

# Statistical Analysis of a Philippine Population Database at STR Locus FGA for Forensic Applications<sup>1</sup>

Kristina A. Tabbada<sup>2</sup>

## Abstract

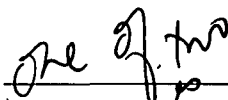
DNA typing is considered the most powerful technique for human identification and maybe used to aid the resolution of criminal and civil cases. Among the most widely used methods of DNA typing is short tandem repeat (STR) loci analysis. To assess the significance of matching profiles or shared alleles, a reference database containing the frequencies of each allele in the population is required. Here, we perform a statistical analysis on the allele frequency distribution for a Filipino population from the National Capital Region (NCR, N=107). DNA was extracted and amplified using standard procedures and analyzed using an automated DNA sequencer (ALFexpress, AP Biotech). Statistical analysis showed that the population conformed to Hardy-Weinberg rules so that the allelic frequencies may be used for forensic calculations. The data used in this study has increased the power of the DNA typing system for use in forensic cases.

## 1. INTRODUCTION

Forensic DNA typing is a scientific procedure for identifying the origins of a human tissue or the genetic relationship of a group of individuals by characterizing (typing) rare features of an individual's genome or hereditary make-up. DNA typing is considered the most powerful technique for human identification and is used in many countries to aid the resolution of criminal (sexual assault, homicide) and civil (questioned paternity, immigration) cases (Kirby, 1992). The power of DNA testing for forensic purposes such as identification is based on the fact that (with the exception of identical twins) each person has a unique DNA sequence. This sequence is composed of a string of nucleotides: adenosine triphosphate (A), cytosine triphosphate (C), guanosine triphosphate (G) and thymine triphosphate (T). However, because sequencing an entire human DNA genome is an immense undertaking, for the purposes of forensic testing it is necessary to test only highly variable regions of the genome, or polymorphic loci (singular, locus), which differ from one person to another. Among the most widely used polymorphisms are short tandem repeat (STR) loci (Hammond et al., 1994). These are small segments of DNA that differ in length due to varying numbers of sequence repeats. Each variation is called an allele, and each allele is designated by the number of repeats it contains. For forensic purposes, STR loci with repeats four nucleotides in length are selected. Hence, if a particular STR has a repeat sequence ACGT, then a five-repeat variation (designated as allele 5) would have the following sequence:

...ACGT ACGT ACGT ACGT ACGT...

whereas a seven repeat variation (designated as allele 7) would have the following sequence:



<sup>1</sup> Winning Entry in 2001 Student Paper Competition in Statistics organized by the Statistical Research and Training Center. This paper was presented during the Student Session of the Eighth National Convention on Statistics (organized by the National Statistical Coordination Board), held on October 1-2, 2001 at the Westin Philippine Plaza.

<sup>2</sup> DNA Analysis Laboratory, Natural Sciences Research Institute, University of the Philippines, Diliman, 1101 Quezon City, Philippines

...ACGT ACGT ACGT ACGT ACGT ACGT ACGT...

Because every person has DNA that is a mixture of DNA inherited from both of his/her parents, everyone has two alleles at each STR locus. The set of alleles that a person has at a locus is called his/her genotype at that locus. (The label "genotype" can be extended to any number of loci and is used synonymously with the lay term "profile".) If person's two alleles at a locus are alike (i.e. both 7 or 7/7), then the person is called homozygous at this locus. If, on the other hand, the alleles are unlike (i.e. 5 and 7 or 5/7), then the person is heterozygous at this locus.

Consider now two situations that are typical of forensic cases. A person is thought to have committed a murder. At the crime scene there is human biological material (blood, hair) that does not belong to the victim (evidence). The question now arises whether the evidence found at the crime scene belongs to the suspect. The genotypes, or DNA profiles (at several STR loci) of the suspect and the evidence are determined and compared. If the profiles do not match (that is, they are different in at least one locus) then it is concluded that the evidence could not have come from the suspect. If, on the other hand, they do match (alike at all loci), then the significance of the match must be evaluated.

The significance of a match will depend upon how common a profile is in the population. If a profile is common in the population, then a match is not of great significance, as it may occur as the result of chance. However, if a profile is rare in the population, then a match becomes highly significant.

In the second situation, a woman is suing a man for support for a child. However, the man (alleged father) claims that the child is not his. The DNA profiles of the alleged father can be taken and compared to resolve this dispute. A child should have inherited one allele at a particular locus from his/her biological father. In cases of disputed paternity this inherited allele is referred to as the shared allele. If the alleged father and child do not share alleles at one or more loci, then the man is not the father of the child. If, on the other hand, the alleged father and child share alleles at all loci tested, then the significance of this allele sharing must be evaluated. Again, the significance of a shared allele will depend upon how common that allele is in the population. If the allele is common in the population, then sharing is not of great significance, as it may occur as the result of chance. However, if an allele is rare in the population, then allele sharing between alleged father and child becomes highly significant.

To assess the significance of matching profiles or shared alleles, a reference database containing the frequencies of each allele in the population is required. It is necessary that the database used should be representative of the population from which the profiles to be tested are derived. This is because allele frequencies vary from one population differing in race or nationality to another (National Research Council, 1996). Thus, the use of an African American, American Caucasian or Taiwanese database to test a Filipino paternity case might result in erroneous conclusions. It is therefore necessary to establish a Filipino population database for forensic STR analysis.

A Filipino population database has previously been established at eight STR loci (Halos et al., 1999). In order to increase the utility of the database in identifying or excluding crime suspects/putative fathers, it is necessary to expand the number of STR loci covered in this database. STR locus FGA is a complex STR found in association with the human alpha fibrinogen locus (Mills et al., 1992). Complex STRs – those which may contain several repeat blocks of variable unit length, along with more or less variable intervening sequences – have the advantage of greater variation due to the range and number of repeat variants. In

the United States and in Europe, STR locus FGA has been widely used in forensic DNA typing due to its high degree of polymorphism as well as its amenability to PCR amplification (Urquhart et al., 1994). The aim of the present study is to apply statistical analysis in assessing the utility of a reference database of STR locus FGA for the Philippine National Capitol Region (NCR) for forensic purposes.

## 2. MATERIALS AND METHODS

### 2.1 Sampling, DNA Extraction and PCR Amplification

Convenience samples of blood samples were collected from 107 unrelated individuals at various locations within the National Capital Region (NCR), including hospitals and schools. DNA was extracted using the method described by Kirby (1992). Extracted DNA was amplified using the following reaction mix: 20 mM KCl, 10 mM Tris HCl, 1.0 mM MgCl<sub>2</sub>, 200 μM each dNTPs, 60 ng/μL bovine serum albumin, 0.02 U/μL Taq polymerase and 0.25 μM of Cy5 labeled and unlabeled primers (primers described by Urquhart et al., 1994). Polymerase chain reaction (PCR) was carried out in a UNO II thermocycler (Biometra). Samples were subjected to an initial heat denaturation step at 96°C for 2 minutes followed by 30 cycles of denaturation at 94°C for one minute, annealing at 60°C for one minute and extension at 72°C for 1 minute 30 seconds. Samples were held at 72°C for an additional 10 minutes to allow complete extension of primers.

### 2.2 Fragment Analysis

PCR products were mixed with 3.0 μL of gel loading buffer and heat denatured at 95°C for three minutes followed by a quick chill at -20°C for seven minutes. Samples were loaded onto a 0.5-mm thick ReProGel High Resolution gel (AP Biotech). Electrophoresis was carried out at 1500 V, 50 mA, 30 W and 40°C on the ALFexpress (AP Biotech) DNA sequencer. Signals were analyzed using ALFwin and Allelelinks (AP Biotech) software programs.

### 2.3 Statistical Analysis

Allele frequencies for FGA were calculated by the gene count method. Possible divergence from Hardy-Weinberg expectations (HWE) and linkage equilibrium (LE) were determined using the exact test from DNAVIEW™ program (Brenner, 1997), with 2000 simulations. The power of discrimination (PD) for FGA alone and combined with the eight other loci (F13A01, FES/FPS, vWA, D8S306, FOLP23, CSF1PO, TH01 and TPOX) in the Philippine (NCR) database were calculated using the Fischer formula (1951). The average power of paternity exclusion (PPE) for FGA alone and FGA combined with the eight other STR loci were also computed.

## 3. RESULTS AND DISCUSSION

A DNA database at STR locus FGA was generated for the National Capital Region (NCR, N=107) population and the allele frequency distribution for this population was determined. Thirteen alleles (see Table 1) were found in the population, ranging from allele 17 to 27 and including rare variants 21.2 and 22.2. The most common allele found was 23 (f=0.21). This allele is also the most common in the Singapore Chinese and Indian populations reported by Fregeau et al. (1998). However, the NCR population differs in this respect from other

populations worldwide, in which other alleles are most common, such as allele 22 in Singapore Malays and USA Caucasians (Fregeau et al., 1998).

*Table 1. Allele frequencies for FGA in the NCR population.*

Allele	Frequency (N = 107)
17	0.0047
18	0.0047
19	0.051
20	0.065
21	0.17
21.2	0.0047
22	0.17
22.2	0.019
23	0.21
24	0.16
25	0.13
26	0.061
27	0.093

### 3.1 Hardy-Weinberg Equilibrium

In order to use a population database composed of FGA allele frequencies for forensic calculations, the population must be in Hardy-Weinberg equilibrium (HWE). Hardy-Weinberg rules assume random mating within the population and the absence of forces such as mutation, selection and migration that would change allele frequencies of FGA from one generation to the next. It is very unlikely that in the NCR population (or any human population, for that matter) these ideal conditions hold true. However, with respect to locus FGA, the population may behave in according to these rules. Obviously, people select their mates independently of the genotype at a STR locus. Thus, it may be said that mating is random *with respect to this locus*. If a population is in HWE, genotypic frequencies can be calculated from allelic frequencies using the equation

$$p^2 + 2pq + q^2 = 1 \quad (1)$$

where  $p$  and  $q$  are allelic frequencies from the population database. Equation (1) states that the proportion of persons with two copies of the same allele is the square of the allele's frequency, and the proportion of persons with two different alleles is twice the product of the two frequencies. This equation can be extended to accommodate any number of alleles at a locus.

Thus, testing for Hardy-Weinberg equilibrium at STR locus FGA is equivalent to asking whether the observed genotypic frequencies of FGA in the sample are close enough to the products of the observed allele frequencies of FGA - that is, the expected genotypic frequencies as calculated in (1) - so that there can be confidence that the same relationship holds for the population frequencies. Of the various methods of testing for conformation to the expectations of Hardy-Weinberg equilibrium, an exact test was used, as this method is most appropriate for small samples (Weir, 1996). The exact test proceeds by looking at all the possible sets of genotypic frequencies for the observed set of allelic frequencies and rejecting the hypothesis of HWE if the observed genotypic frequencies turn out to be unusual under

HWE. The least probable outcomes with a total probability of  $\alpha$  form a rejection region of size  $\alpha$ . If  $p \leq 0.05$ , then the hypothesis of HWE must be rejected at this level, because the outcome (i.e. the genotypic frequencies observed) is expected to occur by chance less than 5% of the time. A Monte-Carlo exact test (Guo and Thompson, 1992) with 2000 simulations showed that the population did not deviate from HWE ( $p=0.7810$ ) at the 95% confidence interval.

### 3.2 Power of Discrimination (PD) of FGA

The finding that FGA is in Hardy-Weinberg equilibrium allows us to use the allelic frequencies to calculate statistics that are significant in forensic science, such as the power of discrimination (PD) and power of paternity exclusion (PPE). Because occurrence of a genotype in one person is a random event that is independent of the occurrence of that genotype in another person, the probability that two randomly-chosen persons will have a particular genotype is the square of its frequency in the population, as expressed in equation (1). The overall probability that two random persons in the population will have the same locus genotype is the sum of the squares of the frequencies of all the genotypes. Expected rather than observed genotypes are used to obtain greater statistical precision. The probability that they will *not* have the same genotype is one minus the sum of the squares of the frequencies of all the genotypes. Thus, the power of discrimination, which is the ability of a locus to discriminate between two random persons in a population, is calculated using Fisher's (1951) formula:

$$1 - [\sum (p_i^2)^2 + \sum (2p_i q_i)^2 + \sum (q_i^2)^2] \quad (2)$$

FGA has a PD of 0.9657. This means that this locus can distinguish between two randomly selected persons in the population 96.57% of the time. This high power of discrimination makes FGA a valuable locus for forensic applications. Recall the case described in the introduction where a suspect's DNA profile is compared to the profile taken from a piece of evidence taken from a crime scene. If the suspect is *not* the source of the evidence, then the suspect and the source of the evidence may be thought of as random persons from the population. The power of discrimination of FGA indicates that this locus will be able to distinguish between the suspect and the evidence in 96.75% of all such cases.

### 3.3 Power of Paternity Exclusion (PPE) of FGA

Power of Paternity Exclusion (PPE) is the ability of a locus to exclude a non-father from paternity (Brenner and Morris, 1989) and can be calculated using the formula:

$$PPE \cong h^2 (1 - 2hH^2) \quad (3)$$

where the rate of homozygosity (H) is

$$H = \sum (p_i^2) \quad (4)$$

and the rate of heterozygosity

$$h = 1 - H. \quad (5)$$

FGA had an average power of paternity exclusion of 0.7185, meaning that it can exclude a non-father from paternity in 71.85% of paternity cases. It is significant that the PPE is a

function of the heterozygosity,  $h(3)$ , which is 0.862. Heterozygosity is a measure of variation of the genotype in the population. Thus, a high level of variation at STR locus FGA in the population – as indicated by heterozygosity - results in a greater power of paternity of exclusion.

### 3.4 Linkage Equilibrium (LE) of FGA

The utility of a locus such as FGA for use in forensic calculations would be increased considerably if used in combination with other loci in the NCR database (Halos et al., 1999), namely: F13A01 (Polymeropoulos et al., 1991c), FES/FPS (Polymeropoulos et al., 1991b), vWA (Kimpton et al., 1992), FOLP23 (Polymeropoulos et al., 1993), D8S306 (Nelson et al., 1993), CSF1PO, TH01 (Polymeropoulos et al., 1991a) and TPOX (Anker et al., 1992). In order to calculate combined PD and PPE, it is necessary to show that these STR loci segregate independently of each other and the occurrence of each genotype at a given locus is independent of the occurrence of a genotype at another locus. This means that the loci are in linkage equilibrium (LE). Linkage equilibrium implies that we are unlikely to observe correlated genotypes at these loci, unless the population is inbred, isolated or highly sub-structured. If linkage disequilibrium were found to exist between a pair of loci, they cannot be used together for forensic calculations, as the occurrence of their genotypes are not independent events. One of the two loci that are not in equilibrium must be discarded. Hence, pair-wise comparisons of FGA and the eight other STR loci were performed to confirm the absence of any association between alleles of different loci (Zaykin et al., 1995). An exact test for linkage disequilibrium indicated that FGA is in equilibrium with the eight STR loci (see Table 2). Thus, all nine loci may be used together for forensic calculations.

*Table 2. Check for linkage disequilibrium between the nine STR loci.*

Loci-Pairs	P-value
F13A01-FGA (N=73)	0.20250
FES/FPS-FGA (N=73)	0.26125
VWA-FGA (N=73)	0.20600
FOLP23-FGA (N=90)	0.33475
D8S306-FGA (N=90)	0.24025
CSF1PO-FGA (N=89)	0.24200
TH01-FGA (N=89)	0.23975
TPOX-FGA (N=89)	0.53625

### 3.5 Combined Power of Discrimination and Power of Paternity Exclusion

Since all nine STR loci in the NCR database - F13A01, FES/FPS, vWA, FOLP23, D8S306, CSF1PO, TH01, TPOX and FGA - are in linkage equilibrium, we can calculate cumulative values for power of discrimination and power of paternity exclusion. Combined power of discrimination (CPD) can be calculated using the following equation:

$$CPD = 1 - (P_1 P_2 \dots P_n) \quad (6)$$

For  $n$  loci, where  $P_i$  is the sum of squares of the genotype frequencies at each locus.

Combined power of paternity exclusion (CPPE) can be calculated using the following equation:

$$CPPE = 1 - [(1 - PPE_1)(1 - PPE_2) \dots (1 - PPE_n)] \quad (7)$$

for  $n$  loci with powers of paternity exclusion  $PPE_n$ .

**Table 3. Summarized statistics of the NCR population across the nine STR loci**

Statistics	F13A01	FES	vWA	D8S306	FOLP23	CSF1PO	TPOX	TH01	FGA
h	0.629	0.671	0.804	0.830	0.732	0.720	0.630	0.730	0.862
PD	0.818	0.830	0.934	0.949	0.890	0.871	0.795	0.883	0.966
PPE	0.327	0.385	0.607	0.655	0.480	0.460	0.330	0.472	0.719
HWE	0.548	0.206	0.406	0.938	0.158	0.568	0.627	0.724	0.781
PPE across all loci					0.9984				
PD across all loci					0.9999999965				

The addition of FGA brought the combined power of paternity exclusion of the nine loci to 0.9984 and the combined power of discrimination to 0.9999999965. The nine-locus system can therefore exclude an unrelated man from paternity in 99.84% of all cases. In addition, the value of PD indicates that the nine STR loci system can discriminate between any two persons in a population of 286,000,000, a population much larger than that of the Philippines. Thus, a person's genotype or profile at the nine STR loci may be said to be unique in the Philippines. As a result, the nine-STR database can distinguish between every Filipino. Thus, statistical analysis indicates that the addition of STR locus FGA to the NCR population database has significantly increased the power of the DNA typing system for use in forensic cases and is now available as a powerful investigative tool in criminal and civil cases.

#### 4. CONCLUSION AND RECOMMENDATIONS

The allele frequency distribution for a Filipino population from the National Capital Region (NCR, N=107) was determined for STR locus FGA. The population conformed to Hardy-Weinberg rules and therefore the allelic frequencies may be used for forensic calculations. FGA had an average power of paternity exclusion of 0.7185 and an index power of discrimination of 0.9657. FGA was found to be in linkage equilibrium with the eight other STR loci currently being used in the laboratory, namely: F13A01, FES/FPS, vWA, FOLP23, D8S306, CSF1PO, TH01 and TPOX; therefore cumulative values for PPE and PD were calculated. The addition of FGA brought the average power of paternity exclusion of the nine loci to 0.9984 and the combined power of discrimination to 0.9999999965. The data obtained in this study has therefore increased the power of the DNA typing system for use in forensic cases.

#### ACKNOWLEDGMENTS

The author is grateful to the Statistical Research and Training Center for organizing the 2001 Student Paper Competition in Statistics. Thanks also to the anonymous referee for comments in improving this paper.

#### References

Anker, R., Steinbreuk, T., and Donnis-Keller, H. (1992). Tetranucleotide repeat polymorphism at the human thyroid peroxidase (hTPO) locus. *Human Molecular Genetics*, 1: 137.

Brenner, C. (1997). DNAVIEW™: A software package for forensic use. California.

Brenner, C. H. and Morris, J. (1989). In Proceedings of the International Symposium on Human Identification, 357-373. Promega Corporation, Madison. Madison, WI.

Fisher, R.A. (1951). Standard calculations for evaluating a blood group system. *Heredity*, 5:95-102.

Fregeau, C. J., Tan-Siew, W.F., Yap, K. H., Carmody, G. R., Chow, S. T., and Fourney, R. M. (1998). Population genetic characteristics of the STR loci D21S11 and FGA in eight diverse human populations. *Human Biology*, 70: 813-844.

Guo, S.W., and Thompson, E.A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*, 48: 361-372.

Hammond, H., Jin, L., Zhong, Y., Caskey, T., and Chakraborty, R. (1994). Evaluation of 13 short tandem repeat loci for use in personal identification applications. *American Journal of Human Genetics*, 55: 175-179.

Halos, S. C., Chu, J. Y., Ferreon, A. C. M., Magno, M.M.F. (1999). Philippine population database at nine microsatellite loci for forensic and paternity applications. *Forensic Science International*, 101: 27-32.

Kimpton, C.P., Walton, A., and Gill, P. (1992). A further tetranucleotide repeat polymorphism in the vWF gene. *Human Molecular Genetics*, 1: 287.

Kirby, L.T. (1992). *DNA Fingerprinting: An Introduction*. W.H. Freeman. New York.

Mills, K. A., Even, D. and Murray, J.C. (1992). Tetranucleotide repeat polymorphism at the human alpha fibrinogen gene (FGA). *Human Molecular Genetics*, 2: 1984.

National Research Council. (1996). *The Evaluation of Forensic DNA Evidence*. National Academy Press, Washington, D.C.

Nelson, L., Riley, R., Lu, J., Robertson, M. and Ward, K. (1993). Tetranucleotide repeat polymorphism at the D8S306. *Human Molecular Genetics*, 2: 1984.

Polymeropolous, M.H., Rath, D.S., Xiao, H. and Merril, C.R. (1991). Tetranucleotide repeat polymorphism at the human tyrosine hydroxylase gene (TH). *Nucleic Acid Research*, 19: 3753.

Polymeropolous, M.H., Rath, D.S., Xiao, H. and Merril, C.R. (1991). Tetranucleotide repeat polymorphism at the human c-fes/fps proto-oncogene (FES). *Nucleic Acid Research*, 19:4018.

Polymeropolous, M.H., Rath, D.S., Xiao, H. and Merril, C.R. (1991). Tetranucleotide repeat polymorphism at the human coagulation factor XIII subunit gene (F13A1). *Nucleic Acid Research*, 19: 4306.

Polymeropolous, M.H., Rath, D.S., Xiao, H. and Merril, C.R. (1993). Tetranucleotide repeat polymorphism at the human dihydrofolate reductase psi-2 pseudogene (DHFRP2). *Nucleic Acid Research*, 21: 4792.



Urquhart, A., Kimpton, C.P., Downes, T.J. and Gill, P. (1994). Variation in short tandem repeat sequences – A survey of twelve microsatellite loci for use as identification markers. *International Journal of Legal Medicine*, 107: 13-20.

Weir, B.S. (1996). *Genetic Data Analysis II: Methods of discrete population genetic data*. Sinauer Associates, Inc. Massachusetts.

Zaykin, D., Zhivotovsky, L. and Weir, B.S. (1995). Exact test for association between alleles at arbitrary numbers of loci. *Genetica*, 96: 169-178.

