ALLELE FREQUENCIES FOR 20 AUTOSOMAL MICROSATELLITE LOCI IN THE KENYAN POPULATION

*Carlo Wamaitha Chege\textsuperscript{1}  \texttt{wamaithachege8@gmail.com}
William Chege Kiarie\textsuperscript{2}  \texttt{chege.william@gmail.com}
Eva Aluvaala Nambati\textsuperscript{2}  \texttt{ealuvaala@kemri.org}
Johnson Kinyua\textsuperscript{1}  \texttt{johnsonkinyua@jkuat.ac.ke}
Eric Lelo\textsuperscript{2}  \texttt{elelo@kemri.org}

* Corresponding author
\textsuperscript{1} Jomo Kenyatta University of Agriculture and Technology
\textsuperscript{2} Kenya Medical Research Institute
Abstract
Samples from 180 unrelated persons of Kenyan descent collected at a DNA testing facility in Nairobi were genotyped using the PowerPlex21® STR kit to generate the first indigenous 20 autosomal STR allele frequency table for use in forensic analysis of human DNA in Kenya. Informed consent for use of the samples for this study was obtained with de-identification procedures employed in accordance with recommendations from the Scientific and Ethics Review Unit at the Kenya Medical Research Institute (KEMRI). The markers amplified for the generation of the allele frequency table were D3S1358, D13S317, PentaE, D16S539, D18S51, D2S1338, CSF1PO, Penta D, THO1, vWA, D21S11, D7S820, TPOX, D8S1179, FGA, D2S1338, D5S818, D6S1043, D12S391, and D19S433. A high degree of gene diversity was observed in this population with average PIC values and heterozygosity of 0.799 and 0.831 respectively across the 20 loci.

Cumulatively, 182 alleles were detected in the Kenyan population analysed across the 20 STR loci. The lowest allele frequency value was 0.003 where one occurrence of the allele was observed while the highest allele frequency was 0.36 for allele 16 marker D3S1358. Polymorphism information content (PIC) results ranged from 0.69 to 0.90 with Penta E returning the highest score. The high PIC score shows that the additional markers offer a more informative value of the genetic markers in this data set. The power of discrimination ranged from 89% to 97% with a combined power of discrimination of 99.99%. The combined match probability, a measure in population genetics that is used to measure the chance of an unrelated person, arbitrarily picked out of the common population and having an identical genotype as that derived from the reference sample or the evidence, was 4.34 x 10^{-26}.

The dataset generated in the present study has been demonstrated to be highly valuable in discriminating between two individual genotypes and greatly amplifies the power of discrimination available to Kenyan forensic DNA testing facilities. The loci included in this dataset comprise the commonly used loci in the US, Europe and Asia and the development of this allele frequency table increases the data sharing possibilities between local and international forensic DNA testing facilities.

Key words: Autosomal STR, variation, Kenyan, statistical parameters, Allele frequency.
1.0 Introduction

Analysis of human DNA for forensic and family relationship testing and identification purposes is currently based on testing of short tandem repeat (STR) markers also called microsatellites. These DNA sequences with a repetitive unit of between two to seven base pairs are used to generate unique DNA profiles composed of genotypes from separate genetic loci. No two individuals possess the same autosomal DNA profile except identical twins. Allele frequency data gathered from a relevant population is used to determine the statistical probability of a random match between two individuals as well as calculate probabilities of relationship between two individuals in question[1].

There are three possible outcomes of a comparison of two STR genotypes in the case of crime scene evidence examination: (a) Exclusion, where there is no match between the two genotypes, (b) an inconclusive result where the relationship between the two profiles is difficult to interpret, or (c) inclusion, where the two STR genotypes in question are a match [2]. Of the three different conclusions, only the latter requires statistical interpretation of the result. The purpose of incorporating statistics is to prove by way of a probability test that the results observed were not a chance event[1]. The match statistics used are usually provided as estimates of the random match probability or the frequency for the particular genotype in a given population [3]. There are however different approaches that may be taken in stating rarity of a match between an unknown and known sample [1].

The only available Kenyan allele frequency table currently has 15 markers comprising of; D8S1179, D21S11, D7S820, CF1PO, D3S1358, TH01, D13S317, D16S539, D2S1338, D19S433, VWA, TPOX, D18S51, D5S818, and FGA loci derived from an older 5 dye STR marker kit [4]. However, recent advances in commercial STR kit development have seen the addition of more STR markers. One of the available commercial STR kits in use, PowerPlex 21 (Promega, Madison, USA), comprises of 20 markers providing additional genotyping data and significantly increasing the power of discrimination available to a forensic analyst when calculating paternity and kinship statistics. This study aimed to generate an additional 5 markers and provide a corresponding allele frequency table that includes D1S1656, D6S1043, D12S391, Penta D, and Penta E to enable utilization of all available genotype data generated from the newer 20 marker STR amplification kits available.

While testing for efficiency of added markers, determination of heterozygosity, defined as the condition at a locus of having two different alleles, is vital [5]. The higher the heterozygosity exhibited the higher the ability to distinguish samples from different individuals. When carrying out paternity testing, the paternity index (PI) is calculated to
measure the likelihood that the alleged father contributed the paternal alleles observed versus an untested, unrelated, random man in the population. An alternative statistic used in paternity testing is the power of exclusion also referred to as probability of exclusion (PE), which is defined as the likelihood that the alleged father contributed the paternal alleles versus an untested, unrelated, random man in the population [6]. The power of exclusion informs the utility of the marker panel used in excluding particular genotypes [7].

The power of discrimination (PD), which has also been referred to as the random match probability is a measure of the probability or chance that two randomly selected individuals have matching genotypes (Planz, 2004). The power of discrimination informs us how powerful the loci are at individualization [7]. Finally, the polymorphism information content (PIC) of a marker is the measure of a marker's usefulness for linkage analysis between parent and offspring [8]. The PIC value is used in genetics as a common measure of polymorphism for an allele locus used in linkage analysis [9]. These statistical parameters were used in the present study to determine the utility of the new 20 marker set.

2.0 Materials and methods

2.1. Study population

The samples used in this study were collected and archived over a period of 5 years from persons visiting the forensic lab for kinship and paternity testing. Informed consent was obtained for use of the samples for extended population studies with de-identification and the protocol was approved following ethical review by the Scientific and Ethics Review Unit-KEMRI. The source population was heterogeneous in ethnic composition and was therefore presumed to be representative of the diversity present across the Kenyan nation. The age and gender of the study participants was not considered during sample selection as these do not influence the allele frequency data. Critically, only samples from confirmed unrelated individuals were included in this study.

2.2 Sample collection

Source DNA was collected from the oral cavity of the donor using sterile foam tipped swabs (COPAN™ Diagnostics Inc). Briefly, the swab was rubbed on the inner lining of the cheek and under the tongue to collect cheek cells with saliva as the transfer medium. The cells on the swab were then transferred from the swab onto indicating nucleic acid cards (COPAN™ Diagnostics Inc) by firmly pressing the swab onto the card for about 30 seconds. Proper sample transfer from the swab onto the card was achieved when the card turned from pink to white. The cards were
then allowed to air dry for about 30 minutes and packaged into individual labelled envelopes and sealed pending DNA profiling. The swabs used to collect the DNA sample were then discarded as bio-hazardous waste.

### 2.3 PCR amplification

The DNA collected on the nucleic acid cards was genotyped using the PowerPlex21® STR kit (Promega, Madison, USA). Briefly, a 1.2mm punch from the nucleic acid card was added into a sterile PCR tube containing 25µl of reconstituted amplification mix prepared by mixing 5 µl of labelled primer mix, 5 µl of master mix and 15 µl of nuclease free water. Positive and negative controls were included in each run to verify the integrity of the PCR reaction as per manufacturer’s instruction. The PCR amplification reaction was conducted on a Veriti™ thermocycler (Applied Biosystems™) with the thermal cycling protocol set as follows: 96°C for 1 minute, 94°C for 10 seconds, 59°C for 1 minute, 72°C for 30 seconds for 25 cycles then 60°C for 20 minutes and a final hold at 4°C.

### 2.4 Fragment analysis

The amplified DNA samples were separated by capillary electrophoresis on a 3500 Genetic Analyzer (Applied Biosystems™). A loading cocktail was prepared on a 96 well plate by combining 0.5µl of WEN ILS 500 size standard (Promega, Madison, USA) and 9.5 µl of highly deionized formamide into which 1 µl of amplified DNA was added. The samples were then centrifuged briefly to remove air bubbles. This was followed by sample denaturation at 95°C for 3 minutes and then immediate chilling on crushed ice. In order to accurately determine the sample genotypes, an allelic ladder was also included in addition to the WEN ILS 500. The allelic ladder was used to accurately size the unknown DNA fragments according to the number of repeats present while the WEN ILS 500 determines the size of the unknown DNA fragments in base pairs. The raw data obtained from the 3500 Genetic Analyzer was processed using GeneMapper® ID-X software version 1.3 (Applied Biosystems™) to generate individual DNA profiles for each sample analysed. The analysed data was exported to MS-Excel which enabled import onto PowerStatsV1.2 software (Promega, Madison, USA) for the generation of allele frequencies.

### 2.5 Data collection procedure

Autosomal STR profiles from the 180 individuals were used to generate an allele frequency table for the 20 markers using PowerStatsV1.2 software. The allele frequencies were determined by counting the number of times each allele was observed in the samples. The STR genotypes collected were converted into observed allele frequencies following the procedure outlined earlier [4]. The number of times each allele is observed in all the samples processed was
counted using PowerStatsV1.2 (Promega, Madison, USA) software and divided by the total number of chromosomes tested, in this case 2 per sample. PowerStats software was also used to calculate the observed heterozygosity (Ho), power of discrimination (PD), probability of exclusion (PE), and polymorphism information content (PIC).

There were cases where some alleles were not sampled sufficiently and the allele was so rare that it was represented only once or a few times in a dataset. In such cases, it is recommended that only alleles observed at least five times be used in forensic calculations [10]. The minimum allele frequency is therefore determined as \(5/(2n)\) where \(n\) is the number of individuals sampled and \(2n\) is the number of chromosomes since autosomes are in pairs due to inheritance of one chromosome from each parent.

Allele independence of the resulting population data was tested through Hardy–Weinberg equilibrium test. Hardy–Weinberg equilibrium (HWE) was used to predict the stability of allele and genotype frequencies from a single generation to the next [10]. For the testing of HWE and to conduct the exact test of population differentiation, GenAlEx version 6.5 was used[11].

3.0 Results

A total of 180 samples were used in this study to generate allele frequencies across 20 microsatellite markers. A total of 182 alleles were detected in the Kenyan population analysed across the 20 STR loci (Table 1). The lowest allele frequency value was 0.003 where one occurrence of the allele was observed while the highest allele frequency was 0.36 for allele 16 marker D3S1358 (Table 1). Since it has been recommended that each allele should be observed a minimum of five times for it be included in accurate and reliable statistical calculations (Statistics, 2013), the lowest observed allele frequency of 0.003 was replaced with the minimum allele frequency of 0.0139 in this population.

| Allele | D3S1358 | D1S1656 | D6S1043 | D13S317 | PENTA E | D16S539 | D18S51 | D2S1338 | CSF1PO | PENTA D |
|--------|---------|---------|---------|---------|---------|---------|--------|---------|--------|---------|
| 2.2    | -       | -       | -       | -       | -       | -       | -      | -       | -      | 0.1444  |
| 3.2    | -       | -       | -       | -       | -       | -       | -      | -       | -      | 0.0222  |

Table 1: Allele frequencies of 20 autosomal STR markers
| ALLELES | TH01 | VWA | D21S11 | D7S820 | D5S818 | TPOX | D8S1179 | D12S391 | D19S433 | FGA |
|---------|------|-----|--------|--------|--------|------|---------|---------|---------|-----|
| 5       | -    | -   | 0.0500 | -      | -      | -    | 0.0167  |         |         |     |
| 6       | -    | -   | -      | -      | -      | -    | 0.0139  |         |         |     |
| 6.4     | -    | -   | -      | -      | -      | -    | 0.0139  |         |         |     |
| 7       | -    | -   | -      | -      | 0.0806 | -    | 0.0500  | 0.0583  |         |     |
| 8       | -    | -   | -      | 0.0472 | 0.1222 | 0.0222| -       | 0.0583  | 0.0639  |     |
| 9       | -    | -   | 0.0139 | 0.0250 | 0.0250 | 0.2028| -       | 0.0778  | 0.1417  |     |
| 9.3     | -    | -   | -      | -      | -      | -    |         |         |         |     |
| 10      | -    | -   | 0.0222 | 0.0278 | 0.0444 | 0.1028| 0.0139  | 0.2222  | 0.1694  |     |
| 10.2    | -    | -   | -      | -      | -      | -    | 0.0139  |         |         |     |
| 11      | -    | 0.0306| 0.0944 | 0.3278 | 0.0694 | 0.3111| -       | 0.2194  | 0.1583  |     |
| 12      | -    | 0.0500| 0.2111 | 0.3556 | 0.0861 | 0.2333| 0.0667  | 0.3222  | 0.0917  |     |
| 12.2    | -    | -   | -      | -      | -      | -    |         |         |         |     |
| 13      | 0.0139| 0.1083| 0.1139 | 0.1694 | 0.1389 | 0.1083| 0.0333  | 0.0500  | 0.0694  |     |
| 13.2    | -    | -   | -      | -      | -      | -    |         |         |         |     |
| 14      | 0.0778| 0.2278| 0.1028 | 0.0472 | 0.0833 | 0.0167| 0.0806  | -       | 0.0306  |     |
| 14.2    | -    | -   | -      | -      | -      | -    |         |         |         |     |
| 14.3    | -    | 0.0167| -      | -      | -      | -    |         |         |         |     |
| 15      | 0.2694| 0.1944| 0.0417 | -      | 0.1417 | 0.0139| 0.1306  | -       | 0.0167  |     |
| 15.2    | -    | -   | -      | -      | -      | -    |         |         |         |     |
| 15.3    | -    | 0.0417| -      | -      | -      | -    |         |         |         |     |
| 16      | 0.3611| 0.1444| 0.0139 | -      | 0.0778 | -    | 0.1639  | 0.0333  | 0.0139  |     |
| 16.2    | 0.0139| -    | -      | -      | -      | -    | 0.0139  | -       | -       |     |
| 16.3    | 0.1028| -    | -      | -      | -      | -    |         |         |         |     |
| 17      | 0.2194| 0.0278| 0.0972 | -      | 0.0417 | -    | 0.1806  | 0.1250  | -       |     |
| 17.2    | 0.0139| -    | -      | -      | -      | -    | 0.0139  | -       | -       |     |
| 17.3    | -    | 0.0306| -      | -      | -      | -    |         |         |         |     |
| 18      | 0.0583| 0.0139| 0.1333 | -      | 0.0167 | -    | 0.1306  | 0.0750  | -       |     |
| 18.2    | 0.0139| -    | -      | -      | -      | -    | 0.0139  | -       | -       |     |
| 18.3    | -    | 0.0139| -      | -      | -      | -    |         |         |         |     |
| 19      | 0.0139| -    | 0.1111 | -      | 0.0139 | -    | 0.0944  | 0.1556  | -       |     |
| 19.1    | -    | -   | -      | -      | -      | -    |         |         |         |     |
| 19.2    | -    | -   | -      | -      | -      | -    |         |         |         |     |
| 20      | -    | -   | 0.0500 | -      | 0.0139 | -    | 0.0444  | 0.1250  | -       |     |
| 21      | -    | -   | 0.0139 | -      | 0.0139 | -    | 0.0222  | 0.1556  | -       |     |
| 21.2    | -    | -   | -      | -      | -      | -    |         |         |         |     |
| 21.3    | -    | -   | 0.0139 | -      | -      | -    |         |         |         |     |
| 22      | -    | -   | -      | 0.0139 | -      | 0.0167 | 0.1028  | -       | -       |     |
| 23      | -    | -   | -      | -      | -      | 0.0139 | 0.0889  | -       | -       |     |
| 24      | -    | -   | -      | -      | -      | -    | 0.0528  | -       | -       |     |
| 24.3    | -    | -   | -      | -      | -      | -    |         |         |         |     |
| 25      | -    | -   | -      | 0.0139 | -      | 0.0500 | -       | -       | -       |     |
| 26      | -    | -   | -      | -      | -      | 0.0167 | -       | -       | -       |     |
| 27      | -    | -   | -      | -      | -      | 0.0194 | -       | -       | -       |     |
| 28      | -    | -   | -      | -      | -      | -    |         |         |         |     |
| 29      | -    | -   | -      | -      | -      | -    |         |         |         |     |
|   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|
| 2.2 | - | - | - | - | - | - | - | - | - |
| 3.2 | - | - | - | - | - | - | - | - | - |
| 5 | - | - | - | - | - | - | - | - | - |
| 6 | 0.1917 | - | - | - | - | 0.0972 | - | - | - |
| 6.4 | - | - | - | - | - | - | - | - | - |
| 7 | 0.3139 | - | 0.0139 | - | 0.0139 | - | - | - | - |
| 8 | 0.2361 | - | 0.2028 | 0.0778 | 0.2944 | - | - | - | - |
| 9 | 0.1833 | - | 0.1083 | 0.0194 | 0.2556 | 0.0139 | - | 0.0139 | - |
| 9.3 | 0.0611 | - | - | - | - | - | - | - | - |
| 10 | 0.0139 | - | 0.3444 | 0.0611 | 0.0694 | 0.0111 | - | 0.0250 | - |
| 10.2 | - | - | - | - | - | - | - | - | - |
| 11 | - | - | - | 0.2222 | 0.2444 | 0.2389 | 0.0389 | - | 0.0639 | - |
| 12 | - | - | - | 0.0944 | 0.3444 | 0.0306 | 0.1417 | - | 0.1500 | - |
| 12.2 | - | - | - | - | - | - | - | - | 0.0167 | - |
| 13 | - | 0.0167 | - | 0.0194 | 0.2389 | 0.0139 | 0.1750 | - | 0.2056 | - |
| 13.2 | - | - | - | - | - | - | - | - | 0.0556 | - |
| 14 | - | 0.1000 | - | - | 0.0139 | - | 0.3222 | 0.0139 | 0.2639 | - |
| 14.2 | - | - | - | - | - | - | - | - | 0.0611 | - |
| 14.3 | - | - | - | - | - | - | - | - | - |
| 15 | - | 0.1917 | - | - | 0.0139 | - | 0.2111 | 0.0889 | 0.0750 | - |
| 15.2 | - | - | - | - | - | - | - | - | 0.0472 | - |
| 15.3 | - | - | - | - | - | - | - | - | - |
| 16 | - | 0.2333 | - | - | - | 0.0694 | 0.0944 | 0.0139 | - |
| 16.2 | - | - | - | - | - | - | - | - | 0.0194 | - |
| 16.3 | - | - | - | - | - | - | - | - | - |
| 17 | - | 0.1861 | - | - | - | - | 0.0222 | 0.1806 | - | 0.0139 |
| 17.2 | - | - | - | - | - | - | - | - | 0.0139 | - |
| 17.3 | - | - | - | - | - | - | - | 0.0139 | - | - |
| 18 | - | 0.1611 | - | - | - | - | 0.0139 | 0.2472 | - | 0.0167 |
| 18.2 | - | - | - | - | - | - | - | - | 0.0139 | - |
| 18.3 | - | - | - | - | - | - | - | - | 0.0139 | - |
| 19 | - | 0.0806 | - | - | - | - | 0.1500 | - | 0.0361 |
| 19.1 | - | - | - | - | - | - | 0.0139 | - | - |
| 19.2 | - | - | - | - | - | - | - | - | 0.0139 | - |
| 20 | - | 0.0306 | - | - | - | - | 0.0889 | - | 0.0361 |
| 21 | - | - | - | - | - | - | 0.0444 | - | 0.0944 |
| 21.2 | - | - | - | - | - | - | - | - | 0.0139 | - |
| 21.3 | - | - | - | - | - | - | - | - | - | - |
The most common alleles among the 20 STR loci typed were allele 7 for TH01, allele 8 for TPOX, allele 10 for Penta D and D7S820, allele 12 for D6S1043, D13S317, D16S539, CSF1PO and D5S818, allele 14 for D1S1656, D8S1179 and D19S433, allele 15 for Penta E, allele 16 for D3S1358 and vWA, allele 17 for D18S51, allele 18 for D12S391, allele 19 and 21 for D2S1338, allele 22 for FGA, and allele 29 for D21S11. The Penta E locus exhibited the highest allelic diversity while D3S1358 had the lowest diversity (Figure 1). Overall, allele 16 at locus D3S1358 was the most common in the population sampled (Table 1). The total number of alleles found in the population tested was 182.
Measure of Hardy Weinberg Equilibrium for the 20 loci showed that only 2 loci (locus D7S820 and FGA), out of the 20 loci had significant departure from HWE (TABLE 2). None of the 20 STR loci were in Linkage disequilibrium.

| LOCUS   | CHISQ  | PROB   | SIGNIFICANCE |
|---------|--------|--------|---------------|
| D3S1358 | 30.409 | 0.084  | Ns            |
| D1S1656 | 23.068 | 0.730  | Ns            |
| D6S1043 | 80.860 | 0.390  | Ns            |
| D13S317 | 17.059 | 0.708  | Ns            |
| PENTA E | 127.735| 0.932  | Ns            |
| D16S539 | 21.777 | 0.792  | Ns            |
| D18S51  | 74.489 | 0.592  | Ns            |
| D2S1338 | 49.645 | 0.933  | Ns            |
| CSF1PO  | 28.321 | 0.131  | Ns            |
| PENTA D | 79.395 | 0.802  | Ns            |
| TH01    | 7.907  | 0.638  | Ns            |
| VWA     | 22.618 | 0.752  | Ns            |
| D21S11  | 79.912 | 0.117  | Ns            |
| D7S820  | 35.009 | 0.028  | *             |
| D5S818  | 33.441 | 0.220  | Ns            |
| TPOX    | 23.875 | 0.688  | Ns            |
Polymorphism information content (PIC) results ranged from 0.69 to 0.90 with Penta E returning the highest score (Table 3). The high PIC score ranging from 0.69 to 0.90 represents the high informative value of the genetic markers in this data set for linkage studies. The power of discrimination ranged from 89% to 97% with a combined power of discrimination of 99.99%. The combined match probability, a measure in population genetics that is used to measure the chance of an unrelated person, arbitrarily picked out of the common population and having an identical genotype as that derived from the reference sample or the evidence, was $4.34 \times 10^{-26}$ (Table 3) which combined with the high power of discrimination demonstrate the utility of the dataset in accurately distinguishing among individuals in the population.

### Table 3: Forensic efficiency parameters of the 20 autosomal STR loci

| LOCUS     | PIC VALUES | POWER OF DISCRIMINATION | POWER OF EXCLUSION | HETEROZYGOSITY | MATCHING PROBABILITY |
|-----------|------------|--------------------------|--------------------|----------------|----------------------|
| D3S1358   | 0.70       | 0.89                     | 0.42               | 0.69           | 0.11                 |
| D1S1656   | 0.84       | 0.96                     | 0.78               | 0.89           | 0.04                 |
| D6S1043   | 0.87       | 0.97                     | 0.73               | 0.87           | 0.03                 |
| D13S317   | 0.69       | 0.89                     | 0.44               | 0.71           | 0.11                 |
| PENTA E   | 0.90       | 0.98                     | 0.82               | 0.91           | 0.02                 |
| D16S539   | 0.75       | 0.92                     | 0.53               | 0.76           | 0.08                 |
| D18S51    | 0.87       | 0.97                     | 0.83               | 0.92           | 0.03                 |
| D2S1338   | 0.88       | 0.97                     | 0.81               | 0.91           | 0.03                 |
| CSF1PO    | 0.75       | 0.92                     | 0.61               | 0.81           | 0.08                 |
| PENTA D   | 0.87       | 0.97                     | 0.83               | 0.92           | 0.03                 |
| TH01      | 0.73       | 0.91                     | 0.58               | 0.79           | 0.09                 |
| VWA       | 0.81       | 0.94                     | 0.74               | 0.87           | 0.06                 |
| D21S11    | 0.85       | 0.96                     | 0.69               | 0.85           | 0.04                 |
| D7S820    | 0.74       | 0.90                     | 0.61               | 0.81           | 0.10                 |
| D5S818    | 0.71       | 0.89                     | 0.51               | 0.75           | 0.11                 |
| TPOX      | 0.74       | 0.91                     | 0.52               | 0.76           | 0.09                 |
| D8S1179   | 0.77       | 0.93                     | 0.63               | 0.82           | 0.07                 |
| D12S391   | 0.84       | 0.96                     | 0.77               | 0.89           | 0.04                 |
| D19S433   | 0.83       | 0.96                     | 0.65               | 0.83           | 0.04                 |
| FGA       | 0.86       | 0.97                     | 0.76               | 0.88           | 0.03                 |
The combined paternity index was $2.59 \times 10^{10}$ which indicates how many times more likely it is that the alleged father in a paternity test is the biological father as opposed to a randomly selected unrelated man from a similar ethnic background. The loci with the highest observed heterozygosity were D18S51 and Penta D with a heterozygosity value of 92%, while locus D3S1358 had the lowest heterozygosity of 69%. The mean heterozygosity of the 20 STR loci was 83%. The high degree of observed heterozygosity corresponds to high allelic diversity and reduces the likelihood of a random match between sample genotypes.

4.0 Discussion and Conclusion
This paper documents the first development of allele frequency data from the Kenyan population targeting 20 autosomal STR loci found in newly developed commercial forensic autosomal STR kits. Prior to this work, the only available allele frequency table for use in forensic analysis of human DNA consisted of 15 markers[4]. However, there have been challenges in the use of this frequency data set since most available commercial makers have now been improved and scaled up by addition of more makers as is the case with the commercial STR kit used in this study. This therefore meant that when calculating match probability statistics, only the 15 makers set were considered since the currently available allele frequency table could only accommodate these. The dataset generated in the present study has been demonstrated to be highly valuable in discriminating between two individual genotypes and greatly amplifies the power of discrimination available to Kenyan forensic DNA testing facilities. The loci included in this dataset comprise the commonly used loci in the US, Europe and Asia and the development of this allele frequency table increases the data sharing possibilities between local and international forensic DNA testing facilities. We believe that the use of the allele frequency table generated in the current study greatly contributes to the refinement of the forensic analysis in relation to paternity testing and kinship study.
5.0 Declaration and acknowledgement

**Abbreviations**

DNA  
Deoxyribonucleic acid

STR  
Short tandem repeats

HWE  
Hardy Weinberg equilibrium

KEMRI  
Kenya Medical Research Institute

**Ethics approval and consent to participate**

Informed consent was obtained for use of the samples for extended population studies with strict de-identification procedures and the protocol was approved following ethical review by the Scientific and Ethics Review Unit-KEMRI under protocol reference KEMRI/SERU/CBRD/198/3897.

**Consent for publication**

Not applicable.

**Availability of data and materials**

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

**Competing Interests**

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

**Funding**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Author's contributions**

CW was extensively involved in proposal development, data collection and analysis as well as preparation of the manuscript. EA significantly assisted in data analysis and manuscript development. WC assisted in data analysis and manuscript development. JK broadly assisted in proposal preparation and data analysis and interpretation. EL provided oversight in proposal development, data analysis, interpretation and preparation of the manuscript.

**Acknowledgements**

Not applicable.
6.0 References

[1] J. M. Butler, “Fundamentals of Forensic DNA Typing,” Fundam. Forensic DNA Typing, 2010, doi: 10.1016/C2009-0-01945-X.

[2] F. R. Bieber, J. S. Buckleton, B. Budowle, J. M. Butler, and M. D. Coble, “Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion,” BMC Genet., vol. 17, no. 1, p. 125, 2016, doi: 10.1186/s12863-016-0429-7.

[3] J. M. Butler, “Genetics and Genomics of Core Short Tandem Repeat Loci Used in Human Identity Testing,” 2006, doi: 10.1111/j.1556-4029.2006.00046.x.

[4] K. J. Kimani, "Kenya population forensic data on a sixteen loci microsatellite DNA system," Aug. 2011, Accessed: Sep. 18, 2018. [Online]. Available: http://ir-library.ku.ac.ke/handle/123456789/919.

[5] F. W. Allendorf, “Convergent Evolution - Evolutionary Biology - Oxford Bibliographies,” 2017, Accessed: Mar. 12, 2020. [Online]. Available: https://www.oxfordbibliographies.com/view/document/obo-9780199941728/obo-9780199941728-0039.xml.

[6] W. K. Fung, Y. ka Chung, and D. man Wong, "Power of exclusion revisited: Probability of excluding relatives of the true father from paternity," Int. J. Legal Med., vol. 116, no. 2, pp. 64–67, 2002, doi: 10.1007/s004140100210.

[7] J. V Planz, “Forensic Statistics From the ground up...,” 2004, Accessed: Mar. 12, 2020. [Online]. Available: https://www.promega.com/-/media/files/resources/conference-proceedings/ishi-15/parentage-and-mixture-statistics-workshop/generalpopulationstats.pdf?la=en.

[8] R. C. Elston, “Polymorphism Information Content,” in Encyclopedia of Biostatistics, John Wiley & Sons, Ltd, 2005.
[9] S. Shete, H. Tiwari, and R. C. Elston, “On estimating the heterozygosity and polymorphism information content value,” *Theor. Popul. Biol.*, vol. 57, no. 3, pp. 265–271, May 2000, doi: 10.1006/tpbi.2000.1452.

[10] J. M. Butler, *Fundamentals of Forensic DNA Typing*. Elsevier Science, 2009.

[11] R. Peakall and P. Smouse, “Methods and Statistics in GenAlEx 6 By Rod Peakall and Peter Smouse,” *Statistics (Ber)*, vol. 502, no. September 2015, pp. 1–23, 2005, doi: 10.1111/j.1471-8286.2005.01155.x.