UNIVERSITI SAINS MALAYSIA

Allele Frequencies of F13A01, FESFPS and vWA STRs in random Dusun population of Malaysia

Dissertation submitted in partial fulfillment for the Degree of Bachelor of Science (Health) in Forensic Science

Mohd Yusmaidie Bin Aziz

School of Health Science
Universiti Sains Malaysia
Health Campus
16150 Kubang Kerian, Kelantan
Malaysia

2007

CERTIFICATE

This is to certify the dissertation entitled

"Allele frequencies of F13A01, FESFPS and vWA STRs in random Dusun population of Malaysia"

Is the bonafide record of research work done by

Mr. Mohd Yusmaidie bin Aziz

During the period 17/12/2006 to 26/03/2007

Under our supervision

ACKNOWLEDGEMENTS

It is my great honour and proud privilege to acknowledge with gratitude, the constant supervision, keen interest and an affectionate guidance rendered to me at every step by my revered teachers, Dean of School of Health Science, Prof. Zainul Fazirruddin Zainuddin, Prof. Dr. Norazmi, the pioneer of forensic programme in USM and also Co supervisor for my project, Dr. Mohamad Hadzri Bin Yacoob, Chairman, Forensic Science Programme, Asso. Prof. Dr. Kuppuswamy, Coordinator, Forensic Science degree student's projects, Mr. Paneerchelvam, supervisor of my project whom made it possible for me to pursue this study and complete this thesis. I am extremely indebted to them for their devotion, valuable time, timely advice and suggestions. It is with great pleasure that I take this opportunity to express my gratitude to Nur Haslindawati, School of Health Science, USM for her guidance, suggestions, the unfailing interest and enthusiasm, for this research. I feel obliged to Seri Mirianti for her encouragement and generous help during this study. I also extend my sincere Puan Hafizah. Thanks are also due to Puan Rosniah who give me the flexibility to use the laboratory. I acknowledge whole heartedly the cooperation of my dear colleagues and other staff members in the department during this work. I am deeply indebted to my respected parents for their constant enthusiasm, support and encouragement in various ways, without which it could not have been possible on my part to complete this study. Last but not the least, I pay my regards to those honorable sample providers, people of Dusun population, without whom I could not have completed my study.

(Mohd Yusmaidie Aziz)

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Short Tandem Repeats (STRs) loci are characterized by high level of length polymorphism within and among populations. This polymorphism makes it becomes a blue print for forensic purposes. In this study, STR is applied for determining the allele's frequencies in Dusun population in Sabah. This is done by using three different STR loci F13A01, FESFPS and vWA. About 100 buccal swab samples from Dusun population were collected. Each sample was separated into different envelopes and sealed. Then DNA was extracted from the samples using simple procedure describe by Promega followed by STR typing on the three locus using multiplex kits. Allele frequencies were calculated and statistical analysis was then carried out. Comparative studies were carried out with referring to allele distribution from other populations. This study produces a database for random Dusun population and hopefully will help in human identification in Malaysia. We also fid out that the genotype frequency distributions do not deviate from Hardy-Weinberg expectations.

INTRODUCTION

Individualization and identification of the evidence materials in the process of law has increased the need for the forensic fields. Forensic scientists have been using polymorphic genetic markers such as blood groups and serum proteins for identification of human remains since they are inherited by simple Mendelian law. The concept of DNA fingerprints (Butler, 2001) has revolutionized human identity testing (Jeffreys AJ et al., 1985). DNA fingerprint is used as a routine procedure in identification of biological materials.

In general all living organisms are composed of cells which are the basic integrated units of biological activity. In human and other higher organisms, cell contains the hereditary material which is known as deoxyribonucleic acid (DNA). DNA is contained in the cell nucleus in the form of chromosomes.

DNA molecules are composed of two chains of repeating nucleotides. Each nucleotide consists of three components. These components are; Phosphate Group, 2-deoxyribose sugar and a nitrogen containing base (cytosine (C), adenine (A), guanine (G) and thymine (T). All these structure construct the DNA structure to resemble a spiral ladder. The pairing of these bases is specific: A always pairs with T and C pairs with G. DNA is a complex, double stranded molecule that is twisted into a helical form and known as the double helix structure.

These four nucleotide bases represent the 'genetic alphabet' and the sequences of base-pairs along the length of the DNA molecule comprise a biochemical vocabulary which encodes the genetic information essential to life processes. The genetic code demonstrates that each of 20 amino acids is coded by a sequence of three bases of the nucleotide. (Marshall Nirenberg, 1963 and Gobind Khorana, 1996) The absolute specificity of base-pairing also provides a mechanism through which "parent" DNA molecules can be copied to form identical "daughter" DNA molecules in the process of replication. The mechanism, known as replication (as opposed to duplication), is possible because the two sides of the parent DNA molecule are complementary rather than identical. In the replication process, the parent DNA molecule "splits" along its length, each side serving as a template for one of the new and identical daughter molecules. (Marshall Nirenberg, 1968, Gobind Khorana, 1996)

Although genes are composed of DNA and contained in the chromosomes in the nucleus of the cell, only a small fraction of that DNA is actually used to form genes. Most of the DNA in our chromosomes has no known function; the portion of such DNA may be more than 95% of genomic DNA. The actual percentage of non-coding DNA is not known with accuracy. The estimate provided in this source is 98.5%. Another source suggests that only 2.5-3.3% of DNA can be expected to be found in the genes. (Paul Berg and Maxine Singer, 1992)

Forensic scientist show more interest in these junk DNA. It is also known that the non-coding DNA contains random sequences and in which are inherited by the individual

from his or her parents just as functional genes are inherited. These tandem repeats are unique for each individual except in the case of identical twins. (Jeffreys et al., 1985) These non-coding base-pair repeated sequences are also known as variable number tandem repeats, (VNTR). (Nakamura Y et al., 1987)

The DNA fragments of interest in forensic-DNA analysis and typing are called restriction fragment length polymorphisms, (RFLP). Where a specific gene, or non-coding DNA sequence, is identical on each of the pair of chromosomes, the condition is said to be homozygous, in contrast to a heterozygous condition, where the two genes or sequences differ in some way.

Short tandem repeats (STR) are variable number tandem repeats (VNTRs), except that the repeated units are much shorter. STRs chosen for forensic use generally have a tandem repeat unit of only three to four base-pairs, which may be repeated in the DNA molecule from a few to dozens of times. The advantage is that only small amounts of even badly degraded DNA may be sufficient for test (Nakamura et al., 1987). The problem is that a very small sample of very short DNA segments - the short tandem repeats - needs to be increased in size for analytical convenience and efficiency. This is done through the use of the polymerase chain reaction.

STRs have an individualistic value. For a given STR, one inherits different number of repeated sequences from each of parents. For example one may have inherited 12 repeats of the CATG sequence on a chromosome from his mother and 5 repeats of this

sequence within the STR on the matching chromosome from his father. There are hundreds of STRs that have been mapped throughout human genome and some of them are used for application of human identity testing. Forensic purpose STRs with tetra nucleotide repeats are due to their fidelity in PCR amplification. The other features of STR included high heterozygosity, regular repeat unit, distinguishable alleles and robust amplification. (www.biotechnologyonline.gov)

The STRBase is information describing each commonly used STR DNA marker. There are over 750 population studies which have been published and now brought together in one place for the first time through STRBase. Most of the studies contain a sample size of greater than 100 samples, which is usually sufficient to make reliable projections about a genotype's frequency in a larger population (Chakraborty, R., 1992). The population data has been sorted by STR locus and is available for CD4, CSF1PO, D21S11, DYS19, F13A1, F13B, FES/FPS, FGA, HPRTB, HUMTH01, LPL, SE33, TPOX, and VWA.

The polymerase chain reaction or PCR, is a technique for increasing the amount of a specific sequence of DNA in a sample and the technique has proven to be invaluable in forensic DNA work. Literally, PCR can make millions or billions of copies of a selected or target DNA sequence in a test tube and accomplish this feature within a matter of a few hours. (Mullis et al., 1980) This is technically called "DNA amplification". The PCR process is similar to the mechanism by which DNA duplicates itself in the cell. There are three steps in the process, as carried out in the laboratory: first,

the double-stranded DNA segment or sequence is separated into two strands by heating; next, the single-stranded segments are prepared through being hybridized with "primers" - short DNA segments - that define the target sequence to be amplified; and third the enzyme DNA polymerase is added to the mixture, along with a quantity of the four nucleotide bases and the replication process begins. The three-step cycle is repeated for 25-30 times.

It has been already established that there are racial differences in allele frequencies (Gerbel and Saferstein, 1997; Butler, 2001). Therefore, there are many construction of database for different racial groups have been done. The utmost particular concern in these databases is the use of general population databases to determine DNA profile probabilities. Population geneticists use statistical probabilities to claim for uniqueness of a forensic DNA profile. Finding that two samples have the same DNA patterns does not necessarily mean they come from the same individual, just as finding two specimens with the same blood type does not mean they come from the same person. RFLP patterns represent only a snapshot of the unique DNA sequence of each individual. The validity of forensic DNA tests does not hinge on population genetics. Interpreting test results, however, depends on population frequencies of the various DNA markers. In other words, population genetics provides meaning numerical weight to DNA patterns obtained by molecular genetics techniques. (United States Congress, 1990)

There are several types of STRs that have been classified by Urquhart et al., (1994) in which three of them are being used in this project. They are simple consisting of one repeated sequence (FESFPS), simple with nonconcensus alleles (HUMTH01 and F13A01) and compound with non consensus alleles (vWA), (Edwards A et al., 1991). In this study, population database for three STRs (F13A01, FESFPS and vWA) for Dusun population living in Sabah, Malaysia was compiled. This multiplex was chosen because the STRs are validated for forensic purposes due to higher polymorphism, have high discriminating power, and have high sensitivity across all of the loci and low intensity stutter peaks and other artifact peaks. The alleles distribution has been checked using Chi-square (X_2)test and the alleles frequency is compared with other populations.

LITERATURE REVIEW

Deoxyribonucleic acid (DNA) typing or DNA profiling has revolutionized the concept of human identity testing (Butler, 2001) and has entered as a routine analytical work in forensic science institutions, making obsolete the use of blood groups and other polymorphic biochemical genetic markers. Sir Alec Jeffrey was the first to describe about DNA fingerprinting in 1985. He discovered that the number of repeated sequences found in certain regions of DNA could differ among individual. These repeated sequences are known as variable number of tandem repeats (VNTR) and in order to examine the length variation of the sequences, RFLP (Restriction Fragment of Length Polymorphism) are used.

Most of our DNA is identical to DNA of others. However, there are inherited regions of our DNA that can vary from person to person. Variations in DNA sequence between individuals are termed 'polymorphisms'. The sequences with the highest degree of polymorphism are very useful for DNA analysis in forensics cases and paternity testing and even for genetic study. The inheritance of a class of DNA polymorphisms is known as 'short tandem repeats', or simply STRs. (Edwards et al., 1991)

Short tandem repeats (STRs) have become popular in forensic laboratories. Since low amount of DNA even in a degraded form can be DNA typed. These STRs contain repeat units that are between 2-6 bp in length and they can be amplified with the polymerase chain reaction (PCR). (Sparkes et al., 1996)

In the early 1990s, STR markers were first described as effective tools for human identity testing. The human identity testing community has settled on a set of core short tandem repeat (STR) loci that are widely used for DNA typing applications. (Edwards et al., 1991)

The four loci of TH01, VWA, FES/FPS, and F13A1 was the first multiplex applied to forensic casework. (Kimpton et al., 1994). A second generation multiplex (SGM) has followed with the loci TH01, VWA, FGA, D8S1179, D18S51, and D21S11. (Kimpton et al., 1996). The amelogenin sex-typing test was then launched in 1995 utilizing the Second Generation Multiplex (SGM) loci. (Werrett DJ., 1997)

The success of STR typing technology had lead the FBI Laboratory of U.S. to establish core STR loci that would form the backbone of CODIS, the U.S. national database system in which a community-wide STR Project was launched in 1996 (Budowle et al., 1998)

This project had involved 22 DNA typing laboratories that collectively evaluated 17 candidate loci, which were available as commercial or preliminary kits from either Promega Corporation (Madison, WI) or Applied Biosystems (Foster City, CA). Performance studies and protocol evaluations were performed, population databases were established, and forensic validation was conducted on the various STR systems investigated. Early work with STRs involved detection on silver-stained polyacrylamide gels (Lins AM, Sprecher CJ, Puers C, Schumm JW., 1996), fluorescence detection

methods involving gel electrophoresis (Edwards A. et al., 1991) and presently capillary electrophoresis with such instruments as the ABI 310 and ABI 3100 Genetic Analyzers are used for STR DNA profiling. (Butler JM, Buel E, Crivellente F, McCord BR., 2004)

The PowerPlex 16 kit, which was released by the Promega Corporation in 2000, amplifies the 13 core loci, amelogenin, and two pentanucleotide loci referred to as Penta D and Penta E (Krenke et al, 2002). Applied Biosystems released their 16plex Identifiler kit in July 2001, which amplifies the 13 core loci, amelogenin, and two tetranucleotide loci D2S1338 and D19S433 (Collins et al., 2004)

In the process of mutational transformations that act upon DNA is tandem duplication, a stretch of DNA is duplicated to produce two or more adjacent copies, resulting in a tandem repeat. Over time, the copies undergo additional mutations so that typically, multiple approximate tandem copies are present. An interesting feature of tandem repeats is that the duplicated copies are preserved together, making it possible to do phylogenetic analysis on a single sequence. This involves using the pattern of mutations among the copies to determine a minimal or a most likely history for the repeat. A history tries to describe the interwoven pattern of substitutions and duplication events in such a way as to minimize the number of identical mutations that arise independently. Because the copies are adjacent and ordered, the history problem can not be solved by standard phylogeny algorithms. (Gary Benson and Lan Dongy, Department of Biomathematical Sciences The Mount Sinai School of Medicine)

In recent years, short tandem repeat (STR) systems have gained importance in forensic analysis of biological specimens as well as in paternity testing, as an alternative to the use of restriction fragment length polymorphism (RFLP) analysis (Edwards et al. 1991; Hammond et al. 1994; Nakamura et al. 1987). PCR-based STRs have several advantages over conventional Southern blotting techniques of the larger variable number tandem repeats (VNTRs). Discrete alleles from STR systems may be obtained due to their smaller size, which puts them in the size range where DNA fragments differing by a single base pair in size may be differentiated. Determination of discrete alleles allows results to be compared easily between laboratories without binning. (http://www.promega.com.). In addition, smaller quantities of DNA, including degraded DNA, may be typed using STRs. Thus, the quantity and integrity of the DNA sample is less of an issue with PCR-based typing methods. (http://www.promega.com.)

There are several types of STRs which have been classified by Urquhart et al (1994).

1. Simple, consisting of 1 repeating sequence: FES/FPS
2. Simple with nonconsensus alleles: HUMTH01, F13A1
3. Compound, consisting of 2 or more different repeat sequences: GABRB15
4. Compound with nonconsensus alleles: vWA
5. Complex repeats: D21S11
6. Hypervariable repeats: SE33

F13A01, FESFPS and vWA were the chosen loci in this study.

F13A1

Other Names	Chromosomal Location	GenBank Accession
FXIIIA01	6p24-p25	M21986; 7 repeats

Repeat: [AAAG] = GenBank top strand (commonly used) (Polymeropoulos, M.H., et al., 1992).

FES/FPS

Other Names	Chromosomal Location	GenBank Accession
FES	15q25-qter; human c-fes/fps proto-oncogene	X06292; has 11 repeats

Repeat: [AAAT] = bottom strand (commonly used); [ATTT] = GenBank top strand (Polymeropoulos, M.H., et al., 1992)

VWA

Other Names	Chromosomal Location	GenBank Accession
vWF, VWA31A	12p12-pter	M25858; has 18 (20) repeat units

Repeat: [AGAT] = bottom strand (commonly used); [TCTA] with [TCTG] and [TCCA] inserts = GenBank top strand (Kimpton et al., 1992)

Due to the fact that DNA typing is only an examination of a DNA sample's sequence and/or length at discrete locations , a match in DNA typing is always a statistical exercise. Currently, time and expense limit an examination of an individual's entire genome, which would show unique identity. In order to determine the probability that a particular genotype might occur at random in a population, population data must be gathered to make an estimate of the frequency of each possible allele and genotype. Usually a sample size of greater than 100 samples is sufficient to make reliable projections about a genotype's frequency in a larger population (Chakraborty, R., 1992).

STR typing can also be done with degraded biospecimen and the amplification product are less than 400 base pairs (Nakamura et al., 1987). Published population database for many different populations are available (Polymeropoulous et al., 1992 Nelleman, 1994, Nagai et al., 1996, Garofano et al., 1998; Sinha et al., 1999; add 2000, 2001, 2002, 2003). Published database on ethnic population groups in Malaysia on STRs

is very few (Seah at al., 2003; Paneerchelvam et al., 2003). Hence in the present study allelic distribution for three validated STRs – F13A01, FESFPs and vWA for Dusun population group was studied.

Although the primary focus of this review is on autosomal STR loci that are widely used for human identity testing, Y chromosome STR loci are growing in popularity. The Y chromosome is found only in males, and therefore genetic markers along the Y chromosome can be specific to the male portion of a male–female DNA mixture such as is common in sexual assault cases. Y chromosome markers can also be useful in missing persons investigations, some paternity testing scenarios, historical investigations, and genetic genealogy, because of the fact that most of the Y chromosome (barring mutation) is passed from father to son without changes. A core set of Y-chromosome STR (Y-STR) loci is widely used in laboratories worldwide for human identity testing and genetic genealogy (Butler JM., 2003).

AIMS AND OBJECTIVES

Study of DNA population in this country has increasingly for several years ago. The government and also the researchers have realized the essential to gather all the data about all the population in this country. Therefore, this project was designed to make up the database of Dusun population, one of the largest ethnic in Sabah, Malaysia. The objectives of the project include;

I. Understand the application of rules in genetics and statistical test.

II. Compiling the distributions pattern of various alleles for F13A01, FESFPS and vWA STRs in random Dusun population of Malaysia for the use on forensics.

Figure 1 : Location of Dusun Population in Sabah, Malaysia.

METHODOLOGY

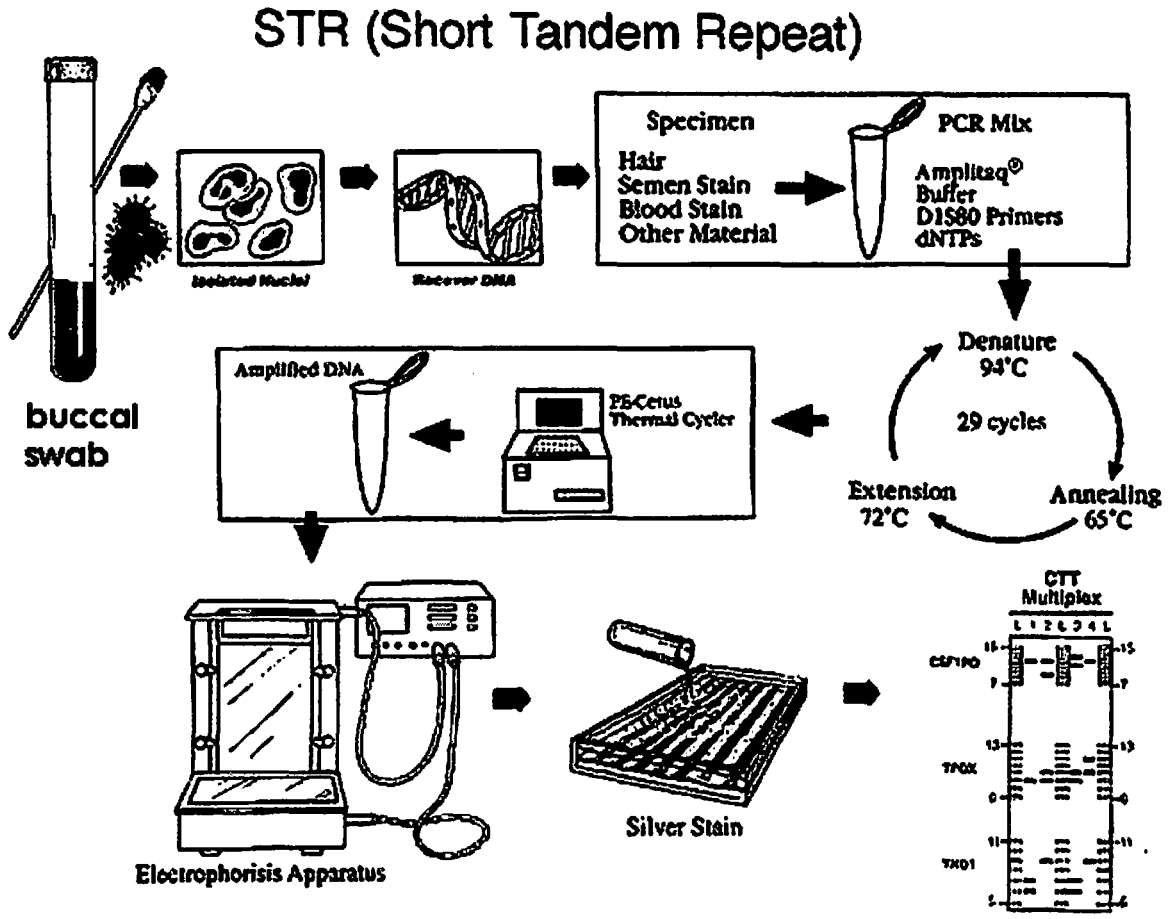

Figure 2: Procedures of STR

100 buccal samples from Dusun population → PCR amplification of extracted DNA samples → Using multiplex PCR primer kits → Electrophoresis of amplified DNA with allelic ladder → Detection by silver staining → Documentation → Statistical analysis

MATERIALS AND METHOD

1.0 Materials

Sterile reagents and materials used in this study. All the procedure of short tandem repeat (STR) typing was practiced according to the guidelines of manufacturer of STR kits (Gene Print STR System Technical Manual D004, Promega)

1.1 Sample Source

Buccal swabs of 100 unrelated ethnic Dusun individuals of Malaysia were collected by using sterile cotton buds. Each individual was asked to swipe the cotton buds inside their mouth at both cheeks for at least 10 seconds. After that, the cotton buds were air dried and kept away from direct sunlight in an envelope with the subject's detail including full name, age, sex and the state they are from.

1.2 Reagents

Digestion buffer (1M Tris HCl pH 7.5, 0.5M EDTA, 20% SDS, 5M NaCl), proteinase K (20µg/µl) (Promega,USA), chloroform-isoamyl-alcohol (24:1), 3M sodium acetate anhydrate (Fluka Garante), 2M sodium acetate, 70% ethanol, Tris-EDTA buffer, 10% NaOH, 0.5% acetic acid in ethanol, acetic acid, bind saline, Q421A, MSDS (Promega,USA), 10% ammonium persulphate, 40% acrylamide, 10X TBE buffer, 0.5X TBE buffer, distilled water, deionized water, FFv Multiplex kit consisting of STR 10X buffer (500mM KCl, 100mM Tris-HCl (pH 9), 15nM MgCl2, 1% Trito X-100,2mM each dNTP), multiplex 10X FFv primer mix, Taq DNA polymerase (1µ/µl), STR 2X

Loading Solution (10X NaOH, 95% formamide, 0.05% bromophenol blue, 0.05% xylene cyanol FF), FFv Allelic Ladder Mix (Promega,USA), Silver staining solution, Fix/Stop solution, developer solution and Rain X (Blue Coral-Slick 50, USA).

1.3 Chemicals

Tris base (Promega,USA), concentrated HCL, Na2EDTA (Promega,USA), NaOH (Merck,Germany), NaCI (Merck,Germany), sodium dodecyl sulphate (SDS) (Bio-Rad lab.), chloroform (Merck,Germany), isoamyl alcohol (Merck,Germany), absolute ethanol (Merck,Germany), glacial acetic acid (Merck,Germany), absolute ethanol (Merck,Germany), EDTA (Promega,USA), ammonium persulphate (Promega,USA), bisacrylamide (Promega,USA), boric acid (Promega,USA), urea (Promega,USA), silver nitrate (Promega,USA), 37% formaladehyde (Promega,USA), sodium thiosulphate (Promega,USA), sodium carbonate (Promega,USA), TEMED (N,N N'N' tetramethylethylene diamine) (Promega,USA), phenol (PIERCE,USA), ethidium bromide (Sigma,USA), Agarose (Promega,USA) and Orange-G dye (Sigma,USA)

1.4 Apparatus

Desiccators Nucerite (Nalgene/syborn Corp), Vortex mixer EVM-6000 (ERLA), Spectrafuge 16M (National Labnet Co., USA), Gilson varying volume pipette (France) (1000µl, 200µl and 10µl), parafilm "M" (Pechiney Plastic packaging, Chicago), MJ Research PTC-200 Peltier Thermal Cycler (Gradient cycler) used for 0.2µl tubes, MJ Research PTC-100TM Peltier Thermal Cycler used for 0.5µl tubes, monopan balance sensitive for smaller weights (DRAGON 204 Mettler Toledo), Fisher & Pagkal N308

Freezer, Hitachi refrigerator, Mammart waterbath, SA 32 Electrophoresis apparatus (GIBCO BRL Sequencing System), High voltage powerpack EC 3000-90 (E-C Apparatus Co., USA), fumehood (Model : RICO), plastic trays for staining, stopwatch, Greiner bio-one filter tip 100, Microwave (Sanyo), Elite 300 Plus (Wealtec), Photo printer P91D UVP (Mitsubishi), Software on Chemi System, UVP (Bioimaging system), UV Transilluminator, EpiChemi Darkroom, UVP (Bioimaging system), pH Cyberscan 1000, Model RS232 Meter (Eutech Instruments, Singapore), Hot plate and magnetic stirrer, EMS-HP-7000 (ERLA), Disposable latex/vinyl examination gloves, MF 22 (Fiocchetti, Italy), DuranFavorit WHL (Genrito Ltd, England), E-pure machine (Barnstead), Diamond aluminium foil (Renaults Consumer Products, USA), Sharps collector (Dispo-Med, Malaysia), 3M ComplyTM Indicator Tape, C-Fold Hantowels (Scott), Penguin Double Clips No.0222, B51 (China), Multipurpose assorted buds (Wang Zheng Co., Malaysia), Eve-Tape, 50ml Syringe W12875 (B-D, Singapore), Pharmacia Gel Electrophoresis Apparatus GNA-100 (Pharmacia Biotech, Sweden), short and long glass plate, Spacers, Combs, Beakers, Conical flask (Asahi Techno Glass), WTF Binder 7200 (Tuttlingen, Germany) and Autoclave: Steam Sterilizer Model MC-30LDP (ALP Co., Japan)