgene.usc.es/snipper/`.

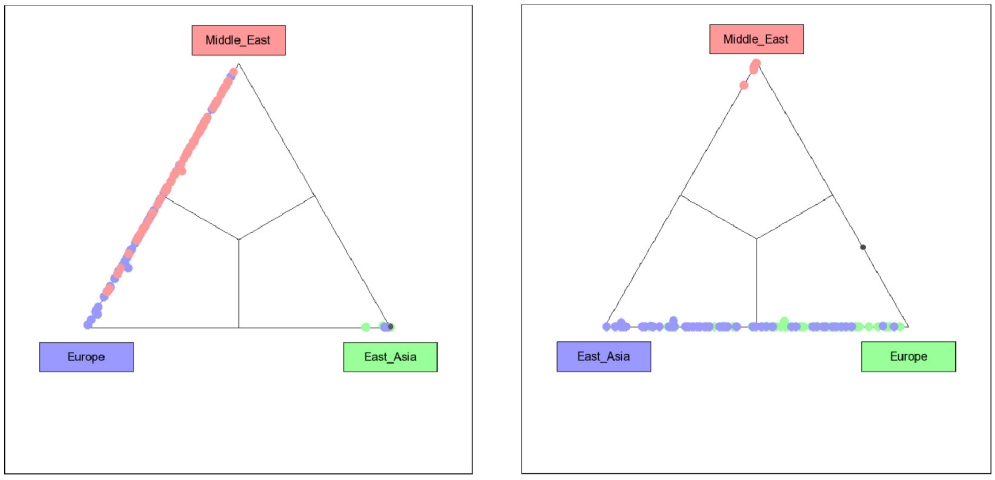
The model-based clustering approach implemented in the STRUCTURE program uses Markov Chain Monte Carlo (MCMC) simulations to simultaneously estimate the most probable allele frequency distribution for a given set of populations k and the most probable population (or populations if the possibility of admixture is taken into account) for each individual sample given the allele frequency distributions estimated and the genotype information of the sample. By repeating the analysis for different numbers of populations and comparing the estimated posterior probabilities of k, the most likely number of populations supported by the available data can be estimated. This approach has clear advantages for many population genetic applications since clustering is performed based solely on the available genotype data with no prior information about population structure necessary [94]. If additional information with possible influence on population structure (sampling location, place of birth, etc.) is available, the model can easily be adopted to take this information into account. Due to its independence from prior information about population structure, this method is ideally suited for assessing the efficiency of a newly designed marker set in estimation of population structure. It is however less suited for the application of ancestry estimation in forensic casework scenarios. While it is possible to use the approach to assign an unknown sample to the most probable population by the analysis of a set of reference samples of known population of origin as a training set together with the unknown sample and checking for the cluster the unknown gets assigned to, this approach is highly impractical. MCMC simulations generally require a large number of iterations (regularly several tens to hundreds of thousands) to converge on a stable estimation of the posterior probabilities in question, resulting in run

times of up to several hours for each analysis [94] even on state-of-the-art computers. Time requirements on this scale are prohibitive for methods used routinely in forensic casework, where results are usually required as fast as possible.

The alternative method implemented in the SNIPPER app suite [90] implements a maximum likelihood calculation using the allele frequencies of a training set of samples with known populations of origin to estimate likelihood parameters for each population and assigning an unknown sample to the population for which the highest posterior probability is calculated. This method allows for the rapid classification of individuals of unknown ancestry into the most likely of a set of known populations. While this approach is suitable for the application in forensic case-work, the requirement for a training set of samples with known populations of origin limits its use in population genetic applications where prior information about the analysed samples of unknown origin is not available. Since the SNIPPER software was initially developed for the analysis of SNP genotypes [90], the data input format is restricted to genotypes consisting of single base letter codes (A, C, T or G and N for missing data), a limitation easily overcome by transcribing the indel genotypes by designating the short allele as "C" and the long allele as "A".

In order to assess the predictive value of the newly developed 21-plex indel AIM assay, a thorough cross-validation was performed, assessing the classification success by classifying each sample in the dataset in turn against a training set consisting of all remaining samples. Leaving out the Indo-Pakistani and Afghan samples due to their small sample size and combining the South-East Asian and East Asian populations into one group as suggested by the results of an AMOVA analysis and the pairwise F_{st} calculations, classification success was tested on the scale of three major continental population groups (Europe, Africa and East Asia) with one intermediate group from the Middle East (mainly from Iran and Iraq).

Example triangle plots visualising the classification are presented in Fig. 3.3. Classification was shown to be highly discriminative for the African and East Asian populations with a prediction success rate of $\sim 97\%$ being only slightly lower than the success rate of almost 100% obtained by Pereira et al. using 46 AIM-indels [88]. Difficulties were encountered in the classification of intermediate populations such as the differentiation between populations from Europe and the Middle East. Classification success for these populations using the 21 indel markers was estimated at $\sim$ 77% for the Middle Eastern sample with



(a) East Asian individual (b) Middle Eastern individual

Figure 3.3: Triangle Plots visualising the classification of an East Asian (Panel a) and a Middle Eastern (Panel b) individual with the SNIPPER software. The triangle vertices represent the three most likely populations of origin for the classified sample with the vertix representing an assignment probability of 1 and the opposite side a probability of 0. Training set samples are represented by dots in the colour of the population label while the test sample is represented as a grey dot. Classification of a randomly selected East Asian sample resulted in a probability of assignment of close to 1 for the East Asian population represented by the grey dot appearing in the East Asian vertex in Panel a. Classification of a randomly selected individual from the Middle Eastern population resulted in the assignment probability for the European population being slightly higher than that for the Middle Eastern population represented by the grey dot located on the Europe–Middle East side of the triangle, about 3/4 towards the European vertex in Panel b.

the misclassified samples mostly being assigned to the European population. European samples showed an even worse performance with a rate of misclassification of almost 60% being assigned as Middle Eastern. This is consistent with the model based approach of ancestry estimation using the STRUCTURE program [34, 35, 94] where European and Middle Eastern populations cluster together (ref. Fig. 4 of the original publication, section 2.2.

While some amount of error in classifying intermediate populations had to be expected, since the markers had been selected on the premise of being able to distinguish large continental populations from each other, these findings are also consistent with the results of the SNP-based 34-plex assay presented by Phillips et al. [90]. Phillips et al. found a slightly different pattern with the European population being more clearly defined and the Middle Eastern and Central South Asian populations being difficult to distinguish using their marker set (reference Fig. 3 in [90]). Some of the observed divergence

can possibly be attributed to differences in the composition of the analysed populations. However, follow-up studies on AIM-SNPs by Phillips et al. and Bulbul et al. shows that discrimination of Middle Eastern, European and Central South Asian populations may not be possible using small scale multiplex assays suitable for a forensic casework scenario. Supplementing the 34-plex assay with an additional 23 AIM-SNPs selected to differentiate between European and South Asian population (termed "Eurasiaplex", [92]) as well as another 32 SNPs consisting of 22 AIMs and 10 SNPs predictive for iris colour [16] resulting in a combined panel of 85 carefully selected SNPs did not improve the resolution of European and Middle Eastern populations significantly (see Fig. 3.4).

Slight improvements of the separation between the South Asian and European sub-populations observed with the use of the full panel of 85 SNPs suggests that further differentiation might still be possible by incorporating additional markers. Since the number of markers that can be analysed in a single PCR/SNaPshot^{TM}reaction is eventually limited by technical constraints, sequential resolution of biogeographic ancestry using a collection of smaller multiplexes may be a viable option, provided that a sufficient amount of sample is available for testing. Using a broad range marker set to distinguish large continental populations followed by the application of specialised marker sets tailored for differentiation of sub-populations within a given continental population, potentially combining different marker classes such as SNPs and indels where applicable, could prove a viable approach. The 21-plex AIM-indel assay presented in section 2.2 could be a suitable starting point for sequential analysis as it is able to clearly distinguish three major continental population groups while leaving sufficient room for extension by only using three of the available dye-channels available in modern capillary electrophoresis instruments. The remaining two dye channels could possibly be used to increase the power of the assay by allowing the simultaneous analysis of a second two-channel multiplex assay selected for specific questions in a mix and match fashion. In any circumstances, extensive research into the identification of additional ancestry informative markers targeted at Eurasian sub-populations will be necessary if the power of differentiation of forensic ancestry informative marker panels is to be improved. Recent advances in next generation sequencing (NGS) approaches offer some very interesting possibilities for forensic genetic applications as well (reviewed in [11]).

Identification of novel markers with high-throughput parallel sequencing ap-

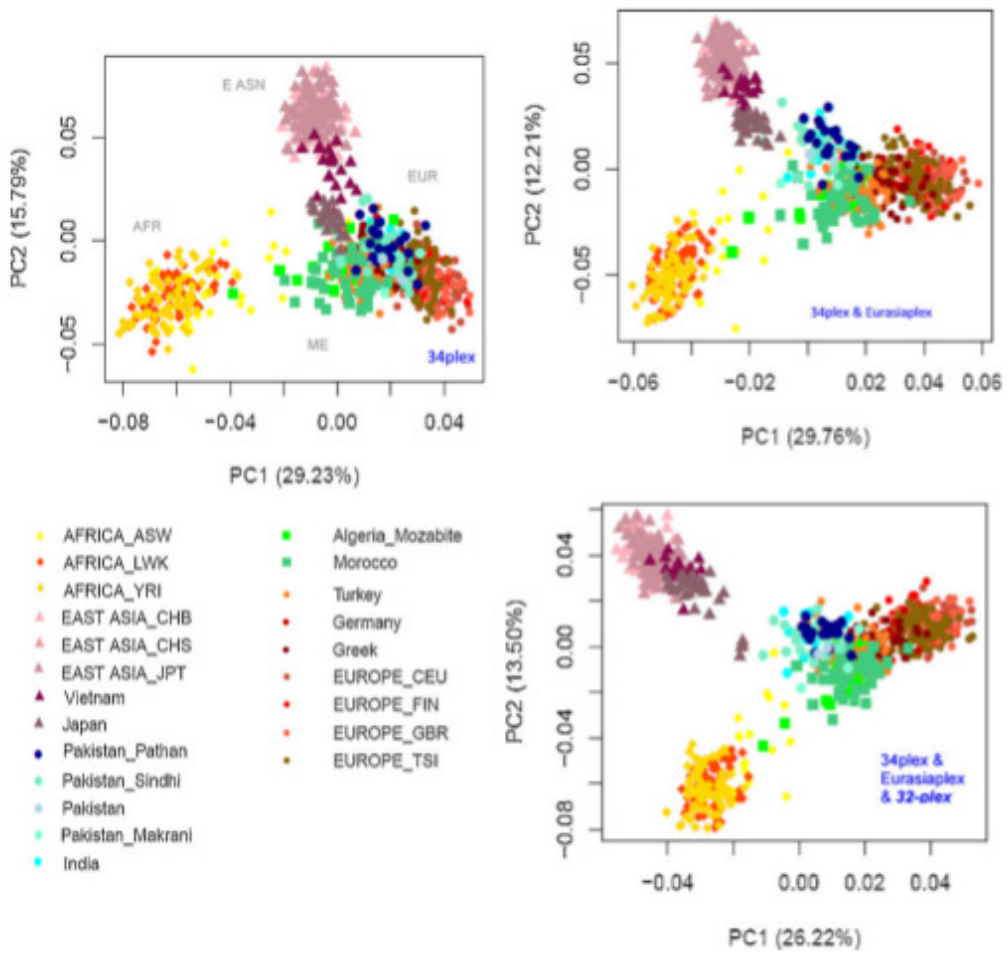


Figure 3.4: Principal component analysis plots depicting the discrimination of 22 populations (HGDP-CEPH and 1000 Genomes reference panels plus 7 Eurasian populations) using 85 ancestry informative and pigmentation related SNPs in various combinations. Taken from [16]

proaches could help to increase the discrimination power of already existing methods. On the other hand, this new technology shows some promising technological advantages for the development of novel forensic genetic methods. Since most currently available NGS technologies produce sequences with read lengths of only 50-100 bp [11], these methods would be ideally suited for the analysis of degraded DNA. While the relatively large amount of 1-5 μg of genomic DNA currently required for NGS applications could prove a limiting factor, the possibility of analysing a large number of different markers for different forensic applications in parallel could well compensate for this drawback. Currently available methods might not meet forensic quality requirements yet [9], but the rapid advances in the field warrant closer investigation of these technologies.

4 Summary

Current forensic genetic research is focussed on novel applications to extract additional information from DNA evidence in cases where classical comparative approaches of DNA analysis are unsuccessful. In cases where no witnesses or suspects are available and comparison of the DNA evidence with DNA databases of reference samples does not provide a match, it would be interesting for investigators to extract information from the DNA sample resulting in new investigative leads needed to solve the case. The presented work explores two such approaches, namely so-called "forensic DNA phenotyping" aiming at the extraction of information about externally visible characteristics from DNA samples, and the estimation of the biogeographic ancestry of the stain donor as a means of excluding large population groups as potential suspects. In a small scale pilot study, the applicability of the analysis of a set of eleven pigmentation-related single nucleotide polymorphisms has been assessed for the estimation of eye- and skin colour. While the newly developed assay was successful in inferring eye- and skin colour on a very broad scale, i.e. brown eye-colour and dark skin vs. non-brown eye-colour and light skin, it also provided information about the biogeographic ancestry of the sample donor.

The estimation of biogeographic ancestry was further explored in the second work presented here using the novel marker class of short insertion/deletion polymorphisms. A novel multiplex-PCR assay containing 21 short ancestry informative indels in a single tube reaction was developed and thoroughly validated. The assay successfully discriminates between three major continental populations (Europe, sub-Saharan Africa and East Asia) while being sensitive and specific enough for the analysis of low quantities of low quality DNA regularly encountered in forensic genetic case work samples.

5 Zusammenfassung in Deutscher Sprache

Die aktuelle forensisch-genetische Forschung beschäftigt sich verstärkt mit der Erlangung zusätzlicher Informationen aus DNA-Proben in Fällen, in denen die klassischen, vergleichenden Analyse-Ansätze nicht erfolgreich sind. In Kriminalfällen ohne Zeugen und Verdächtige, in welchen der Vergleich der verfügbaren DNA-Profile mit DNA-Datenbanken keinen Treffer erzielt, ist das Ziel die Erlangung von Informationen, die geeignet sind, die Anzahl möglicher Verdächtiger auf ein kriminalistischer Arbeit zugängliches Maß zu reduzieren. Die vorliegende Arbeit untersucht zwei solcher Ansätze, nämlich die Bestimmung äußerlicher phänotypischer Merkmale und der biogeographischen Herkunft eines unbekannten Spurenlegers. Der erste Teil der Arbeit beschreibt die Entwicklung einer Methode zur genetischen Typisierung von elf Pigmentierungsassoziierten Einzelnukleotid-Polymorphismen in einer Multiplex-PCR-Reaktion. Es konnte gezeigt werden, dass die Bestimmung von Augen- und Hautfarbe zumindest für die Extremwerte "braune Augen" und "dunkle Hautfarbe" gegenüber "helle Augen" und "helle Hautfarbe" mit dieser Methode grundsätzlich möglich ist. Weiter wurde ermittelt, dass die gewählten genetischen Marker geeignet sind, zwischen eropäischer und nicht-europäischer Abstammung eines Spurenlegers zu unterscheiden. Der zweite Teil der Arbeit verfolgt den Aspekt der biogeographischen Herkunft weiter und konzentriert sich dabei auf die noch relativ neue Marker-Klasse der kurzen Insertions/Deletions-Polymorphismen (Indel). Ein neu entwickelter Assay zur Bestimmung von 21 kurzen Indels mit Informationsgehalt für die biogeographische Herkunft wurde mit Hinblick auf eine forensische Anwendung validiert. Die Methode ist geeignet, drei große kontinentale Bevölkerungsgruppen (Europa, Afrika südlich der Sahara und Ost-Asien) voneinander zu unterscheiden. Das Verfahren ist dabei ausreichend spezifisch und sensitiv für die Analyse von geringen Mengen von DNA schlechter Qualität, wie sie in forensisch genetischer Fallarbeit regelmäßig vorliegen kann.

6 References

[1] dbSNP, Short Genetic Variations. URL http://www.ncbi.nlm.nih.gov/snp/. last visited: 09.05.2013.

[2] A. Amorim and L. Pereira. Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. *Forensic Sci Int*, 150(1):17–21, May 2005. doi: 10.1016/j.forsciint.2004.06.018. URL http://dx.doi.org/10.1016/j.forsciint.2004.06.018.

[3] S. Anderson, A. T. Bankier, B. G. Barrell, M. H. de Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. Smith, R. Staden, and I. G. Young. Sequence and organization of the human mitochondrial genome. *Nature*, 290(5806):457–465, Apr 1981.

[4] R. M. Andrews, I. Kubacka, P. F. Chinnery, R. N. Lightowlers, D. M. Turnbull, and N. Howell. Reanalysis and revision of the cambridge reference sequence for human mitochondrial DNA. *Nat Genet*, 23(2):147, Oct 1999. doi: 10.1038/13779. URL http://dx.doi.org/10.1038/13779.

[5] W. Arber and D. Dussoix. Host specificity of DNA produced by Escherichia coli: I. Host controlled modification of bacteriophage λ. *Journal of Molecular Biology*, 5(1):18 – 36, 1962. ISSN 0022-2836. doi: 10.1016/S0022-2836(62)80058-8. URL http://www.sciencedirect.com/science/article/pii/S0022283662800588.

[6] K. G. Ardlie, L. Kruglyak, and M. Seielstad. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet*, 3(4):299–309, Apr 2002. doi: 10.1038/nrg777. URL http://dx.doi.org/10.1038/nrg777.

[7] J.W. Bacher, C. Helms, H. Donis-Keller, L. Hennes, N. Nassif, and J.W. Schumm. Chromosome localization of CODIS loci and new pentanucleotide repeat loci. *Prog. Forensic Genet.*, 8:33–36, 2000.

[8] H-J. Bandelt and W. Parson. Consistent treatment of length variants in the human mtDNA control region: a reappraisal. *Int J Legal Med*, 122(1):11–21, Jan 2008. doi: 10.1007/s00414-006-0151-5. URL http://dx.doi.org/10.1007/s00414-006-0151-5.

[9] Hans-Jürgen Bandelt and Antonio Salas. Current next generation sequencing technology may not meet forensic standards. *Forensic Sci Int Genet*, 6(1):143–145, Jan 2012. doi: 10.1016/j.fsigen.2011.04.004. URL http://dx.doi.org/10.1016/j.fsigen.2011.04.004.

[10] L. Bastos-Rodrigues, J. R. Pimenta, and S. D. J. Pena. The genetic structure of human populations studied through short insertion-deletion polymorphisms. *Ann Hum Genet*, 70(Pt 5):658–665, Sep 2006. doi: 10.1111/j.1469-1809.2006.00287.x. URL http://dx.doi.org/10.1111/j.1469-1809.2006.00287.x.

[11] Eva C Berglund, Anna Kiialainen, and Ann-Christine Syvänen. Next-generation sequencing technologies and applications for human genetic history and forensics. *Investig Genet*, 2:23, 2011. doi: 10.1186/2041-2223-2-23. URL http://dx.doi.org/10.1186/2041-2223-2-23.

[12] C. W. Birky. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. *Proc Natl Acad Sci U S A*, 92(25):11331–11338, Dec 1995.

[13] C. W. Birky. The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. *Annu Rev Genet*, 35:125–148, 2001. doi: 10.1146/annurev.genet.35.102401.090231. URL http://dx.doi.org/10.1146/annurev.genet.35.102401.090231.

[14] B. Brinkmann. Is the amelogenin sex test valid? *Int J Legal Med*, 116(2):63, Apr 2002.

[15] A. J. Brookes. The essence of SNPs. *Gene*, 234(2):177–186, Jul 1999.

[16] O. Bulbul, G. Filoglu, H. Altuncul, A. Freire Aradas, Y. Ruiz, M. Fondevila, C. Phillips, Á. Carracedo, A.K. Kriegel, and P.M. Schneider. A SNP multiplex for the simultaneous prediction of biogeographic ancestry and pigmentation type. *Forensic Science International: Genetics Supplement Series*, 3(1):e500 – e501, 2011. ISSN 1875-1768. doi: 10.1016/j.fsigss.2011.10.001. URL http://www.sciencedirect.com/science/article/pii/S1875176811002496.

[17] J. M. Butler. *Forensic DNA Typing*. Elsevier Academic Press, 2005.

[18] J. M. Butler. Genetics and genomics of core short tandem repeat loci used in human identity testing. *J Forensic Sci*, 51(2):253–265, Mar 2006. doi: 10.1111/j.1556-4029.2006.00046.x. URL http://dx.doi.org/10.1111/j.1556-4029.2006.00046.x.

[19] R. Chakraborty, D. N. Stivers, B. Su, Y. Zhong, and B. Budowle. The utility of short tandem repeat loci beyond human identification: Implications for development of new DNA typing systems. *Electrophoresis*, 20(8):1682–1696, 1999. ISSN 1522-2683. doi: 10.1002/(SICI)1522-2683(19990101)20:8⟨1682::AID-ELPS1682⟩3.0.CO;2-Z.

[20] T. M. Clayton, J. P. Whitaker, and C. N. Maguire. Identification of bodies from the scene of a mass disaster using dna amplification of short tandem repeat (STR) loci. *Forensic Sci Int*, 76(1):7–15, Nov 1995.

[21] J. R. Collins, R. M. Stephens, B. Gold, B. Long, M. Dean, and S. K. Burt. An exhaustive DNA micro-satellite map of the human genome using high performance computing. *Genomics*, 82(1):10–19, Jul 2003.

[22] H. E. Collins-Schramm, C. M. Phillips, D. J. Operario, J. S. Lee, J. L. Weber, R. L. Hanson, W. C. Knowler, R. Cooper, H. Li, and M. F. Seldin. Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. *Am J Hum Genet*, 70(3):737–750, Mar 2002. doi: 10.1086/339368. URL `http://dx.doi.org/10.1086/339368`.

[23] National Research Council Committee on DNA Forensic Science: An Update. *The Evaluation of Forensic DNA Evidence*. The National Academies Press, 1996. ISBN 9780309121941. URL `http://www.nap.edu/openbook.php?record_id=5141`.

[24] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, Oct 2004.

[25] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, Jun 2007. doi: 10.1038/nature05911. URL `http://dx.doi.org/10.1038/nature05911`.

[26] R. Daniel, J. J. Sanchez, N. T. Nassif, A. Hernandez, and S. J. Walsh. SNPs associated with physical traits: A valuable tool for the inference of biogeographical ancestry. *Forensic Science International: Genetics Supplement Series*, 1(1): 538 – 540, 2008. ISSN 1875-1768. doi: 10.1016/j.fsigss.2007.10.165. URL `http://www.sciencedirect.com/science/article/pii/S1875176808001959`.

[27] P. de Knijff, M. Kayser, A. Caglià, D. Corach, N. Fretwell, C. Gehrig, G. Graziosi, F. Heidorn, S. Herrmann, B. Herzog, M. Hidding, K. Honda, M. Jobling, M. Krawczak, K. Leim, S. Meuser, E. Meyer, W. Oesterreich, A. Pandya, W. Parson, G. Penacino, A. Perez-Lezaun, A. Piccinini, M. Prinz, and L. Roewer. Chromosome Y microsatellites: population genetic and evolutionary aspects. *Int J Legal Med*, 110(3):134–149, 1997.

[28] G. Destro-Bisol, I. Boschi, A. Caglià, S. Tofanelli, V. Pascali, G. Paoli, and G. Spedini. Microsatellite variation in Central Africa: an analysis of intrapopulational and interpopulational genetic diversity. *Am J Phys Anthropol*, 112(3):319–337, Jul 2000. doi: 3.0.CO;2-F. URL `http://dx.doi.org/3.0.CO;2-F`.

[29] L.A. Dixon, C.M. Murray, E.J. Archer, A.E. Dobbins, P. Koumi, and P. Gill. Validation of a 21-locus autosomal SNP multiplex for forensic identification purposes. *Forensic Science International*, 154(1):62 – 77, 2005. ISSN 0379-0738. doi: DOI:10.1016/j.forsciint.2004.12.011. URL `http://www.sciencedirect.com/science/article/pii/S0379073804008175`.

[30] J. Draus-Barini, S. Walsh, E. Pośpiech, T. Kupiec, H. Glab, W. Branicki, and M. Kayser. Bona fide colour: Dna prediction of human eye and hair colour from ancient and contemporary skeletal remains. *Investig Genet*, 4(1):3, 2013. doi: 10.1186/2041-2223-4-3. URL http://dx.doi.org/10.1186/2041-2223-4-3.

[31] D. L. Duffy, G. W. Montgomery, W. Chen, Z. Z. Zhao, L. Le, M. R. James, N. K. Hayward, N. G. Martin, and R. A. Sturm. A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. *Am J Hum Genet*, 80(2):241–252, Feb 2007. doi: 10.1086/510885. URL http://dx.doi.org/10.1086/510885.

[32] A. Edwards, A. Civitello, H. A. Hammond, and C. T. Caskey. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *Am J Hum Genet*, 49 (4):746–756, Oct 1991.

[33] H. Eiberg, J. Troelsen, M. Nielsen, A. Mikkelsen, J. Mengel-From, K. W. Kjaer, and L. Hansen. 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(2):177–187, Mar 2008. doi: 10.1007/s00439-007-0460-x. URL http://dx.doi.org/10.1007/s00439-007-0460-x.

[34] D. Falush, M. Stephens, and J. K. Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164 (4):1567–1587, Aug 2003.

[35] D. Falush, M. Stephens, and J. K. Pritchard. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes*, 7(4): 574–578, Jul 2007. doi: 10.1111/j.1471-8286.2007.01758.x. URL http://dx.doi.org/10.1111/j.1471-8286.2007.01758.x.

[36] M. Fondevila, C. Phillips, N. Naveran, M. Cerezo, A. Rodriguez, R. Calvo, L. M. Fernandez, Á. Carracedo, and M. V. Lareu. Challenging DNA: Assessment of a range of genotyping approaches for highly degraded forensic samples. *Forensic Science International: Genetics Supplement Series*, 1(1):26–28, 2008.

[37] M. Fondevila, C. Phillips, N. Naveran, L. Fernandez, M. Cerezo, A. Salas, Á. Carracedo, and M. V. Lareu. Case report: identification of skeletal remains using short-amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur. *Forensic Sci Int Genet*, 2(3):212–218, Jun 2008. doi: 10.1016/j.fsigen.2008.02.005. URL http://dx.doi.org/10.1016/j.fsigen.2008.02.005.

[38] P. Frost. European hair and eye color: A case of frequency-dependent sexual selection? *Evolution and Human Behavior*, 27(2):85 – 103, 2006. ISSN 1090-5138. doi: 10.1016/j.evolhumbehav.2005.07.002. URL http://www.sciencedirect.com/science/article/pii/S1090513805000590.

[39] T. Frudakis, K. Venkateswarlu, M. J. Thomas, Z. Gaskin, S. Ginjupalli, S. Gunturi, V. Ponnuswamy, S. Natarajan, and P. K. Nachimuthu. A classifier for the SNP-based inference of ancestry. *J Forensic Sci*, 48(4):771–782, Jul 2003.

[40] G. Geserick and I. Wirth. Genetic Kinship Investigation from Blood Groups to DNA Markers. *Transfus Med Hemother*, 39(3):163–175, Jun 2012. doi: 000338850. URL `http://dx.doi.org/000338850`.

[41] P. Gill. An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. *Int J Legal Med*, 114(4-5):204–210, 2001.

[42] P. Gill, A. J. Jeffreys, and D. J. Werrett. Forensic application of DNA 'fingerprints'. *Nature*, 318(6046):577–579, 1985.

[43] P. Gill, L. Fereday, N. Morling, and P. M. Schneider. The evolution of DNA databases-recommendations for new European STR loci. *Forensic Sci Int*, 156(2-3):242–244, Jan 2006. doi: 10.1016/j.forsciint.2005.05.036. URL `http://dx.doi.org/10.1016/j.forsciint.2005.05.036`.

[44] P. Gill, L. Fereday, N. Morling, and P. M. Schneider. New multiplexes for Europe-amendments and clarification of strategic development. *Forensic Sci Int*, 163(1-2): 155–157, Nov 2006. doi: 10.1016/j.forsciint.2005.11.025. URL `http://dx.doi.org/10.1016/j.forsciint.2005.11.025`.

[45] P. Gill, C. Phillips, C. McGovern, J.-A. Bright, and J. Buckleton. An evaluation of potential allelic association between the STRs vWA and D12S391: implications in criminal casework and applications to short pedigrees. *Forensic Sci Int Genet*, 6(4): 477–486, Jul 2012. doi: 10.1016/j.fsigen.2011.11.001. URL `http://dx.doi.org/10.1016/j.fsigen.2011.11.001`.

[46] W. Goodwin, A. Linacre, and S. Hadi. *An Introduction to Forensic Genetics*. John Wiley & Sons, Ltd., Chichester, England, 2007.

[47] J. Graf, R. Hodgson, and A. van Daal. Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation. *Hum Mutat*, 25(3): 278–284, Mar 2005. doi: 10.1002/humu.20143. URL `http://dx.doi.org/10.1002/humu.20143`.

[48] E. Hagelberg, I. C. Gray, and A. J. Jeffreys. Identification of the skeletal remains of a murder victim by DNA analysis. *Nature*, 352(6334):427–429, Aug 1991. doi: 10.1038/352427a0. URL `http://dx.doi.org/10.1038/352427a0`.

[49] G. H. Hardy. Mendelian proportions in a mixed population. *Science*, 28(706):49–50, Jul 1908. doi: 10.1126/science.28.706.49. URL `http://dx.doi.org/10.1126/science.28.706.49`.

[50] A. J. Jeffreys, J. F. Brookfield, and R. Semeonoff. Positive identification of an immigration test-case using human DNA fingerprints. *Nature*, 317(6040):818–819, 1985.

[51] A. J. Jeffreys, V. Wilson, and S L Thein. Hypervariable 'minisatellite' regions in human DNA. *Nature*, 314(6006):67–73, March 1985. URL `http://dx.doi.org/10.1038/314067a0`.

[52] A. J. Jeffreys, V. Wilson, and S. L. Thein. Individual-specific 'fingerprints' of human DNA. *Nature*, 316(6023):76–79, 1985.

[53] A. J. Jeffreys, M. J. Allen, E. Hagelberg, and A. Sonnberg. Identification of the skeletal remains of Josef Mengele by DNA analysis. *Forensic Sci Int*, 56(1):65–76, Sep 1992.

[54] M. A. Jobling and P. Gill. Encoded evidence: DNA in forensic analysis. *Nat Rev Genet*, 5(10):739–751, October 2004. ISSN 1471-0056. URL `http://dx.doi.org/10.1038/nrg1455`.

[55] M. A. Jobling and C. Tyler-Smith. The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet*, 4(8):598–612, Aug 2003. doi: 10.1038/nrg1124. URL `http://dx.doi.org/10.1038/nrg1124`.

[56] Y. W. Kan and A. M. Dozy. Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. *Proc Natl Acad Sci U S A*, 75(11):5631–5635, Nov 1978.

[57] T. M. Karafet, F. L. Mendez, M. B. Meilerman, P. A. Underhill, S. L. Zegura, and M. F. Hammer. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res*, 18(5):830–838, May 2008. doi: 10.1101/gr.7172008. URL `http://dx.doi.org/10.1101/gr.7172008`.

[58] M. Kayser and P. de Knijff. Improving human forensics through advances in genetics, genomics and molecular biology. *Nat Rev Genet*, 12(3):179–192, Mar 2011. doi: 10.1038/nrg2952. URL `http://dx.doi.org/10.1038/nrg2952`.

[59] M. Kayser and P. M. Schneider. DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations. *Forensic Sci Int Genet*, 3(3):154–161, Jun 2009. doi: 10.1016/j.fsigen.2009.01.012. URL `http://dx.doi.org/10.1016/j.fsigen.2009.01.012`.

[60] M. Kayser, A. Caglià, D. Corach, N. Fretwell, C. Gehrig, G. Graziosi, F. Heidorn, S. Herrmann, B. Herzog, M. Hidding, K. Honda, M. Jobling, M. Krawczak, K. Leim, S. Meuser, E. Meyer, W. Oesterreich, A. Pandya, W. Parson, G. Penacino, A. Perez-Lezaun, A. Piccinini, M. Prinz, C. Schmitt, and L. Roewer. Evaluation of Y-chromosomal STRs: a multicenter study. *Int J Legal Med*, 110(3):125–33, 141–9, 1997.

[61] M. Kayser, F. Liu, A. C. J. W. Janssens, F. Rivadeneira, O. Lao, K. van Duijn, M. Vermeulen, P. Arp, M. M. Jhamai, W. F. J. van Ijcken, J. T. den Dunnen, S. Heath, D. Zelenika, D. D. G. Despriet, C. C. W. Klaver, J. R. Vingerling, P. T. V. M. de Jong, A. Hofman, Y. S. Aulchenko, A. G. Uitterlinden, B. A. Oostra, and C. M. van Duijn. Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. *Am J Hum Genet*, 82(2):411–423, Feb 2008. doi: 10.1016/j.ajhg.2007.10.003. URL `http://dx.doi.org/10.1016/j.ajhg.2007.10.003`.

[62] Paula Kersbergen, Kate van Duijn, Ate D Kloosterman, Johan T den Dunnen, Manfred Kayser, and Peter de Knijff. Developing a set of ancestry-sensitive dna markers reflecting continental origins of humans. *BMC Genet*, 10:69, 2009. doi: 10.1186/1471-2156-10-69. URL `http://dx.doi.org/10.1186/1471-2156-10-69`.

[63] J. R. Kidd, F. R. Friedlaender, W. C. Speed, A. J. Pakstis, F. M. de la Vega, and K. K. Kidd. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *Investig Genet*, 2(1):1, 2011. doi: 10.1186/2041-2223-2-1. URL `http://dx.doi.org/10.1186/2041-2223-2-1`.

[64] C. P. Kimpton, P. Gill, A. Walton, A. Urquhart, E. S. Millican, and M. Adams. Automated DNA profiling employing multiplex amplification of short tandem repeat loci. *PCR Methods Appl*, 3(1):13–22, Aug 1993.

[65] B. E. Krenke, L. Viculis, M. L. Richard, M. Prinz, S. C. Milne, C. Ladd, A. M. Gross, T. Gornall, J. R. H. Frappier, A. J. Eisenberg, C. Barna, X. G. Aranda, M. S. Adamowicz, and B. Budowle. Validation of male-specific, 12-locus fluorescent short tandem repeat (STR) multiplex. *Forensic Sci Int*, 151(1):111–124, Jun 2005.

[66] K. Landsteiner. Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe. *Centralblatt für Bacteriologie*, 27:357–362, 1900.

[67] Oscar Lao, Kate van Duijn, Paula Kersbergen, Peter de Knijff, and Manfred Kayser. Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. *Am J Hum Genet*, 78(4):680–690, Apr 2006. doi: 10.1086/501531. URL `http://dx.doi.org/10.1086/501531`.

[68] B. L. LaRue, J. Ge, J. L. King, and B. Budowle. A validation study of the Qiagen Investigator DIPplex® kit; an INDEL-based assay for human identification. *Int J Legal Med*, 126(4):533–540, Jul 2012. doi: 10.1007/s00414-012-0667-9. URL `http://dx.doi.org/10.1007/s00414-012-0667-9`.

[69] Jun Z Li, Devin M Absher, Hua Tang, Audrey M Southwick, Amanda M Casto, Sohini Ramachandran, Howard M Cann, Gregory S Barsh, Marcus Feldman, Luigi L Cavalli-Sforza, and Richard M Myers. Worldwide human relationships inferred from

genome-wide patterns of variation. *Science*, 319(5866):1100–1104, Feb 2008. doi: 10.1126/science.1153717. URL http://dx.doi.org/10.1126/science.1153717.

[70] *ABI PRISM® SNaPshot™ Multiplex Kit Protocol Rev. B.* Life Technologies, Darmstadt, Germany, 2010.

[71] M. Litt and J. A. Luty. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am J Hum Genet*, 44 (3):397–401, Mar 1989.

[72] A. L. Lowe, A. Urquhart, L. A. Foreman, and I. W. Evett. Inferring ethnic origin by means of an STR profile. *Forensic Sci Int*, 119(1):17–22, Jun 2001.

[73] A. Mannucci, K. M. Sullivan, P. L. Ivanov, and P. Gill. Forensic application of a rapid and quantitative DNA sex test by amplification of the X-Y homologous gene amelogenin. *Int J Legal Med*, 106(4):190–193, 1994.

[74] J. Mengel-From, C. Børsting, J. J. Sanchez, H. Eiberg, and N. Morling. Human eye colour and HERC2, OCA2 and MATP. *Forensic Sci Int Genet*, 4(5):323–328, Oct 2010. doi: 10.1016/j.fsigen.2009.12.004. URL http://dx.doi.org/10.1016/j.fsigen.2009.12.004.

[75] R. D. Miller, M. S. Phillips, I. Jo, M. A. Donaldson, J. F Studebaker, N. Addleman, S. V. Alfisi, W. M. Ankener, H. A. Bhatti, C. E. Callahan, B. J. Carey, C. L. Conley, J. M. Cyr, V. Derohannessian, R. A. Donaldson, C. Elosua, S. E. Ford, A. M. Forman, C. A. Gelfand, N. M. Grecco, S. M. Gutendorf, C. R. Hock, M. J. Hozza, S. Hur, S. M. In, D. L. Jackson, S. A. Jo, S.-C. Jung, S. Kim, K. Kimm, E. F. Kloss, D. C Koboldt, J. M. Kuebler, F.-S. Kuo, J. A. Lathrop, J.-K. Lee, K. L. Leis, S. A. Livingston, E. G. Lovins, M. L. Lundy, S. Maggan, M. Minton, M. A. Mockler, D. W. Morris, E. P. Nachtman, B. Oh, C. Park, C.-W. Park, N. Pavelka, A. B. Perkins, S. L. Restine, R. Sachidanandam, A. J. Reinhart, K. E. Scott, G. J. Shah, J. M. Tate, S. A. Varde, A. Walters, J. R. White, Y.-K. Yoo, J.-E. Lee, M. T. Boyce-Jacino, P.-Y. Kwok, and S. N. P. Consortium Allele Frequency Project. High-density single-nucleotide polymorphism maps of the human genome. *Genomics*, 86(2):117–126, Aug 2005.

[76] R. E. Mills, C. T. Luttig, C. E. Larkins, A. Beauchamp, C. Tsui, W. S. Pittard, and S. E. Devine. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.*, 16(9):1182–1190, Sep 2006. doi: 10.1101/gr.4565806. URL http://dx.doi.org/10.1101/gr.4565806.

[77] R. E. Mills, W. S. Pittard, J. M. Mullaney, U. Farooq, T. H. Creasy, A. A. Mahurkar, D. M. Kemeza, D. S. Strassler, C. P. Ponting, C. Webber, and S. E. Devine. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res*, 21(6):830–839, Jun 2011. doi: 10.1101/gr.115907.110. URL http://dx.doi.org/10.1101/gr.115907.110.

[78] J. J. Mulero, C. W. Chang, L. M. Calandro, R. L. Green, Y. Li, C. L. Johnson, and L. K. Hennessy. Development and validation of the AmpFlSTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system. *J Forensic Sci*, 51(1):64–75, Jan 2006. doi: 10.1111/j.1556-4029.2005.00016.x. URL http://dx.doi.org/10.1111/j.1556-4029.2005.00016.x.

[79] M. W. Nachman and S. L. Crowell. Estimate of the Mutation Rate per Nucleotide in Humans. *Genetics*, 156(1):297–304, 2000. URL http://www.genetics.org/content/156/1/297.abstract.

[80] K. Nakayama, S. Fukamachi, H. Kimura, Y. Koda, A. Soemantri, and T. Ishida. Distinctive distribution of aim1 polymorphism among major human populations with different skin color. *J Hum Genet*, 47(2):92–94, 2002. doi: 10.1007/s100380200007. URL http://dx.doi.org/10.1007/s100380200007.

[81] M. M. Nass and S. Nass. Intramitochondrial fibers with DNA characteristics. I. fixation and electron staining reactions. *J Cell Biol*, 19:593–611, Dec 1963.

[82] M. Nothnagel, R. Szibor, O. Vollrath, C. Augustin, J. Edelmann, M. Geppert, C. Alves, L. Gusmão, M. Vennemann, Y. Hou, U.-D. Immel, S. Inturri, H. Luo, S. Lutz-Bonengel, C.. Robino, L. Roewer, B. Rolf, J. Sanft, K.-J. Shin, J. E. Sim, P. Wiegand, C. Winkler, M. Krawczak, and S. Hering. Collaborative genetic mapping of 12 forensic short tandem repeat (str) loci on the human x chromosome. *Forensic Sci Int Genet*, 6(6):778–784, Dec 2012. doi: 10.1016/j.fsigen.2012.02.015. URL http://dx.doi.org/10.1016/j.fsigen.2012.02.015.

[83] K. Lewis O'Connor, C. R. Hill, P. M. Vallone, and J. M. Butler. Linkage disequilibrium analysis of D12S391 and vWA in U.S. population and paternity samples. *Forensic Sci Int Genet*, 5(5):538–540, Nov 2011. doi: 10.1016/j.fsigen.2010.09.003. URL http://dx.doi.org/10.1016/j.fsigen.2010.09.003.

[84] L. M. Pardo, I. MacKay, B. Oostra, C. M. van Duijn, and Y. S. Aulchenko. The effect of genetic drift in a young genetically isolated population. *Ann Hum Genet*, 69(Pt 3):288–295, May 2005. doi: 10.1046/j.1529-8817.2005.00162.x. URL http://dx.doi.org/10.1046/j.1529-8817.2005.00162.x.

[85] W. Parson. Bedeutung der mtDNA-Analyse für forensische Fragestellungen. *Rechtsmedizin*, 19(3):183–194, 2009. ISSN 0937-9819. doi: 10.1007/s00194-009-0594-3. URL http://dx.doi.org/10.1007/s00194-009-0594-3.

[86] W. Parson and A. Dür. EMPOP - A forensic mtDNA database. *Forensic Science International: Genetics*, 1(2):88 – 92, 2007. ISSN 1872-4973. doi: 10.1016/j.fsigen.2007.01.018. URL http://www.sciencedirect.com/science/article/pii/S1872497307000555.

[87] R. Pereira, C. Phillips, C. Alves, A. Amorim, Á. Carracedo, and L. Gusmão. A new multiplex for human identification using insertion/deletion polymorphisms. *Electrophoresis*, 30(21):3682–3690, Nov 2009. doi: 10.1002/elps.200900274. URL http://dx.doi.org/10.1002/elps.200900274.

[88] R. Pereira, C. Phillips, N. Pinto, C. Santos, S. E. B. dos Santos, A. Amorim, Á. Carracedo, and L. Gusmão. Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. *PLoS One*, 7(1):e29684, 2012. doi: 10.1371/journal.pone.0029684. URL http://dx.doi.org/10.1371/journal.pone.0029684.

[89] C. Phillips, R. Fang, D. Ballard, M. Fondevila, C. Harrison, F. Hyland, E. Musgrave-Brown, C. Proff, E. Ramos-Luis, B. Sobrino, Á. Carracedo, M. R. Furtado, D. Syndercombe Court, and P. M. Schneider. Evaluation of the Genplex SNP typing system and a 49plex forensic marker panel. *Forensic Science International: Genetics*, 1 (2):180 – 185, 2007. ISSN 1872-4973. doi: DOI:10.1016/j.fsigen.2007.02.007. URL http://www.sciencedirect.com/science/article/pii/S1872497307000610.

[90] C. Phillips, A. Salas, J. J. Sánchez, M. Fondevila, A. Gómez-Tato, J. Alvarez-Dios, M. Calaza, M. Casares de Cal, D. Ballard, M. V. Lareu, Á. Carracedo, and SNPforID Consortium. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int Genet*, 1(3-4):273–280, Dec 2007.

[91] C. Phillips, D. Ballard, P. Gill, D. Syndercombe Court, Á. Carracedo, and M. V. Lareu. The recombination landscape around forensic strs: Accurate measurement of genetic distances between syntenic str pairs using hapmap high density snp data. *Forensic Sci Int Genet*, 6(3):354–365, May 2012. doi: 10.1016/j.fsigen.2011.07.012. URL http://dx.doi.org/10.1016/j.fsigen.2011.07.012.

[92] C. Phillips, A. Freire Aradas, A. K. Kriegel, M. Fondevila, O. Bulbul, C. Santos, F. Serrulla Rech, M. D Perez Carceles, Á. Carracedo, P. M. Schneider, and M. V. Lareu. Eurasiaplex: A forensic SNP assay for differentiating European and South Asian ancestries. *Forensic Sci Int Genet*, 7(3):359–366, May 2013. doi: 10.1016/j.fsigen.2013.02.010. URL http://dx.doi.org/10.1016/j.fsigen.2013.02.010.

[93] M. Prinz, K. Boll, H. Baum, and B. Shaler. Multiplexing of Y chromosome specific STRs and performance for mixed samples. *Forensic Sci Int*, 85(3):209–218, Mar 1997.

[94] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, Jun 2000.

[95] H. Pulker, M. V. Lareu, C. Phillips, and Á Carracedo. Finding genes that underlie physical traits of forensic interest using genetic tools. *Forensic Sci Int Genet*, 1(2): 100–104, Jun 2007. doi: 10.1016/j.fsigen.2007.02.009. URL http://dx.doi.org/10.1016/j.fsigen.2007.02.009.

[96] L. Roewer, M. Krawczak, S. Willuweit, M. Nagy, C. Alves, A. Amorim, K. Anslinger, C. Augustin, A. Betz, E. Bosch, A. Cagliá, Á. Carracedo, D. Corach, A. F. Dekairelle, T. Dobosz, B. M. Dupuy, S. Füredi, C. Gehrig, L. Gusmaõ, J. Henke, L. Henke, M. Hidding, C. Hohoff, B. Hoste, M. A. Jobling, H. J. Kärgel, P. de Knijff, R. Lessig, E. Liebeherr, M. Lorente, B. Martínez-Jarreta, P. Nievas, M. Nowak, W. Parson, V. L. Pascali, G. Penacino, R. Ploski, B. Rolf, A. Sala, U. Schmidt, C. Schmitt, P. M. Schneider, R. Szibor, J. Teifel-Greding, and M. Kayser. Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. *Forensic Sci Int*, 118(2-3):106–113, May 2001.

[97] N. A. Rosenberg, J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman. Genetic structure of human populations. *Science*, 298(5602):2381–2385, Dec 2002. doi: 10.1126/science.1078311. URL `http://dx.doi.org/10.1126/science.1078311`.

[98] N. A. Rosenberg, S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, and M. W. Feldman. Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. *PLoS Genet.*, 1(6):e70, 12 2005. doi: 10.1371/journal.pgen.0010070. URL `http://dx.plos.org/10.1371%2Fjournal.pgen.0010070`.

[99] R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P. Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander, D. Altshuler, and International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–933, Feb 2001.

[100] R. K. Saiki, S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich, and N. Arnheim. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230(4732):1350–1354, 1985. doi: 10.1126/science.2999980. URL `http://www.sciencemag.org/content/230/4732/1350.abstract`.

[101] R. K. Saiki, D. H. Gelfand, S. Stoffel, S .J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239(4839):487–491, 1988. doi: 10.1126/science.2448875. URL `http://www.sciencemag.org/content/239/4839/487.abstract`.

[102] J. N. Sampson, K. K. Kidd, J. R. Kidd, and H. Zhao. Selecting SNPs to identify ancestry. *Ann Hum Genet*, 75(4):539–553, Jul 2011. doi: 10.1111/j.1469-1809.2011.00656.x. URL `http://dx.doi.org/10.1111/j.1469-1809.2011.00656.x`.

[103] J. J. Sanchez, C. Phillips, C. Børsting, K. Balogh, M. Bogus, M. Fondevila, C. D. Harrison, E. Musgrave-Brown, A. Salas, D. Syndercombe-Court, P. M. Schneider, Á. Carracedo, and N. Morling. A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis*, 27(9):1713–1724, 2006. ISSN 1522-2683. doi: 10.1002/elps.200500671. URL http://dx.doi.org/10.1002/elps.200500671.

[104] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12):5463–5467, Dec 1977.

[105] F. R. Santos, A. Pandya, and C. Tyler-Smith. Reliability of DNA-based sex tests. *Nat Genet*, 18(2):103, Feb 1998. doi: 10.1038/ng0298-103. URL http://dx.doi.org/10.1038/ng0298-103.

[106] N. P. C. Santos, E. M. Ribeiro-Rodrigues, A. K. C. Ribeiro-Dos-Santos, R. Pereira, L. Gusmão, A. Amorim, J. F. Guerreiro, M. A. Zago, C. Matte, M. H. Hutz, and S. E. B. Santos. Assessing individual interethnic admixture and population substructure using a 48-insertion-deletion (INSEL) ancestry-informative marker (AIM) panel. *Hum Mutat*, 31(2):184–190, Feb 2010. doi: 10.1002/humu.21159. URL http://dx.doi.org/10.1002/humu.21159.

[107] G. Schatz, E. Haslbrunner, and H. Tuppy. Deoxyribonucleic acid associated with yeast mitochondria. *Biochemical and Biophysical Research Communications*, 15(2): 127 – 132, 1964. ISSN 0006-291X. doi: 10.1016/0006-291X(64)90311-0. URL http://www.sciencedirect.com/science/article/pii/0006291X64903110.

[108] P. M. Schneider. Beyond strs: The role of diallelic markers in forensic genetics. *Transfus Med Hemother*, 39(3):176–180, Jun 2012. doi: 000339139. URL http://dx.doi.org/000339139.

[109] P. M. Schneider and P. D. Martin. Criminal DNA databases: the European situation. *Forensic Sci Int*, 119(2):232–238, Jun 2001.

[110] M. D. Shriver, M. W. Smith, L. Jin, A. Marcini, J. M. Akey, R. Deka, and R. E. Ferrell. Ethnic-affiliation estimation by use of population-specific DNA markers. *Am. J. Hum. Genet.*, 60(4):957–964, Apr 1997.

[111] M. D. Shriver, E. J. Parra, S. Dios, C. Bonilla, H. Norton, C. Jovel, C. Pfaff, C. Jones, A. Massac, N. Cameron, A. Baron, T. Jackson, G. Argyropoulos, L. Jin, C. J. Hoggart, P. M. McKeigue, and R. A. Kittles. Skin pigmentation, biogeographical ancestry and admixture mapping. *Hum Genet*, 112(4):387–399, Apr 2003. doi: 10.1007/s00439-002-0896-y. URL http://dx.doi.org/10.1007/s00439-002-0896-y.

[112] M. Soejima and Y. Koda. Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2. *Int J Legal Med*, 121(1):36–39, Jan 2007. doi: 10.1007/s00414-006-0112-z. URL http://dx.doi.org/10.1007/s00414-006-0112-z.

[113] E. M. Southern. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol*, 98(3):503–517, Nov 1975.

[114] R. Sparkes, C. Kimpton, S. Watson, N. Oldroyd, T. Clayton, L. Barnett, J. Arnold, C. Thompson, R. Hale, J. Chapman, A. Urquhart, and P. Gill. The validation of a 7-locus multiplex STR test for use in forensic casework. (I). Mixtures, ageing, degradation and species studies. *Int J Legal Med*, 109(4):186–194, 1996.

[115] O. Spichenok, Z. M. Budimlija, A. A. Mitchell, A. Jenny, L. Kovacevic, D. Marjanovic, T. Caragine, M. Prinz, and E. Wurmbach. Prediction of eye and skin color in diverse populations using seven snps. *Forensic Sci Int Genet*, 5(5):472–478, Nov 2011. doi: 10.1016/j.fsigen.2010.10.005. URL `http://dx.doi.org/10.1016/j.fsigen.2010.10.005`.

[116] C. Stern. The Hardy-Weinberg Law. *Science*, 97(2510):137–138, Feb 1943. doi: 10.1126/science.97.2510.137. URL `http://dx.doi.org/10.1126/science.97.2510.137`.

[117] M. Stoneking, D. Hedgecock, R. G. Higuchi, L. Vigilant, and H. A. Erlich. Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. *Am J Hum Genet*, 48(2):370–382, Feb 1991.

[118] R. A. Sturm, D. L. Duffy, Z. Z. Zhao, F. P. N. Leite, M. S. Stark, N. K. Hayward, N. G. Martin, and G. W. Montgomery. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *Am J Hum Genet*, 82(2):424–431, Feb 2008. doi: 10.1016/j.ajhg.2007.11.005. URL `http://dx.doi.org/10.1016/j.ajhg.2007.11.005`.

[119] P. Sulem, D. F. Gudbjartsson, S. N. Stacey, A. Helgason, T. Rafnar, K. P. Magnusson, A. Manolescu, A. Karason, A. Palsson, G. Thorleifsson, M. Jakobsdottir, S. Steinberg, S. Pálsson, F. Jonasson, B. Sigurgeirsson, K. Thorisdottir, R. Ragnarsson, K. R. Benediktsdottir, K. K. Aben, L. A. Kiemeney, J. H Olafsson, J. Gulcher, A. Kong, U. Thorsteinsdottir, and K. Stefansson. Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet*, 39(12):1443–1452, Dec 2007. doi: 10.1038/ng.2007.13. URL `http://dx.doi.org/10.1038/ng.2007.13`.

[120] R. Szibor. X-chromosomal markers: past, present and future. *Forensic Sci Int Genet*, 1(2):93–99, Jun 2007. doi: 10.1016/j.fsigen.2007.03.003. URL `http://dx.doi.org/10.1016/j.fsigen.2007.03.003`.

[121] R. Szibor, M. Krawczak, S. Hering, J. Edelmann, E. Kuhlisch, and D. Krause. Use of X-linked markers for forensic purposes. *Int J Legal Med*, 117(2):67–74, Apr 2003. doi: 10.1007/s00414-002-0352-5. URL `http://dx.doi.org/10.1007/s00414-002-0352-5`.

[122] G. A. Thorisson and L. D. Stein. The SNP consortium website: past, present and future. *Nucleic Acids Res*, 31(1):124–127, Jan 2003.

[123] C. Tian, P. K. Gregersen, and M. F. Seldin. Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet*, 17(R2):R143–R150, Oct 2008. doi: 10.1093/hmg/ddn268. URL http://dx.doi.org/10.1093/hmg/ddn268.

[124] G. Tully. Genotype versus phenotype: human pigmentation. *Forensic Sci Int Genet*, 1 (2):105–110, Jun 2007. doi: 10.1016/j.fsigen.2007.01.005. URL http://dx.doi.org/10.1016/j.fsigen.2007.01.005.

[125] P. A. Underhill and T. Kivisild. Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu Rev Genet*, 41:539–564, 2007. doi: 10.1146/annurev.genet.41.110306.130407. URL http://dx.doi.org/10.1146/annurev.genet.41.110306.130407.

[126] C.o.t.E. Union. Enfopol 287 crimorg 170. 15870/09:1–7., 2009.

[127] P. Valverde, E. Healy, I. Jackson, J. L. Rees, and A. J. Thody. Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nat Genet*, 11(3):328–330, Nov 1995. doi: 10.1038/ng1195-328. URL http://dx.doi.org/10.1038/ng1195-328.

[128] M. van Oven and M. Kayser. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat*, 30(2):E386–E394, Feb 2009. doi: 10.1002/humu.20921. URL http://dx.doi.org/10.1002/humu.20921.

[129] E. von Dungern and L. Hirschfeld. Ueber Vererbung gruppenspezifischer Strukturen des Blutes. *Z Immun Forsch.*, 6:284–292, 1910.

[130] S. Walsh, A. Lindenbergh, S. B. Zuniga, T. Sijen, P. de Knijff, M. Kayser, and K. N. Ballantyne. Developmental validation of the irisplex system: determination of blue and brown iris colour for forensic intelligence. *Forensic Sci Int Genet*, 5(5):464–471, Nov 2011. doi: 10.1016/j.fsigen.2010.09.008. URL http://dx.doi.org/10.1016/j.fsigen.2010.09.008.

[131] S. Walsh, F. Liu, K. N. Ballantyne, M. van Oven, O. Lao, and M. Kayser. 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(3):170–180, Jun 2011. doi: 10.1016/j.fsigen.2010.02.004. URL http://dx.doi.org/10.1016/j.fsigen.2010.02.004.

[132] S. Walsh, F. Liu, A. Wollstein, L. Kovatsi, A. Ralf, A. Kosiniak-Kamysz, W. Branicki, and M. Kayser. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Sci Int Genet*, 7(1):98–115, Jan 2013. doi: 10.1016/j.fsigen.2012.07.005. URL http://dx.doi.org/10.1016/j.fsigen.2012.07.005.

[133] J. L. Weber and P. E. May. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet*, 44(3):388–396, Mar 1989.

[134] J. L. Weber, D. David, J. Heil, Y. Fan, C. Zhao, and G. Marth. Human diallelic insertion/deletion polymorphisms. *Am. J. Hum. Genet.*, 71(4):854–862, Oct 2002. doi: 10.1086/342727. URL http://dx.doi.org/10.1086/342727.

[135] P. F. Wieacker, I. Knoke, and S. Jakubiczka. Clinical and molecular aspects of androgen receptor defects. *Exp Clin Endocrinol Diabetes*, 106(6):446–453, 1998. doi: 10.1055/s-0029-1212014. URL http://dx.doi.org/10.1055/s-0029-1212014.

[136] N. Yang, H. Li, L. A. Criswell, P. K. Gregersen, M. E. Alarcon-Riquelme, R. Kittles, R. Shigeta, G. Silva, P. I. Patel, J. W. Belmont, and M. F. Seldin. Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. *Hum Genet*, 118(3-4):382–392, Dec 2005. doi: 10.1007/s00439-005-0012-1. URL http://dx.doi.org/10.1007/s00439-005-0012-1.

[137] D. Zaumsegel, M. A. Rothschild, and P. M. Schneider. SNPs for the analysis of human pigmentation genes - A comparative study. *Forensic Science International: Genetics Supplement Series*, 1(1):544 – 546, 2008. ISSN 1875-1768. doi: 10.1016/j.fsigss.2007.11.016. URL http://www.sciencedirect.com/science/article/pii/S1875176808000401.

[138] D. Zaumsegel, M. A. Rothschild, and P. M. Schneider. A 21 marker insertion deletion polymorphism panel to study biogeographic ancestry. *Forensic Sci Int Genet*, 7(2):305–312, Feb 2013. doi: 10.1016/j.fsigen.2012.12.007. URL http://dx.doi.org/10.1016/j.fsigen.2012.12.007.

7 Lebenslauf

Personal Data

Place and Date of Birth	Troisdorf, Germany, April 03rd 1978
Place of Residence	Cologne, Germany

Research Experience

Current *May 2013*	Institute for Legal Medicine, Medical Faculty, University of Cologne, Germany
Aug 2005 - Sep 2006	Undergraduate student project Institute for Physiological Chemistry, Medical Faculty, University of Utrecht, The Netherlands Topic: "Applications of Biotinylation Tagging in *Saccharomyces Cervisiae*"
Jul 2004 - Jul 2005	Undergraduate student project Institute for Molecular Genetics, Biological Faculty, University of Utrecht, The Netherlands Topic: "Genetic Mapping of the Loci Conferring Broad Resistance to *Hyaloperonospora parasitica* in *Arabidopsis thaliana* accession C24"
Mar 2003 - Aug 2003	Bachelor thesis project Institute for Clinical Research and Development (ikfe GmbH), Mainz, Germany Topic: "Establishment of Methods for High Throughput SNP Screening in Genes for Drug Metabolising Enzymes with Real-Time PCR"

Education

October 2006	Master of Science in Biomedical Sciences Masters Course "Cancer Genomics and Developmental Biology" **University of Utrecht**, Utrecht, The Netherlands Thesis: "Dynamics of Histone Modifications in Differentiating Embryonic Stem Cells"
September 2003	Bachelor of Science in Biology **Bonn-Rhein-Sieg University of Applied Sciences**, Rheinbach, Germany Thesis: "Establishment of Methods for High Throughput SNP Screening in Genes for Drug Metabolising Enzymes with Real-Time PCR"
June 2003	Bachelor of Science in Biosciences with Biomedical Sciences Second Class Honours (1st Division) **The Robert Gordon University**, Aberdeen, Scotland
June 1999	High School Graduation (German Abitur) **Heinrich-Böll-Gymnasium**, Troisdorf, Germany
1995 - 1996	High School Student Exchange **Lilydale High School**, Lilydale, Melbourne, Australia