7

DNA Biometrics

Masaki Hashiyada
Division of Forensic Medicine, Department of Public Health and Forensic Medicine,
Tohoku University Graduate School of Medicine
Japan

1. Introduction

The biometric authentication technologies, typified by fingerprint, face recognition and iris scanning, have been making rapid progress. Retinal scanning, voice dynamics and handwriting recognition are also being developed. These methods have been commercialized and are being incorporated into systems that require accurate on-site personal authentication. However, these methods are based on the measurement of similarity of feature-points. This introduces an element of inaccuracy that renders existing technologies unsuitable for a universal ID system. Among the various possible types of biometric personal identification system, deoxyribonucleic acid (DNA) provides the most reliable personal identification. It is intrinsically digital, and does not change during a person's life or after his/her death. This chapter addresses three questions: First, how can personally identifying information be obtained from DNA sequences in the human genome? Second, how can a personal ID be generated from DNA-based information? And finally, what are the advantages, deficiencies, and future potential for personal IDs generated from DNA data (DNA-ID)?

2. Human identification based on DNA polymorphism

A human body is composed of approximately of 60 trillion cells. DNA, which can be thought of as the blueprint for the design of the human body, is folded inside the nucleus of each cell. DNA is a polymer, and is composed of nucleotide units that each has three parts: a base, a sugar, and a phosphate. The bases are adenine, guanine, cytosine and thymine, abbreviated A, G, C and T, respectively. These four letters represent the informational content in each nucleotide unit; variations in the nucleotide sequence bring about biological diversity, not only among human beings but among all living creatures. Meanwhile, the phosphate and sugar portions form the backbone structure of the DNA molecule. Within a cell, DNA exists in the double-stranded form, in which two antiparallel strands spiral around each other in a double helix. The bases of each strand project into the core of the helix, where they pair with the bases of the complementary strand. A pairs strictly with T, and C with G (Alberts, 2002; Watson, 2004).
Within human cells, DNA found in the nucleus of the cell (nuclear DNA) is divided into chromosomes. The human genome consists of 22 matched pairs of autosomal chromosomes and two sex-determining chromosomes, X and Y. In other words, human cells contain 46 different chromosomes. Males are described as XY since they possess a single copy of the X

chromosome and a single copy of the Y chromosome, while females possess two copies of the X chromosome and are described as XX.

The regions of DNA that encode and regulate the synthesis of proteins are called genes; these regions consist of exons (protein-coding portions) and introns (the intervening sequences) and constitute approximately 25% of the genome (Jasinska & Krzyzosiak, 2004). The human genome contains only 20,000–25,000 genes (Collins et al., 2004; Lander et al., 2001; Venter et al., 2001). Therefore, most of the genome, approximately 75%, is extragenic. These regions are sometimes referred to as 'junk' DNA; however, recent research suggests that they may have other essential functions. Markers commonly used to identify individual human beings are usually found in the noncoding regions, either between genes or within genes (i.e., introns).

2.1 Sort tandem repeat (STR)

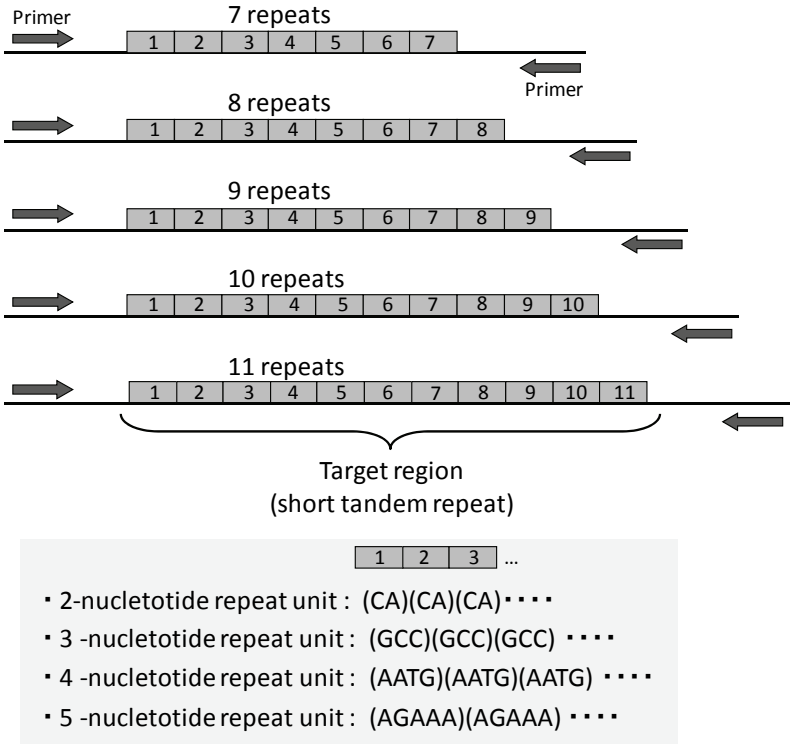


Fig. 1. The structure of Short Tandem Repeat (STR)

In the extragenic region of eukaryotic genome, there are many repeated DNA sequences (approximately 50% of the whole genome). These repeated DNA sequences come in all sizes, and are typically designated by the length of the core repeat unit and either the number of contiguous repeat units or the overall length of the repeat region. These regions are referred to as satellite DNA (Jeffreys et al., 1995). The core repeat unit for a medium-length repeat, referred to as a minisatellite or VNTR (variable number of tandem repeats), is in the range of approximately 8–100 bases in length (Jeffreys et al., 1985). DNA regions with

repeat units that are 2–7 base pairs (bp) in length are called microsatellites, simple sequence repeats (SSRs), or most commonly short tandem repeats (STRs) (Clayton et al. ,1995; Hagelberg et al., 1991;Jeffreys et al., 1992)(Fig. 1). STRs have become popular DNA markers because they are easily amplified by the polymerase chain reaction (PCR) and they are spread throughout the genome, including both the 22 autosomal chromosomes and the X and Y sex chromosomes. The number of repeats in STR markers can vary widely among individuals, making the STRs an effective means of human identification in forensic science (Ruitberg et al., 2001). The location of an STR marker is called its "locus." The type of STR is represented by the number of repeat called 'allele' which is taken from biological father and mother. When an individual has two copies of the same allele for a given marker, they are homozygous; when they have two different alleles, they are heterozygous.

2.1.1 DNA sample collection

DNA can be easily obtained from a variety of biological sources, not only body fluid but also nail, hair and used razors (Anderson et al., 1999; Lee et al., 1998; Lee & Ladd, 2001). For biometric applications, a buccal swab is the most simple, convenient and painless sample collection method (Hedman et al., 2008). Buccal cell collection involves wiping a small piece of filter paper or a cotton swab against the inside of the subject's cheek, in order to collect shed epithelial cells. The swab is then air dried, or can be pressed against a treated collection card in order to transfer epithelial cells for storage purposes.

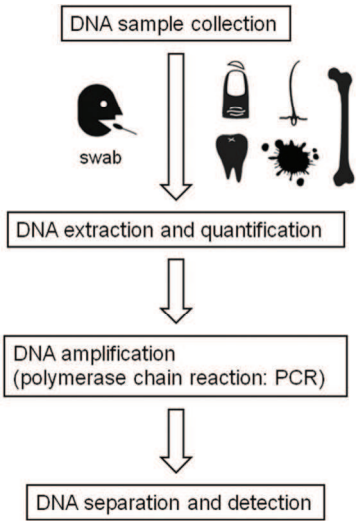


Fig. 2. The flow of DNA polymorphism analysis

2.1.2 DNA extraction and quantification

There are many methods available for extracting DNA (Butler, 2010). The choice of which method to use depends on several factors, especially the number of samples, cost, and speed. Extraction time is the critical factor for biometric applications. The author has already reported the "5-minute DNA extraction" using an automated procedure (Hashiyada, 2007a). The use of large quantities of fresh buccal cells made it possible to extract DNA in a short time.

In forensic cases, DNA quantitation is an important step (Butler, 2010). However, this step can be omitted in biometrics because a relatively large quantity of DNA can be recovered from fresh buccal swab samples.

2.1.3 DNA amplification (polymerase chain reaction: PCR)

The field of molecular biology has greatly benefited from the discovery of a technique known as the polymerase chain reaction, or PCR (Mullis et al., 1986; Mullis & Faloona, 1987; Saiki et al., 1986). First described in 1985 by Kary Mullis, who received the Novel Prize in Chemistry in 1993, PCR has made it possible to make hundreds of millions of copies of a specific sequence of DNA in a few hours. PCR is an enzymatic process in which a specific region of DNA is replicated over and over again to yield many copies of a particular sequence. This molecular process involves heating and cooling samples in a precise thermal cycling pattern for approximately 30 cycles. During each cycle, a copy of the target DNA sequence is generated for every molecule containing the target sequence. In recent years, it has become possible to PCR amplify 16 STRs, including the gender assignment locus called 'amelogenin,' in one tube (Kimpton et al., 1993; Kimpton et al., 1996). Such multiplex PCR is enabled by commercial typing kits, such as AmpFlSTR® Identifiler® (Applied Biosystems, Foster City, CA, USA) and PowerPlex® 16 (Promega, Madison, WI, USA).

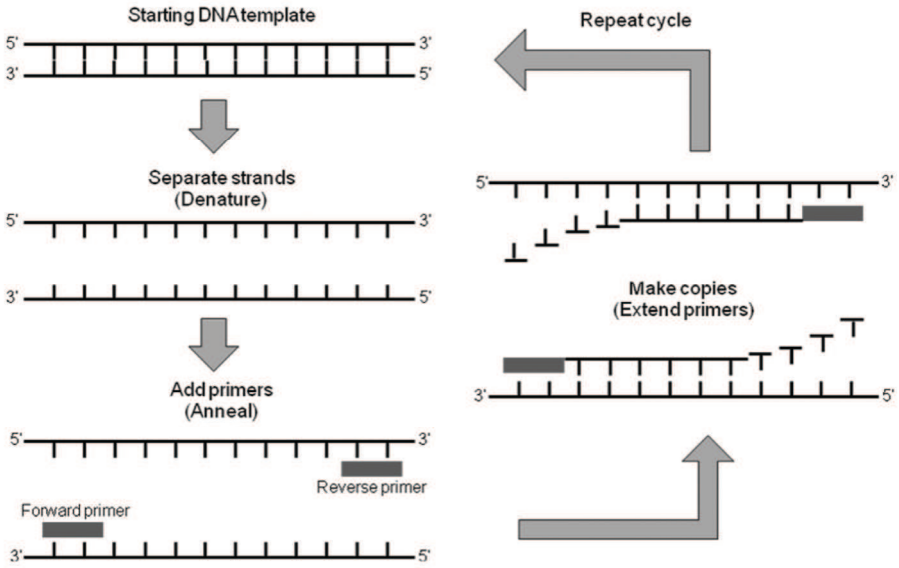


Fig. 3. DNA amplification with polymerase chain reaction (PCR)

2.1.4 DNA separation and detection

After STR polymorphisms have been amplified using PCR, the length of products must be measured precisely; some STR alleles differ by only 1 base-pair. Electrophoresis of the PCR products through denaturing polyacrylamide gels can be used to separate DNA molecules from 20–500 nucleotides in length with single base pair resolution (Slater et al., 2000). Recently, the fluorescence labelling of PCR products followed by multicolour detection has

been adopted by the forensic science field. Up to five different dyes can be used in a single analysis. Electrophoresis platforms have evolved from slab-gels to capillary electrophoresis (CE), which use a narrow glass filled with an cross-linked polymer solution to separate the DNA molecules (Butler et al., 2004). After data collection by the CE, the alleles (i.e., the type or the number of STR repeat units), are analyzed by the software that accompanies the CE machine.
It takes around four hours, starting with DNA extraction, to obtain data from 16 STRs including the sex determination locus.

2.2 Single nucleotide polymorphism (SNP)

The simplest type of polymorphism is the single nucleotide polymorphism (SNP), a single base difference at a particular point in the sequence of DNA (Brookes, 1999). SNPs normally have just two alleles, e.g., one allele is a cytosine (C) and the other is a thymine (T) (Fig. 4). SNPs therefore are not highly polymorphic and do not possess ideal properties for DNA polymorphism to be used in forensic analysis. However, SNPs are so abundant throughout the genome that it is theoretically possible to type hundreds of them. Furthermore, sample processing and data analysis may be more fully automated because size-based separation is not required. Thus, SNPs are prospective new bio-markers in clinical medicine (Sachidanandam et al., 2001; Stenson et al., 2009).

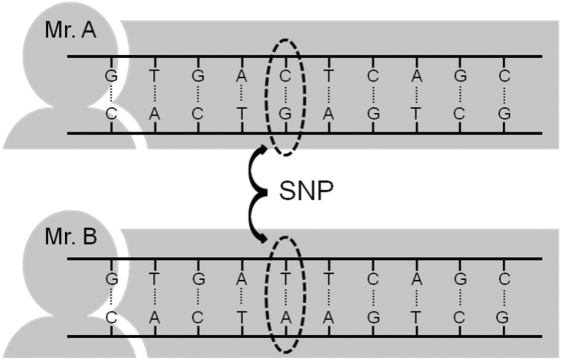


Fig. 4. The schema of Single nucleotide polymorphism (SNP)

2.2.1 SNP detection methods

Several SNP typing methods are available, each with its own strengths and weaknesses, unlike the STR analysis (Butler, 2010). In order to achieve the same power of discrimination as that provided by STRs, it is necessary to analyse many more SNPs. 40 to 50 SNPs must be analyzed in order to obtain reasonable powerful discrimination and define the unique profile of an individual (Gill, 2001). Importantly, however, we can count on the development of new SNP detection technologies, capable of high-throughput analysis, in the near future.

2.3 Lineage markers

Autosomal DNA markers are shuffled with each generation, which means that half of an individual's genetic information comes from his or her father and the other half from his or her mother. However, the Y chromosome (Chr Y) and mitochondrial DNA (mtDNA)

markers are called "lineage markers" because they are passed down from generation to generation without changing (except for mutational events). Maternal lineages can be followed using mtDNA sequence information (Anderson et al., 1981; Andrews et al., 1999) and whereas paternal lineages can be traced using Chr Y markers (Jobling & Tyler–Smith, 2003; Kayser et al., 2004). The analysis of lineage markers does not have the discriminatory power of autosomal markers. Even so, there are some features of both Chr Y and mtDNA that make them valuable forensic tools.

3. DNA polymorphism for biometric source

The most commonly studied or implemented biometrics are fingerprinting, face, iris, voice, signature, retina and the patterns of vein and hand geometry (Shen & Tan, 1999; Vijaya Kumar et al., 2004). No one model is best for all situations. In addition, these technologies are based on the measurement of similarity of features. This introduces an element of inaccuracy that renders the existing technologies unsuitable for a universal ID system. However, DNA polymorphism information, such as STRs and SNPs, could provide the most reliable personal identification. This data can be precisely defined the most minute level, is intrinsically digital, and does not change during a person's life or after his/her death. Therefore, DNA identification data is utilized in the forensic sciences. On the negative side, the biggest problem in using DNA is the time required for the extraction of nucleic acid and the evaluation of STR or SNP data. In addition, there are several other problems, such as the high cost of analysis, issues raised by monozygotic twins, and ethical concerns.
This section describes a method for generation of DNA personal ID (DNA-ID) based on STR and SNP data, specifically. In addition, by way of example, the author proposes DNA INK for authentic security.

3.1 DNA personal ID using STR system

We will refer to repeat counts of alleles obtained by STR analysis, as described in section 2.1, as (*j*, *k*). Each locus is associated with two alleles with distinct repeat counts (*j*, *k*), as shown in Fig. 2: one allele is inherited from the father, and the other from the mother. Before (*j*, *k*) can be applied to a DNA personal ID, it is necessary to statistically analyze how the distribution of (*j*, *k*) varies at a given locus based on actual data.
We can generate a DNA-ID, α_X, that includes allelic information about STR loci. The loci are incorporated in the following sequence. The repeat counts for the pair of alleles at each locus are arranged in ascending order.
Step 1. Measure the STR alleles at each locus.
Step 2. Obtain STR count values for each locus; express these in ascending order.

$$L : j \| k, j \le k$$

Depending on the measurement, the same person's STR count may appear as (*j*, *k*) or (*k*, *j*). Therefore, *j* and *k* are expressed in an ascending order, i.e., using (*j*, *k* | *j* ≤ *k*), in order to establish a one-to-one correspondence for each individual. This step is referred to as a ordering operation.
Step 3. Generate a DNA-ID α_X according to the following series, L_i (*j*, *k*):

$$\alpha_X = L_1 \| L_2 \| L_3 \| \ldots \| L_n$$

where L_i indicates the ith STR count (j, k).
For example, suppose that Mr. M has the following alleles at the respective loci;

$$\begin{aligned}\alpha_X &= \text{D3S1358}\| \text{D13S317}\| \text{D18S51}\| \text{D21S11}\| \ldots\| \text{D16S539} \\ &= (12,14)\| (8,11)\| (13,15)\| (29,32.2)\| \ldots\| (10,10)\end{aligned}$$

The α_X was thus defined as follows.

$$\alpha_X = 1214811131529322 \ldots\ldots 1010$$

When the STR number of an allele had a fractional component, such as allele32.2 in D21S11, the decimal point was removed, and all of the numbers, including those after the decimal point, were retained.
Finally, α_X is generated number with several tends of digits, and becomes a personal identification information that is unique with a certain probability predicted by statistical and theoretical analysis.

3.1.1 Establishment of the identification format

Because α_X contains personal STR information, it must be encrypted to protect privacy. This can be achieved using a one-way function that also reduces the data length of the DNA-ID. This one-way function, the secure hash algorithm-1 (SHA-1), produces an ID with a length δ_X of 160 bits, according to the following transformation:

$$\delta_X = h(\alpha_X)$$

3.2 Statistical and theoretical analysis of DNA-ID

3.2.1 Matching probability at locus L

The probability that a STR allele (j, k) at locus L will occur in this combination is denoted as p_{jk}. The individual occurrence probabilities of j and k are denoted as p_j and p_k, respectively. Here, j and k are sequenced in ascending order to make the choice of generated ID unambiguous, using the STR analysis system described above. After the STR analysis, the probability that (j, k) occurs is p_{jk} plus the probability p_{kj} that (k, j) occurs. The reason for this is as follows. Even if (k, j) occurs in the same person during measurement, it is treated as (j, k) by rewriting it as (j, k) if $k > j$.
Therefore, p_{jk} is expressed as follows when $j \neq k$ ($j < k$):

$$p_{jk} = p_j \bullet p_k + p_k \bullet p_j$$

$$= 2p_j p_k$$

If $j = k$,

$$p_{jk} = p_j \bullet p_j$$

3.2.2 Probability of a match between any two persons' DNA-ID

Probability p that the STR count at the same locus is identical for any two persons can be expressed as follows:

$$\text{When } j = k,\ \sum_{j=1}^{m}\left(p_j \cdot p_k\right)^2$$

$$\text{When } j \neq k,\ \sum_{1\le j<k\le m}^{m}\left(2p_j \cdot p_k\right)^2$$

$$\therefore p = \sum_{j=1}^{m}\left(p_j\right)^4 + \sum_{1\le j<k\le m}^{m} 4\left(p_j \cdot p_k\right)^2$$

Here, m is the upper limit of j and k, and the information reported so far indicates m= 60. Next, a determination is made of the DNA-ID matching probability p_n, where n loci were used to generate the ID. The probability that the STR counts at the i^{th} locus will match for any two persons is denoted as p_i. When n loci are used, the probability p_n that the DNA-IDs of any two persons will match (the DNA-ID matching probability) is as follows:

$$p_n = \prod_{i=1}^{n} p_i$$

Here, it is assumed that there is no correlation among the STR loci.

3.2.3 Verification using validation experiment (STR)

As a validation experiment, we studied the genotype and distribution of allele frequencies at 18 STRs in 526 unrelated Japanese individuals. Data was obtained using three commercial STR typing kits: PowerPlex™ 16 system (Promega), PowerPlex SE33 (Promega), and AmpFlSTR Identifiler™ (Applied biosystems) (Hashiyada, 2003a; 2003b). Information about the 18 target STRs is described in Table 1.

Step 1. Perform DNA extraction, PCR amplification and STR typing

Step 2. Perform the exact test (the data were shuffled 10,000 times), the homozygosity, and likelihood ratio tests using STR data for each STR locus in order to evaluate Hardy–Weinberg equilibrium (HWE). HWE provides a simple mathematical representation of the relationship among genotype and allele frequencies within an ideal population, and is central to forensic genetics. Importantly, when a population is in HWE, the genotype frequencies can be predicted from the allele frequencies.

Step 3. Calculate parameters, the matching probability, the expected and observed heterozygosity, the power of discrimination, the polymorphic information content, the mean exclusion chance, in order to estimate the polymorphism at each STR locus.

There are some loci on the same chromosomes (chr) such as D21S11 and Penta D on chr 21, D5S818 and CSF1PO on chr 5, and TPOX and D2S1338 on chr 2. No correlation was found between any sets of loci on the same chromosome, which means they are statistically independent. In addition, the statistical data for the 18 analyzed STRs, excluding the Amelogenin locus, were analyzed and showed a relatively high rate of matching probability; no significant deviation from HWE was detected. The combined mean exclusion chance was 0.9999998995 and the combined matching probability was 1 in 9.98×10^{21}, i.e., 1.0024×10^{-22}. These values were calculated using polymorphism data from Japanese subjects; it is likely that different values would be obtained using data compiled from different ethnic groups, e.g., Caucasian or African.

Locus	Chromosome Location	Repeat Motif*	Locus	Chromosome Location	Repeat Motif*
TPOX	2 q 25.3	GAAT	TH01	11 p 15.5	TCAT
D2S1338	2 q 35	TGCC/TTCC	VWA	12 p 13.31	TCTG/TCTA
D3S1358	3 p 21.31	TCTG/TCTA	D13S317	13 q 31.1	TATC
FGA	4 q 31.3	CTTT/TTCC	Penta E	15 q 26.2	AAAGA
D5S818	5 q 23.2	AGAT	D16S539	16 q 24.1	GATA
CSF1PO	5 q 33.1	TAGA	D18S51	18 q 21.33	AGAA
SE33	6 q 14	AAAG	D19S433	19 q 12	AAGG/TAGG
D7S820	7 q 21.11	GATA	D21S11	21 q 21.1	TCTA/TCTG
D8S1179	8 q 24.13	TCTA/TCTG	Penta D	21 q 22.3	AAAGA

* Two types of motif means a compound or complex repeat sequence

Table 1. Information about autosomal STR loci

3.2.4 The "Birthday Paradox" of DNA-ID

In principle, the low matching probability of STR-based IDs would allow absolute and unequivocal discrimination between individuals. However, if STRs are to be used as an authentication system in our society, we must investigate the probability of two or more randomly selected people having an identical DNA- ID. The most well-known simulation of this probability is "the birthday paradox". Of 40 students in a class, the probability that at least two students have the same birthday is approximately 0.9. This result seems counterintuitive, and is called a "paradox," because for any single pair of students, the probability that they have the same birthday is 1/365 (0.0027). The paradox arises when we forget to consider that we are selecting samples randomly out of the members in a group.

In two randomly selected individuals, the probability that one STR locus is different and that all STR loci are identical is $(1\text{-}P_M)^{L(L\text{-}1)/2}$ and $1\text{-}(1\text{-}P_M)^{L(L\text{-}1)/2}$, respectively, where L is the population size. However, the formula, $1\text{-}(1\text{-}P_M)^{L(L\text{-}1)/2}$, is beyond the ability of personal computers, so we use the expected value, $L(L\text{-}1)/2 \cdot P_M$, to estimate two persons having the same STR genotype. This formula can use an approximate value of $1\text{-}(1\text{-}P_M)^{L(L\text{-}1)/2}$. This is because L^2 is much smaller than $1/ P_M$ when L is small, and because $1\text{-}(1\text{-}P_M)^{L(L\text{-}1)/2}$ is smaller than $L(L\text{-}1)/2 \cdot P_M$ when L is not small. In this report, the value, $L(L\text{-}1)/2 \cdot P_M$, is defined as the practical matching probability (P_{PM}). The matching probability (P_M) for 18 STRs is 1.0024×10^{-22}, as described above. When P_{PM} multiplied by the population size is less than 1, each person in the population could have a unique DNA-ID. Therefore, when using 18 loci, a population of tens of millions could be expected to include pairs of individuals with identical STR alleles. If the frequencies of STR alleles are similar among all ethnic groups, each person in Japan (or the world) could have a unique DNA-ID if the P_{PM} of the STR system were approximately 10^{-24} and 10^{-30}, respectively. As the number of people in a community increases, the more the practical matching probability increases.

This number can be applied for unrelated persons; however, we also need to consider P_{PM} between related individuals. For instance, between two first cousins, if 41 STR loci are analyzed, we can obtain a unique DNA-ID. In addition, discrimination between half siblings requires analysis of 57 STR loci guarantee a unique DNA-ID. Thus, when using DNA identification systems such as STR systems for DNA-personal-IDs, the P_{PM} should be considered for both related and unrelated individuals (Hashiyada, 2007b).

3.3 DNA personal ID using SNP system

The vast majority of SNPs are biallelic, meaning that they have two possible alleles and therefore three possible genotypes. For example, if the alleles for a SNP locus are R and S (where 'R' and 'S' could represent a A(adenine), G(guanine), C(cytosine) and T(thymine) nucleotide), three possible genotypes would be RR, RS (SR) or SS. Because a single biallelic SNP by itself yields less information than a multiallelic STR marker, it is necessary to analyze a larger number of SNPs in order to obtain a reasonable power of discrimination to define a unique profile. Computational analysis have shown that on average, 25 to 45 SNP loci are needed in order to yield equivalent random match probabilities comparable to those obtained with the 13 core STR loci that have been adopted by the FBI's DNA database (COmbined DNA Index System, CODIS).

The steps of creating a DNA-ID using SNPs are as follows;

Step 1. Define alleles 1 and 2 for each SNP locus. Since DNA has a double helix structure, the single nucleotide polymorphism of A or G is the same polymorphism of T or C, respectively (Fig. 4). In other words, it is important to specify which strand of the double helix is to be analyzed, and to define allele 1 and allele 2 at the outset.

Step 2. Analyze the SNP loci and place them in the following order.

$$\text{L : allele 1} \parallel \text{allele 2}$$

Step 3. Generate the DNA-ID α_X according to the following series of L_i (allele1, allele2):

$$\alpha_X = L_1 \parallel L_2 \parallel L_3 \parallel \ldots \parallel L_n$$

where L_i indicates the i^{th} SNP nucleotide (allele1, allele2).

For example, suppose that a person has the following alleles at the respective loci;

$$\alpha_X = \text{SNP 1} \parallel \text{SNP 2} \parallel \text{SNP 3} \parallel \text{SNP 4} \parallel \ldots \parallel \text{SNP 50}$$

$$= (A,A) \parallel (C,T) \parallel (T,C) \parallel (C,C \parallel \ldots\ldots \parallel (G,A)$$

Then α_X would be defined as follows.

$$\alpha_X = \text{AACTTCCC}\ldots\ldots\text{GA}$$

Next, the four types of nucleotide, A, G, C and T, are translated into binary notation.

$$A=00,\ G=01,\ C=10,\ T=11$$

Finally, the α_X is described as a string of 100 bits (digits of value 0 or 1).

$$\alpha_X = 0000101111101010\ldots\ldots0100$$

This α_X must be encrypted for privacy protection using the secure hash algorithm-1 (SHA-1) for the same reasons as described above for STRs. The resulting DNA-ID (SNP) has a length δ_X of 160 bits, according to the following transformation:

$$\delta_X = h(\alpha_X)$$

3.3.1 Verification using validation experiment (SNP)

As a validation experiment, the author analyzed 120 autosomal SNPs in 100 unrelated Japanese subjects using the TaqMan® method (Applied Biosystems), and built a Japanese SNP database for identification. Although several SNPs were located on the same autosomal chromosome, no correlation was found between alleles at any SNP loci. Furthermore, no significant deviation from Hardy–Weinberg Equilibrium (HWE) was detected. The macthing probability (MP) of each SNP ranged from 0.375–0.465 (Hashiyada, 2007a). The MP for 41 SNPs (3.63×10^{-18}), which have high MP in each loci, was very similar to the MPs obtained with the current STR multiplex kits, PowerPlex™ 16 System(Promega) and AmpFlSTR Identifiler (Applied Biosystems), which were 5.369×10^{-18} and 1.440×10^{-17}, respectively in Japanese population.

3.4 Rapid analysis system of SNP

A reduction of the time required for DNA analysis is necessary in order to make practical use of DNA biometrics. In the STR system, it is difficult to decrease the analysis time because it is necessary to perform electrophoresis after PCR amplification. From DNA extraction to STR typing, the entire process takes 4–5 hours. However, there are many methods for analyzing SNPs that do not demand such a lengthy process. The author developed the SNP typing methodology using the modified TaqMan® method, which is capable of amplifying the DNA and typing the SNPs at the same time. The author modified the number of PCR cycles and the annealing/extension time, and selected SNP loci that yield successful results under the modified PCR conditions. This new method is capable of detecting and typing 96 SNPs within 30 minutes (Hashiyada et al., 2009).

3.5 DNA INK

In this paragraph, the author demonstrates an example of an application of STR polymorphism information, specifically the authentication of rare or expensive goods using the DNA-ID. The author outlines the development of biometric ink containing DNA whose sequence is based on personal STR information. The "DNA INK" is made of synthetic DNA and printing ink.

Step 1. Perform STR analysis by the method described above.

Step 2. Generate the DNA-ID, δ_X, consisting of 160 bits, as described above.

Step 3. Extract one-quarter of the data in the DNAI-ID (δ_X) in order to reduce costs and improve practicality. The original 160-bit length was defined as

$$\delta_{X\,i} = \delta_{X\,1} | \ \delta_{X2} | \ \delta_{X\,3} | \ \delta_{X4} |$$

where $\delta_{X\,1}$, δ_{X2}, $\delta_{X\,3}$ and δ_{X4} refer to the identification, ID, containing of the first, second, third and fourth 40 bits of δ_X. Each set of 8 data bits is extended by two redundant bits known as the shift and check bits, which serve not only as check but also as limiting factors in the latter stages of DNA sequence generation. These limiting factors are necessary in DNA sequence analysis in order to exclude five or more repetitions of the same base. The extracted 40-bit data as follows;

$$\delta_{X\,1} = 1001100110011101011010101001011110100010$$

$$\delta_{X\,1} = 10011001\ [\underline{10}]\ 10011101\ [\underline{00}]\ 01101010\ [\underline{01}]\ 10010111\ [\underline{00}]\ 10100010\ [\underline{11}]$$

(Shift and check bits show as square brackets with underlines.)

Step 4. Transform the bit series generated above into base sequences according to the following scheme. We called this step the "Encodeed Base Array" method.

00=A(adenine), 01=C(cytosine), 10=G(guanine), 11=T(thymine)

δ_{X1} =10011001 [10]10011101 [00] 01101010 [01] 10010111 [00] 10100010 [11]

=GCGC [G] GCTC [A] CGGG [C] GCCT [A] GGAG [T]

Step 5. Define the identification data format by adding a header (H, 10 bits) and a serial number (N, 30 bits) to δ_{X1} (40 + 10 = 50 bits). The resulting DNA sequence, consisting of H (5-bp), N (15-bp) and δ_{X1} (25-bp) would then be flanked by two 20-bp–long primer sequences. This synthetic DNA could be amplified by PCR, and only those who know the primer sequences would be able to analyze the intervening sequence. Figure 5 shows the structure of the 85-bp synthetic DNA sequence.

Step 6. Synthesize the complementary strand. Synthetic single-strand DNA is more economical to produce than double-strand DNA, but much less physically stable; therefore, double-strand PCR-amplified DNA should be used for incorporation into the DNA ink.

Step 7. Mix 3 mg of double-strand DNA with 100 ml of ink. The ink itself is composed of a colorless transparent pigment, so that it is invisible to the naked eye, but contains an IR color former that enables easy detection of the printed mark. In addition, add dummy DNA in order to make the DNA-ID sequence difficult to analyze by someone who does not know the primer sequences.

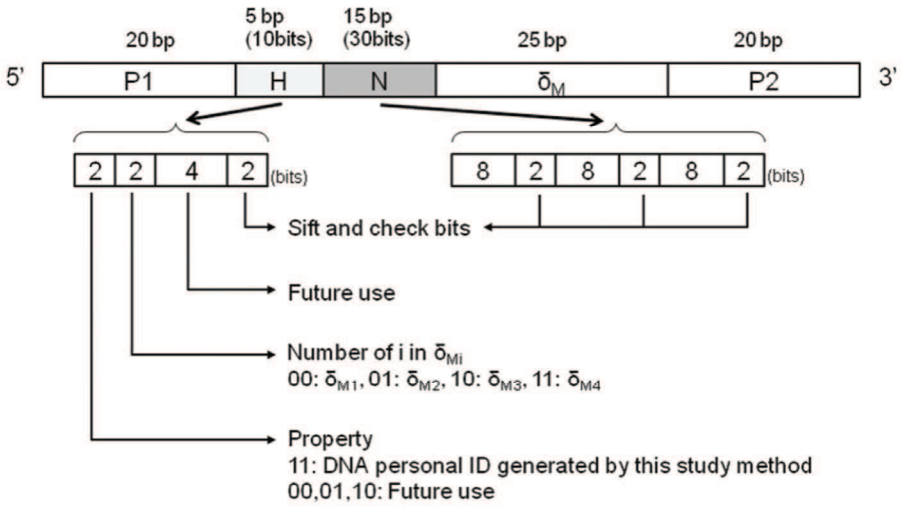


Fig. 5. Sequence structure of the 85-bp single-strand DNA-ID
P1, P2: Primer sequences are designed so as not to anneal to the human genome
H: Header, N: Serial number

The several types of resistance tests, by heat, acids, alkalis, alcohol, ultraviolet (UV) and sunlight, were used to ascertain the durability of DNA ink for practical use. Samples printed using DNA ink were covered with zinc oxide (ZnO) on the surface in order to enhance resistance to UV light, which is the major cause of DNA degradation.
The target DNA sequence was detected successfully in all resistance tests except for the UV exposure test. However, the durability improved when the ink was covered by ZnO, allowing successful amplification even after 40 hours of UV exposure. Finally, the DNA ink was proved as a sort of biological memory which could print the polymorphism information created by DNA, on the surface of everything excluding the air and water.

4. Problems of DNA biometrics

There can be no doubt that DNA-ID is potentially useful as a biometric. It has many advantages, including accuracy, strictness, discriminatory power (and ease of increasing this power), and the ability to use the same analysis platform all over the world. However, DNA polymorphism information is not widely used in biometrics at this point. The weak points of DNA-ID are discussed below.

4.1 Time required for DNA analysis

The most serious flaw is that DNA analysis is time-consuming compared to other authentication methods. It takes at least 4 hours to get STR identification data by common methods used in forensic science. Most of the time required for DNA analysis is taken up by PCR amplification and electrophoresis. It is impossible to dramatically shorten the duration of these steps using existing technologies. SNP analysis may be faster, however: it is possible to analyze 96 SNPs within 30 minutes (Hashiyada, Itakura et al., 2009). Thus, a SNP system could use a specific usage, for example in passports or in very large-scale mercantile transactions.

4.2 Ethical concerns

The polymorphic target region in DNA used to create the DNA-ID does not relate to a person's physical characteristics or disease factors, since the STRs and the SNP loci were selected from the extragenic regions. However, because the DNA-ID system involves handling information that can identify each individual, it should be strictly supervised in order to protect privacy. Once the DNA-ID has been generated, the one-way encryption described above makes it impossible to recover any of the original DNA information (3.1, 3.3). Therefore, raw materials like buccal swab should be especially tightly controlled in order to prevent spoofing.

4.3 Monozygotic twins and DNA chimeras

Monozygotic twins, or more commonly referred to as identical twins, begin life as a single egg, which is fertilized by one sperm but then splits into two eggs early in the gestational period. Therefore, the twins share a precisely duplicated whole genome, and can't be distinguished by DNA polymorphism. However, sometimes one member of a pair of identical twins can develop cancer or schizophrenia while the other does not (Zwijnenburg et al., 2010). A recent "twin study" has revealed that twin pairs have significant differences in their DNA sequence, and furthermore that environmental factors can change gene expression and susceptibility to disease by affecting epigenetics, i.e., changes in the DNA

that do not alter its sequence (Haque et al., 2009). Such data will hopefully aid development of tools that allow discrimination between the identical twins in the near future.

A DNA chimera refers to a recombinant molecule of DNA composed of segments from more than one source. The author has observed chimerism in a case of allogeneic bone marrow transplantation (BMT).The recipient had suffered from acute promyelocytic leukemia and received a BMT from a healthy donor, resulting in complete remission of the leukemia. Samples of peripheral blood leukocytes (PBL), buccal mucosa, hair follicles and fingernails were collected from the transplant recipient. DNA analysis revealed that the STR profile of PBL of the recipient had completely converted to donor type, whereas the hair follicles and fingernails were recipient-derived. DNA patterns of the buccal mucosa appeared chimeric, i.e., they had qualities of both the recipient and donor. Neutrophilic leukocytes were observed in smear specimens from buccal swabs of the recipient, indicating that the buccal cells were not truly chimeric but were instead merely contaminated with leukocytes.

4.4 Cost

DNA analysis requires a high capital cost in order to buy and maintain equipment as well as purchase commercial kits. In addition, it is necessary to equip a laboratory and employ specialists in molecular biology. These high costs may pose a barrier to entry of venture capitals. The more popular such DNA techniques become, however, the lower the unit costs of the apparatus and reagents will become.

5. Conclusion

Development of biometric authentication technologies has progressed rapidly in the last few years. Personal identification devices based on unique patterns of fingerprints, iris, or subcutaneous veins in the finger have all been commercialized. All of these methods of verification are based on matching analog patterns or feature-point comparisons. Because they lack absolute accuracy, they have not yet achieved a universal standard. Among the various types of biometric information source, the DNA-ID is thought to be the most reliable method for personal identification. DNA information is intrinsically digital, and does not change either during a person' life or after his/her death. The discriminatory power of the data can be enhanced by increasing the number of STR or SNP loci. The DNA-ID could be encrypted via the one-way function (SHA-1) to protect privacy and to reduce data length. Using the STR system, it is currently difficult to complete analysis within 3 hours; however, using the SNP system, it is possible to analyse 96 SNPs within 30 minutes. Both systems yielded verifiable results in validation experiments. The author also introduced the idea of DNA-INK as a practical application of DNA-ID.

DNA-ID has some disadvantages, as well, including long analysis time, ethical concerns, high cost, and the impossibility of discrimination of monozygotic twins. However, the author believes that the DNA-ID must be employed as a biometric methodology, using breakthrough methods developed in the near future.

6. Acknowledgments

I am grateful to Dr. Yukio Itakura for his extensive support, and I give special thanks to my colleagues at Div. Forensic Medicine, Tohoku University. I also thank Prof. M. Funayama for reading the manuscript and giving me helpful advice.

7. References

Alberts, B., Jhonson, A., Lewis, J., Raff, M., Roberts, K., Walter P. (2002). *Molecular biology of THE CELL* NY, USA: Garland Science.

Anderson, S., et al. (1981). Sequence and organization of the human mitochondrial genome. Nature, 290(5806): p. 457-65.

Anderson, T.D., et al. (1999). A validation study for the extraction and analysis of DNA from human nail material and its application to forensic casework. J Forensic Sci, 44(5): p. 1053-6.

Andrews, R.M., et al. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet, 23(2): p. 147.

Brookes, A.J. (1999). The essence of SNPs. Gene, 234(2): p. 177-86.

Butler, J.M., et al. (2004). Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. Electrophoresis, 25(10-11): p. 1397-412.

Butler, J.M. (2010). *Fundamemntals of Forensic DNA Typipng*: ELSERVIER.

Clayton, T.M., et al. (1995). Identification of bodies from the scene of a mass disaster using DNA amplification of short tandem repeat (STR) loci. Forensic Sci Int, 76(1): p. 7-15.

Collins, F.S., et al. (2004). Finishing the euchromatic sequence of the human genome. Nature, 431(7011): p. 931-45.

Gill, P. (2001). An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. Int J Legal Med, 114(4-5): p. 204-10.

Hagelberg, E., et al. (1991). Identification of the skeletal remains of a murder victim by DNA analysis. Nature, 352(6334): p. 427-9.

Haque, F.N., et al. (2009). Not really identical: epigenetic differences in monozygotic twins and implications for twin studies in psychiatry. Am J Med Genet C Semin Med Genet, 151C(2): p. 136-41.

Hashiyada, M., et al. (2009). Development of a spreadsheet for SNPs typing using Microsoft EXCEL. Leg Med (Tokyo), 11 Suppl 1: p. S453-4.

Hashiyada, M., Itakura, Y., Nagashima, T., Nata, M., Funayama, M. (2003a). Polymorphism of 17 STRs by multiplex analysis in Japanese population. Forensic Sci Int, 133(3): p. 250-3.

Hashiyada, M., Itakura, Y., Nagasima, T., Sakai, J., Funatyama, M. (2007a). High-throughput SNP analysis for human identification. DNA Polymorphism

Official Journal of Japanese Society for DNA Polymorphism Resarch, 15: p. 3.

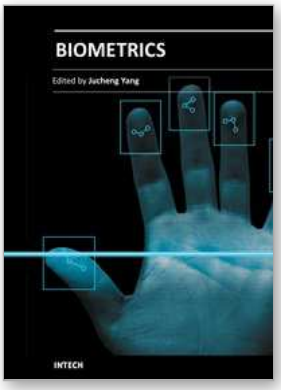
Hashiyada, M., Matsuo, S., Takei, Y., Nagasima, T.,Itakura, Y., Nata, M., Funatyama, M. (2003b). The length polymorphism od SE33(ACTBP2) locus in Japanese population. Practice in Forensuic Medicine, 46: p. 4.

Hashiyada, M., Sakai, J., Nagashima, T., Itakura Y., Kanetake, J., Takahashi, S., Funayama, M. (2007b). The birthday paradox in the biometric personal authentication system using STR polymorphism -Practical matching probabilities evaluate the DNA psesonal ID system-. The Research and practice in forensic medicine, 50: p. 5.

Hedman, J., et al. (2008). A fast analysis system for forensic DNA reference samples. Forensic Sci Int Genet, 2(3): p. 184-9.

Jasinska, A.&W.J. Krzyzosiak (2004). Repetitive sequences that shape the human transcriptome. FEBS Lett, 567(1): p. 136-41.

Jeffreys, A.J., et al. (1985). Hypervariable 'minisatellite' regions in human DNA. Nature, 314(6006): p. 67-73.
Jeffreys, A.J., et al. (1992). Identification of the skeletal remains of Josef Mengele by DNA analysis. Forensic Sci Int, 56(1): p. 65-76.
Jeffreys, A.J., et al. (1995). Mutation processes at human minisatellites. Electrophoresis, 16(9): p. 1577-85.
Jobling, M.A.&C. Tyler-Smith (2003). The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet, 4(8): p. 598-612.
Kayser, M., et al. (2004). A comprehensive survey of human Y-chromosomal microsatellites. Am J Hum Genet, 74(6): p. 1183-97.
Kimpton, C.P., et al. (1993). Automated DNA profiling employing multiplex amplification of short tandem repeat loci. PCR Methods Appl, 3(1): p. 13-22.
Kimpton, C.P., et al. (1996). Validation of highly discriminating multiplex short tandem repeat amplification systems for individual identification. Electrophoresis, 17(8): p. 1283-93.
Lander, E.S., et al. (2001). Initial sequencing and analysis of the human genome. Nature, 409(6822): p. 860-921.
Lee, H.C., et al. (1998). Forensic applications of DNA typing: part 2: collection and preservation of DNA evidence. Am J Forensic Med Pathol, 19(1): p. 10-8.
Lee, H.C.&C. Ladd (2001). Preservation and collection of biological evidence. Croat Med J, 42(3): p. 225-8.
Mullis, K., et al. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol, 51 Pt 1: p. 263-73.
Mullis, K.B.&F.A. Faloona (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol, 155: p. 335-50.
Ruitberg, C.M., et al. (2001). STRBase: a short tandem repeat DNA database for the human identity testing community. Nucleic Acids Res, 29(1): p. 320-2.
Sachidanandam, R., et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature, 409(6822): p. 928-33.
Saiki, R.K., et al. (1986). Analysis of enzymatically amplified beta-globin and HLA-DQ alpha DNA with allele-specific oligonucleotide probes. Nature, 324(6093): p. 163-6.
Shen, W.&T. Tan (1999). Automated biometrics-based personal identification. Proc Natl Acad Sci U S A, 96(20): p. 11065-6.
Slater, G.W., et al. (2000). Theory of DNA electrophoresis: a look at some current challenges. Electrophoresis, 21(18): p. 3873-87.
Stenson, P.D., et al. (2009). The Human Gene Mutation Database: 2008 update. Genome Med, 1(1): p. 13.
Venter, J.C., et al. (2001). The sequence of the human genome. Science, 291(5507): p. 1304-51.
Vijaya Kumar, B.V., et al. (2004). Biometric verification with correlation filters. Appl Opt, 43(2): p. 391-402.
Watson, J., Baker, T., Bell, S., Gann, A., Levine, M., Losick R. (2004). *Molecular Biology of the Gene*, San Francisco, CA, USA: Benjamin Cummings, Cold Spring Harbor Laboratory Press.
Zwijnenburg, P.J., et al. (2010). Identical but not the same: the value of discordant monozygotic twins in genetic research. Am J Med Genet B Neuropsychiatr Genet, 153B(6): p. 1134-49.



Biometrics

Edited by Dr. Jucheng Yang

ISBN 978-953-307-618-8
Hard cover, 266 pages
Publisher InTech
Published online 20, June, 2011
Published in print edition June, 2011

Biometrics uses methods for unique recognition of humans based upon one or more intrinsic physical or behavioral traits. In computer science, particularly, biometrics is used as a form of identity access management and access control. It is also used to identify individuals in groups that are under surveillance. The book consists of 13 chapters, each focusing on a certain aspect of the problem. The book chapters are divided into three sections: physical biometrics, behavioral biometrics and medical biometrics. The key objective of the book is to provide comprehensive reference and text on human authentication and people identity verification from both physiological, behavioural and other points of view. It aims to publish new insights into current innovations in computer systems and technology for biometrics development and its applications. The book was reviewed by the editor Dr. Jucheng Yang, and many of the guest editors, such as Dr. Girija Chetty, Dr. Norman Poh, Dr. Loris Nanni, Dr. Jianjiang Feng, Dr. Dongsun Park, Dr. Sook Yoon and so on, who also made a significant contribution to the book.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Masaki Hashiyada (2011). DNA biometrics, Biometrics, Dr. Jucheng Yang (Ed.), ISBN: 978-953-307-618-8, InTech, Available from: http://www.intechopen.com/books/biometrics/dna-biometrics

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821