II. LITERATURE REVIEWS

STR AND DNA FINGERPRINTING

There are $3.3x10^9$ base pairs on the haploid genome of humans which are composed of 25,000-100,000 genes in different sizes between 5-10 kb pairs. Each one has exons and signal sequences for every process in the cells such as gene expression regulation, chromosome packaging and chromosome segregation. All of them are unique sequences that are present on about 70-80% of human genomes. The rest, about 20-30%, are repetitive DNA, which can be classified into 2 kinds:

1.) Tandem repetitive DNA

The head to tail repeated sequences are composed of core sequences of about 2-250 base pairs for each one. That is about 1/3 of repetitive DNA or 10% of human genomes.

2.) Interspersed repetitive DNA

These DNA are not presented in core sequences that are repeated from head to tail like tandem repetitive DNA. They are found to be dispersed on about 15-20% of human genomes. Other eukaryotic genomes can also be found. This kind of repetitive DNA may vary in size from <500 to >5000 base pairs[35].

If extracted DNA from eukaryotic cells are centrifuged by using density gradient ultracentrifugation, satellite DNA will be found with a white color portion, which is composed of 3 kinds of repetitive DNA[36]:

A.) MACROSATELLITES

The hypervariable DNA that is located near the centromeres and telomere. Because of their long size in megabases, the detection of them needs the pulse field gel electrophoresis technique so that rare applications from DNA polymorphisms of macrosatellites can take place, especially in forensic cases that have a partial decay of biological evidence.

B.) MINISATELLITES

The head to tail repetitive sequence DNA that lower in size than macrosatellites and the high polymorphic DNA fragments have been found from the analysis of these minisatellite DNA by using the restriction fragment length polymorphism (RFLP) technique. These fragments are several kilobases in length. This kind of polymorphic DNA may be located in a single locus (i.e. variable number of tandem repeats, VNTR) or a multilocus on a human genome.

C.) MICROSATELLITES

The head to tail repetitive sequence DNA that lines continuously on the human genome between areas of conserved DNA sequence. These repeated segments (i.e. core units) range in length from two to seven base pairs in double stranded DNA[8-10, 19-20]. Alleles are distinguished by different numbers of copies of the repeat. One hundred -500 base pairs have been found for each region, which make them lower in size than macrosatellites and minisatellites. Among them, the dinucleotide repeats are by far the most abundant. Three-(tri-), four-(tetra-) and five-(penta-) nucleotide repeats are also found at reasonable frequencies throughout the

human genome. Collectively, microsatellite markers that have from two to seven nucleotide repeats are often called short or simple tandem repeats (STR)[14].

At present, several studies use STR loci for several applications[3-4]. That is because of these advantages:

1. They are highly abundant in the human genome, occuring every 6-10 kilobases, so they can be selected as a suitable region for specific purposes.

2. Most of them are polymorphic.

3. Multiplex PCR could be designed in order to provide convenience and time saving.

4. DNA from degraded biological evidences can be successfully amplified by PCR reaction.

STR ALLELE DETERMINATION

It has been discovered that the DNA patterns of individuals in STR loci occures according to Mendelian inheritances. The genotype of each one is composed of 2 alleles in each locus. One allele is received from the father and the other from the mother. The number of core units of tandemly repetitive sequences will be the name of the allele. For example, a person whose genotype in a Tho1 locus is 9-9, that results from a number of 9 core units in each allele at that locus and has at least 1 allele the same as both parents (allele 9), which is homozygote. On the other hand, a person whose genotype in a THO1 locus is 9-6, which means that the

number of core units are six in one allele and nine in another (9 from the father and 6 from the mother), is heterozygote.

CHARACTERISTICS OF STUDIED STR

1. CSF1PO. The human c-fms proto-oncogene for the CSF-1 receptor gene (HUMCSF1PO) is located on chromosome5 at q33.3-34. It is tetranucleotide repeat, which has four base AGAT repeated sequences 5'-3'. At present, 9 alleles have been found according to repeated numbers of repetitive units between the two primers, (5'-AACCTGAGTCTGCCAAG GACTAGC-3' and 5'-TTCCACACACCACCACTGGCCATCTTC-3'). These have been designed for the generation of an amplification product size of between 295-327 base pairs,[14] as follows :

allele	size	allele	size
7	295	12	315
8	299	13	319
9	303	14	323
10	307	15	327
11	311		

2. TPOX. The human thyroid peroxidase gene[37] is located on chromosome2 at 2p23-2pter. It is a tetranucleotide repeat, which has four base AATG repeated sequences 5'-3'. At present, 8 alleles have been found according to repeated numbers of repetitive units between two primers designed for the generation of an amplification product size of between 224-

252 base pairs, as follows :

allele	size	allele	size
6	224	10	240
7	228	11	244
8	232	12	248
9	236	13	252

3. THO1. The human tyrosine hydroxylase gene (HUMTHO1) is located on chromosome11 at 11p15.5. It is a tetranucleotide repeat, which has four base AATG repeated sequences 5'-3'. At present, 8 alleles have been found according to repeated numbers of repetitive units between the two primers, (5'-GTGGGCTGAAAAGCTCCCGATTAT-3' and 5'-GTGA TTCCCATTGGCCTGTTCCTC-3'). These have been designed for the generation of an amplification product size of between 155-179 base pairs,[38] as follows :

allele	size	allele	size
5	155	9	171
6	159	9.3	174
7	163	10	175
8	167	11	179

4. F13AO1. The human coagulation factor XIIIa subunit gene (HUMF13AO1)[39] is located on chromosome6 at 6p24-25. It is a tetranucleo- tide repeat, which has four base AAAG repeated sequences 5'-

3'. At present, 14 alleles have been found according to repeated numbers of repetitive units between the two primers, (5'-GAGGTTGCACT CGAGCCTTTGCAA-3' and 5'-TTCCTGAATCATCCCAGAGCCACA-3'). These have been designed for the generation of an amplification product size of between 281-331 base pairs,[14] as follows :

allele	size	allele	size
3.2	281	10	307
4	283	11	311
5	287	12	315
6	291	13	319
7	295	14	323
8	299	15	327
9	303	16	331

5. FESFPS. The human c-fes/fps proto-oncogene (HUMFESFPS)[40] is located on chromosome15 at 15q25-qter. It is a tetranucleotide repeat, which has four base AAAT repeated sequences 5'-3'. At present, 8 alleles have been found according to repeated numbers of repetitive units between the two primers, (5'-GCTTGTTAATTCATGTAGGGAAGGC-3' and 5'-GTAGTCCCAGCTACTTGGCTACTC-3'). These have been designed for the generation of an amplification product size of between 222-250 base pairs,[14] as follows :

allele	size	allele	size
7	222	11	238

allele	size	allele	size
8	226	12	242
9	230	13	246
10	234	14	250

6. vWA. The human von Willebrand factor gene (HUMVWFA31) is located on chromosome12 at 12p12-pter. It is a tetranucleotide repeat, which has four base AGAT repeated sequences 5'-3'. At present, 12 alleles have been found according to repeated numbers of repetitive units between the two primers (5'-CCCTAGTGGATGATAATAAGAATAATCAGTATG-3' and 5'-GGACAGATGATAAATACATAGGATGGATGG-3'). These have been designed for the generation of an amplification product size of between 123-167 base pairs,[41-42] as follows :

allele	size	allele	size
10	123	16	147
11	127	17	151
12	131	18	155
13	135	19	159
14	139	20	163
15	143	21	167

FORENSIC TERMS AND HOW TO CALCULATE

The important values used in forensic sciences are:

1. **Power of discrimination** (P.D.)

P.D. plays the role of calculation index for displaying the efficiency of individual identification. More P.D. lessens the chance of the same DNA typing being found between two unrelated individuals. For example, the P.D. of the CSF1PO locus is found to be 0.87, which provides an 87% difference among two individual persons or 13% the same DNA typing at that region. There fore, in the example of the calculation of P.D., the TPOX locus is demonstrated by using this Formula: $\mathbf{P.D. = 1- \Sigma(Pi)^2}$

where as Pi is the expected genotype frequencies,[33] which can be calculated by dividing each of the expected genotype number by the total number of all genotypes in each locus; while the expected genotype number can be determined by $1 \times$ number of sample $\times$ (frequency of allele)^2 in the case of the homozygous, but $2 \times$ number of sample $\times$ frequency of $allele_1$ and $\times$ frequency of $allele_2$ in the heterozygous.

There are step by step in calculation as these follows:

1.1) List of expected qenotype frequencies (Pi)

Genotype	Frequency(Pi)
8-8	0.3371
8-9	0.1686
8-10	0.0094
8-11	0.2809
8-12	0.0281
9-9	0.0211
9-10	0.0024
9-11	0.0702

Genotype	Frequency(Pi)
9-12	0.0070
10-10	0.0001
10-11	0.0039
10-12	0.0004
11-11	0.0585
11-12	0.0117
12-12	0.0006

1.2) Determination of $(Pi)^2$ in each expected genotype

Genotype	$(Pi)^2$
8-8	0.1136
8-9	0.0284
8-10	0.0001
8-11	0.0789
8-12	0.0008
9-9	0.0004
9-10	0.0000
9-11	0.0049

Genotype	$(Pi)^2$
9-12	0.0000
10-10	0.0000
10-11	0.0000
10-12	0.0000
11-11	0.0034
11-12	0.0001
12-12	0.0000

1.3) Determination for the summation of $(Pi)^2$

$$\sum(Pi)^2 = 0.1136+0.0284+0.0001+0.0789+0.0008+0.0004+0.0000+0.0049$$
$$+0.0000+0.0000+0.0000+0.0000+0.0034+0.0001+0.0000$$
$$= 0.2309$$

1.4) Filling the value of $\sum(Pi)^2$ in the equation

$$\text{P.D.} = 1 - 0.2309 = \mathbf{0.7691}$$

2. Power of exclusion (P.E.)

P.E. plays the role of calculation index for displaying the efficiency of parantage proving. More P.E. provides a better chance to determine real parentage. There are two formulae for analysing P.E., as follows:

2.1) P.E. (no parent)$^{(34)}$ $= \sum Pi^2(1\text{-}Pi)^2 + \sum 2PiPj(1\text{-}Pi\text{-}Pj)^2$

2.2) P.E (one parent)$^{(34)}$ $= \sum Pi(1\text{-}Pi)^2 + \sum (PiPj)^2(3Pi+3Pj\text{-}4)$

whereas Pi is the most common allele frequencies, and Pj is the next lower common one in each pair of alleles.

P.E. (no parent) is used for parantage proving among alleged persons without any known parent.

P.E. (one parent) is used for father proving with a known mother or vice versa.

The methods of calculating the two equations are represented here step by step by P.E. (no parent) from the CSF1PO locus:

2.1.1) Make a list of allele frequencies (Pi) by arranging from a high to low frequency.

Allele	All-freq.(Pi)
12	0.3387
10	0.2903
11	0.2419
9	0.0726
13	0.0403
15	0.0081
14	0.0081

2.1.2) Fill the Pi value into the first part of the equation for each allele. Then find the summation of them, $\sum Pi^2(1\text{-}Pi)^2$.

Allele	$Pi^2(1-Pi)^2$
12	$(0.3387)^2(1-0.3387)^2 = 0.0502$
10	$(0.2903)^2(1-0.2903)^2 = 0.0424$
11	$(0.2419)^2(1-0.2419)^2 = 0.0336$
9	$(0.0726)^2(1-0726)^2 = 0.0045$
13	$(0.0403)^2(1-0403)^2 = 0.0015$
15	$(0.0081)^2(1-0.0081)^2 = 0.0001$
14	$(0.0081)^2(1-0.0081)^2 = 0.0001$
	$\sum Pi^2(1-Pi)^2 = 0.1324$

2.1.3) Fill the Pi,Pj value into the second part of the equation for each pair of allele by starting from the first pair to the last one. Then find the summation of them, $\sum 2PiPj(1-Pi-Pj)^2$.

Frequency of Allele 12 was Pj for each pair

$2(0.3387x0.2903)(1-0.3387-0.2903)^2 = 0.0271$
$2(0.3387x0.2419)(1-0.3387-0.2419)^2 = 0.0288$
$2(0.3387x0.0726)(1-0.3387-0.0726)^2 = 0.0170$
$2(0.3387x0.0403)(1-0.3387-0.0403)^2 = 0.0105$
$2(0.3387x0.0081)(1-0.3387-0.0081)^2 = 0.0023$
$2(0.3387x0.0081)(1-0.3387-0.0081)^2 = 0.0023$
Total = 0.0881

Frequency of Allele 10 was Pj for each pair

$2(0.2903x0.2419)(1-0.2903-0.2419)^2$ =	0.0307
$2(0.2903x0.0726)(1-0.2903-0.0726)^2$ =	0.0171
$2(0.2903x0.0403)(1-0.2903-0.0403)^2$ =	0.0105
$2(0.2903x0.0081)(1-0.2903-0.0081)^2$ =	0.0023
$2(0.2903x0.0081)(1-0.2903-0.0081)^2$ =	0.0023
Total =	0.0630

Frequency of Allele 11 was Pj for each pair

$2(0.2419x0.0726)(1-0.2419-0.0726)^2$ =	0.0165
$2(0.2419x0.0403)(1-0.2419-0.0403)^2$ =	0.0100
$2(0.2419x0.0081)(1-0.2419-0.0081)^2$ =	0.0022
$2(0.2419x0.0081)(1-0.2419-0.0081)^2$ =	0.0022
Total =	0.0310

Frequency of Allele 9 was Pj for each pair

$2(0.0726x0.0403)(1-0.0726-0.0403)^2$ =	0.0046
$2(0.0726x0.0081)(1-0.0726-0.0081)^2$ =	0.0010
$2(0.0726x0.0081)(1-0.0726-0.0081)^2$ =	0.0010
Total =	0.0066

Frequency of Allele 13 was Pj for each pair

$2(0.0403x0.0081)(1-0.0403-0.0081)^2$ =	0.0006
$2(0.0403x0.0081)(1-0.0403-0.0081)^2$ =	0.0006
Total =	0.0012

Frequency of Allele 15 was Pj for each pair

$$2(0.0081 \times 0.0081)(1-0.0081-0.0081)^2 = 0.0001$$

$$\sum 2PiPj(1-Pi-Pj)^2 = 0.0881 + 0.0630 + 0.0310 + 0.0066 + 0.0012 + 0.0001$$

$$= 0.1900$$

2.1.4) Determination for summation of the first part of equation, $\sum Pi^2(1-Pi)^2$ and the second part, $\sum 2PiPj(1-Pi-Pj)^2$.

P.E. = 0.1324 + 0.1900 = 0.3224

In the same way as the previous equation, P.E. (one parent) can be determined.

HARDY-WEINBERG EQUILIBRIUM

In 1908, G. H. Hardy, a British mathematician, and W. Weinberg, a German physician, discovered independently a rule that relates allelic and genotypic frequencies in a population of diploid, sexually reproducing individuals if that population has random mating, a large size, no mutation or migration, and no selection. The rule has three aspects:[43]

1. The allelic frequencies at an autosomal locus in a population will not change from one generation to the next (allelic frequency equilibrium).
2. The genotypic frequencies of the population are determined in a predictable way by the allelic frequencies (genotypic-frequency equili-

brium).

3. The equilibrium is neutral, in the case it is perturbed, it will be reestablished within one generation of random mating, although at a new allelic frequency (if all the other requirements are maintained).

• ASSUMPTION OF HARDY-WEINBERG EQUILIBRIUM

For considering a population of diploid, sexually reproducing organisms with a single autosomal locus or multiple loci, the following five major assumptions are necessary for the Hardy-Weinberg equilibrium to hold:

1.) RANDOM MATING

This assumption means that the probability of two genotypes mating is the product of the frequencies (or probabilities) of the genotypes in the population. Deviations from random mating come about for two reasons, choice or circumstance. If members of a population choose individuals of a particular phenotype as mates, more or less often than at random, the population is engaged in assortative mating. If individuals with similar phenotypes are mating more often than at random, positive assortative mating is in force. If matings occur between individuals with dissimilar phenotypes more often than at random, negative assortative mating, or disassortative mating, is at work. Deviations from random mating also arise when mating individuals are either more closely related genetically or more distantly related than individuals chosen at random from the population. **Inbreeding** is the mating of individuals who are related

(e.g.,cousins) and **outbreeding** is the mating of genetically unrelated individuals.

One of the first counterintruitive observations of population genetics is that **deviations from random mating alter genotypic frequencies, but not allelic frequencies.** Envision a population in which every individual is the parent of two children. On average, each individual will pass on one copy of each of his or her alleles. Assortative mating and inbreeding will change the zygotic (genotypic) combinations from one generation to the next, but will not change which alleles are passed into the next generation. Thus genotypic, but not allelic, frequencies change under nonrandom mating.

2.) LARGE POPULATION SIZE

Although an extremely large number of gametes is produced each generation, each successive generation is the result of a sampling of a relatively small portion of the gametes of the previous generation. A sample may not be an accurate representation of a population, especially if it is small. Thus, the second assumption of the Hardy-Weinberg equilibrium is that the population is infinitely large. A large population produces a large sample of successful gametes. The larger the sample, the greater the probability that the allelic frequencies of the offspring will accurately represent those in the parent population. **When populations are small or when alleles are rare, changes in allelic frequencies take place due to change alone.** These changes are referred to as **random genetic drift,** or just genetic drift.

3. & 4.) NO MUTATION and NO MIGRATION

Allelic and genotypic frequencies may be changed by the loss or addition of alleles through mutation, or through the migration (immigration or emigration) of individuals from or into a population. **The third and fourth Hardy-Weinberg assumptions are that there is no such allelic loss or addition in the population due to mutation or migration.**

5.) NO NATURAL SELECTION

This assumption is necessary for the Hardy-Weinberg equilibrium: **no individual will have a reproductive advantage over another because of its genotype**. In other words, there is no natural selection occuring (artificial selection, as practiced by animal breeders and plant growers will also perturb the Hardy-Weinberg equilibrium).

In summary, the Hardy-Weinberg equilibrium holds for an infinitely large, randomly mating population in which mutation, migration, or selection does not occur. In view of the assumptions, it seems that such an equilibrium would not be characteristic of natural populations. However, this is not the case. Hardy-Weinberg's equilibrium is approximated in natural populations for two major reasons. **First**, the consequences of violating some of the assumptions, such as no mutation or an infinitely large population size, are small. Mutation rates, for example, are on the order of one change per locus per generation per 10^5 gametes. Thus, there is virtually no measurable effect of mutation in a single generation. In addition, populations do not have to be infinitely large to act as if they were. As can be seen, a relatively small population can still closely approximate the

Hardy-Weinberg equilibrium. In other words, minor deviations from the other assumptions can still result in a good fit to the equilibrium; only major deviations can be detected statistically. **Second**, the Hardy-Weinberg equilibrium is extremely resilient to change because, regardless of the perturbation, the equilibrium is usually reestablished after only one generation of random mating. The new equilibrium will be, however, at new allelic frequencies. The Hardy-Weinberg equilibrium does not "return" to previous allelic values.

• PROOF OF THE HARDY-WEINBERG EQUILIBRIUM

The three properties of the Hardy-Weinberg equilibrium already mentioned, are (1) allelic frequencies do not change from generation to generation, (2) allelic frequencies determine genotypic frequencies, and (3) the equilibrium is achieved in one generation of random mating. All three properties of the Hardy-Weinberg equilibrium can be proven in the example of one-locus, 7 allele (9-15) case of CSF1PO locus in sexually reproducing diploids, by simply observing the **offspring** of a randomly mating, infinitely large population. The initial frequencies of the 28 genotypes are allowed to be any value that sum to **1**; then, the proportions of offspring after one generation of random mating are:

(9-9) :(9-10) : (9-11) : (9-12) : (9-13) : (9-14) : (9-15) : **(10-10)** : (10-11) : (10-12) : (10-13) : (10-14) : (10-15) : **(11-11)** : (11-12) : (11-13) : (11-14) : (11-15) : **(12-12)** : (12-13) : (12-14) : (12-15) : **(13-13)** : (13-14) : (13-15) :

(14-14) : (14-15) : **(15-15)** = **1** : 2 : 2 : 2 : 2 : 2 : 2 : **1** : 2 : 2 : 2 : 2 : 2 : **1** : 2 : 2 : 2 : 2 : **1** : 2 : 2 : 2 : **1** : 2 : 2 : **1** : 2 : **1**.

It is obvious that the proportions of **homozygouses** : **heterozygouses** are **1 : 2** . Then, the expected number of genotypes can be calculated from these equations:

Expected number of any **homozygous** $= 1 \times N \times f_i^2$ (Where N is sample size and f_i is the frequency of any allele from a studied locus).

Expected number of any **heterozygous** $= 2 \times N \times f_{i1} \times f_{i2}$ (Where f_{i1} is the frequency of one allele and f_{i2} is frequency of another one).

Studied alleles and frequencies from CSF1PO locus were found as follows :

Allele	Frequency
9	0.0726
10	0.2903
11	0.2419
12	0.3387
13	0.0403
14	0.0081
15	0.0081

The expected genotypes and frequencies were:

Genotypes	Exp.freq.
9-9	0.0053
9-10	0.0422
9-11	0.0351
9-12	0.0492
9-13	0.0059
9-14	0.0012
9-15	0.0012
10-10	0.0843
10-11	0.1404
10-12	0.1966
10-13	0.0234
10-14	0.0047
10-15	0.0047
11-11	0.0585

Genotype	Exp.freq.
11-12	0.1639
11-13	0.0195
11-14	0.0039
11-15	0.0039
12-12	0.1147
12-13	0.0273
12-14	0.0055
12-15	0.0055
13-13	0.0016
13-14	0.0007
13-15	0.0007
14-14	0.0001
14-15	0.0001
15-15	0.0001

frequencies were changed to Numbers because this chi-square test could be applied only to enumeration data, not frequencies, then the observed genotype number (O) was compared with the expected genotype number (E).

Genotypes	Ob.number	Exp.number
9-9	0	0.3268
9-10	1	2.6134
9-11	1	2.1777
9-12	7	3.0491
9-13	0	0.3628
9-14	0	0.0729
9-15	0	0.0729

Genotypes	Ob.number	Exp.number
10-10	6	5.2250
10-11	10	8.7077
10-12	12	12.1923
10-13	0	1.4507
10-14	1	0.2916
10-15	0	0.2916
11-11	5	3.6280
11-12	6	10.1595
11-13	2	1.2088
11-14	0	0.2430
11-15	1	0.2430
12-12	7	7.1125
12-13	3	1.6926
12-14	0	0.3402
12-15	0	0.3402
13-13	0	0.1007
13-14	0	0.0405
13-15	0	0.0405
14-14	0	0.0041
14-15	0	0.0081
15-15	0	0.004
	N=62	N=62

The number of any observed or expected genotypes, which was less than 5, was pooled together to provide one group[44]. In this presented case, genotypes 9-9, 9-10, 9-11, 9-12, 9-13, 9-14, 9-15, 10-13, 10-14, 10-15, 11-11, 11-13, 11-14, 11-5, 12-13, 12-14, 12-15, 13-13, 13-14, 13-15, 14-14, 14-15, 15-15 were pooled to provide observed and expected groups of 21.0000 and 18.6030, respectively. Then the observed and expected

number of genotypes between each group (6 pairs) was compared as follows :

Genotypes	Ob.number	Exp.number	$(O-E)^2/E$
pooled	21.0000	18.6030	0.3088
10-10	6.0000	5.2250	0.1150
10-11	10.0000	8.7077	0.1918
10-12	12.0000	12.1923	0.0030
11-12	6.0000	10.1595	1.7030
12-12	7.0000	7.1125	0.0018
		total	**2.3234**

After the chi-square value has been calculated as above, reference must be made to a table of chi-squares in order to obtain the corresponding probability value (P). Use of such a table requires that another measurement, the degrees of freedom (df), be established first. For these kinds of data, the number of degrees of freedom is always one less than the number of classes in the ratio, which is 5 in this presented case. To determine whether the deviations observed are considered too large to be attributed to chance (expected), the calculated chi-square value is compared with those given horizontally in the table of chi-square [45] for 5 degrees of freedom (n-1). When a reasonable match is found, the heading at the top of the appropriate column of chi squares is consulted for the corresponding probability value. If the chi-square value equals or exceeds the value in the P = 0.05 column, then the deviations of the actual numbers from the theoretically expected ones are statistically significant, and the conclusion is that the hypothesis is wrong, and therefore rejected. In this case, the chi-square of 2.3234 corresponds to an approximate probability of more than

0.75, but less than 0.90. Thus, a probability of more than 0.75 indicates that the Hardy-Weinberg data fit the proposed ratio very well indeed. This means that the population is in the Hardy-Weinberg equilibrium.