GENETIC PROFILING OF FIVE MAJOR GHANA POPULATION USING SHORT TANDEM REPEAT AND MITOCHONDRIAL DNA SEQUENCES FOR FORENSIC AND ANCESTRY STUDY PURPOSES

EDWARD KOFI ABBAN

UNIVERSITI SAINS MALAYSIA

2023

GENETIC PROFILING OF FIVE MAJOR GHANA POPULATION USING SHORT TANDEM REPEAT AND MITOCHONDRIAL DNA SEQUENCES FOR FORENSIC AND ANCESTRY STUDY PURPOSES

by

EDWARD KOFI ABBAN

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

February 2023

ACKNOWLEDGEMENT

My furthermost appreciation goes to the Almighty God for the opportunity to undertake this study. I also thank my principal supervisor Assoc. Prof. Dr Edinur Hisham Atan for readily accepting to oversee this study and assisting with the planning and design of the entire study.

I thank my second supervisor Dr Abd Rashid Nur Haslindawaty for her direction during the preparation and completion of the mtDNA project of this study. I likewise thank my third and fourth supervisors; Professor Shaharum Shamsuddin and Dr Anita Ghansah for consenting unreservedly to co-supervise this study.

I also thank Assoc. Prof. Dr. Ahmad Fahmi Lim Abdullah for assisting me with admissions to this program and to SUPT./Mr Hakim Mohd Hashom and Hajar who in many ways assisted me to complete this study.

I am also thankful to the Inspector General of Police, Ghana, for granting me the permission to undertake this study and to Inspector General of the Royal Malaysian Police for granting me the permission to use their facility at the DNA Databank.

I thank my main sponsors the Ghana Educational Trust Fund for providing funds for this study and to the Universiti Sains Malaysia for the research grants.

I devote this work to my folks Henry Richard Abban and Augustina Sika-Goh, my dazzling spouse Heckel Dokyi Amoabeng.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AMOVA	Analysis of Molecular Variance
AP	Acid phosphate
bp	Base pair
CID	Criminal Investigation Department
CLAMPK	Custer Markov Packager Across K
CODIS	Combined DNA Index System
CR	Control Region
CRS	Cambridge Reference Sequence
CTPI	Combined Typical paternity index
DC	Discrimination Capacity
ddH_2O	Deionized distilled water
DNA	Deoxyribonucleic acid
dNTPs	Deoxynucleotides triphosphate
EDNAP	European DNA Profiling Group
EDTA	Ethylene diamine tetra-acetic acid
EMPOP	European Mitochondrial DNA Population Database
ENFSI	European Network of Forensic Science Institutes
EtOH	Absolute ethanol
EU	European Union
FDNAL	Forensic DNA Laboratory
FORSTAT	Forensic Statistics Analysis Toolbox
FSL	Forensic Science Laboratory
F_{ST}	Fixation index
F_{STi}	Population specific indices
GD	Genetic Diversity
gDNA	Genomic DNA
GPS	Ghana Police Service
HD	Haplotype diversity
HF	Haplotype frequency
HMP	Haplotype Match Probability
HMW	High Molecular Weight

HVS-I	Hypervariable segment I
HVS-II	Hypervariable segment II
HVS-III	Hypervariable segment III
HWE	Hardy-Weinberg equilibrium
INDELS	Insertion/deletion polymorphisms
INFORMM	Institute for Research into Molecular Medicine
INTERPOL	International Police
IPS	Institut Pengajian Siswazah
ISFG	International Science for Forensic Genetics
LD	Linkage Disequilibrium
LE	Linkage Equilibrium
MDS	Metric Multidimensional Scaling
$MgCl_2$	Magnesium Chloride
MLP	Multi Locus Probe
MP	Matching Probability
MSY	Male-Specific Region
MVSP	Multivariate Statistical Package
Na_2EDTA	Disodium salt of ethylene diamine tetra-acetic acid
NMIMR	Noguchi Memorial Institute of Medical Research
NRPY	Non-recombining portion of Y chromosome
PAR	Pseudoautosomal Region
PCoA	Principal Coordinate Analysis
PCR	Polymerase Chain Reaction
PD	Power of Discrimination
PE	Power of exclusion
PIC	Polymorphic Information Content
PSA	Prostate-specific antigens
psi	Pound per square inch
RFLP	Restriction fragment length polymorphism
RMP	Random Match Probability
RNA	Ribonucleic acid
Sec	Second
SLP	Single Locus Probes
SNP	Single Nucleotide Polymorphism

LTRs	Simple tandem repeats
SWGDAM	Scientific Working Group—DNA Analysis Methods
TBE	tris borate EDTA
TPI	Typical Paternity Index
USM	Universiti Sains Malaysia
VNTR	Variable number tandem repeats
YHRD	Y-STR Haplotype Reference Database

LIST OF APPENDICES

PEMPROFILAN GENETIK LIMA POPULASI UTAMA DI GHANA MENGGUNAKAN PENJUJUKAN PENDEK BERULANG DAN JUJUKAN DNA MITOKONDRIA UNTUK APLIKASI FORENSIK DAN KAJIAN ASAL USUL KETURUNAN

ABSTRAK

Set-set data populasi bagi jujukan pendek berulang (STR) dan bahagian DNA mitokondrion (mtDNA) yang relevan dengan forensik daripada suatu kumpulan populasi tertentu diperlukan sebelum sebarang keputusan ujian profil DNA boleh diserahkan ke mahkamah sebagai bukti. Set-set data populasi ini terbukti berguna untuk mengira kebarangkalian padanan dalam kes-kes forensik atau pertikaian kebapaan, danuntuk kajian genetik sesebuah populasi. Walaubagaimanapun, set data populasi yang mewakili populasi Afrika termasuk di Ghana kurang berbanding populasi-populasi di Eropah dan Asia. Oleh itu, kajian ini dijalankan untuk menyediakan buat kali pertamanya set data populasi STR dan mtDNA yang disaring daripada set individu yang sama yang mewakili subpopulasi Akans, Ewe, Ga-Dangbe, Mole-Dagbon dan Guang di Ghana. Dua puluh satu lokus autosomal STR di dalam 515 sampel DNA daripada sub-populasi ini telah dijeniskan menggunakan *kit amplifikasi PCR Investigator 24plex*. Kebarangkalian padanan identiti berjulat antara 1 dalam 1.30×10^{-25} hingga 6.28×10^{-26} dan paduan kuasa diskriminasi berjulat antara 0.9999999999468 hingga 0.99989999 untuk 5 sub-populasi utama di Ghana. Jarak genetik, serta analisa pokok filogenetik dan penyelaras utama (PCoA) mendedahkan yang kelima-lima populasi lebih dekat antara satu sama lain dan dengan populasi jiran, berbanding populasi di lokasi yang lebih jauh. Individu lelaki yang tidak mempunyai pertalian (n=268) kemudiannya dicirikan secara genetik untuk 23 lokus kromosom Y STR

menggunakan kit *Powerplex Y23 STR*. Kepelbagaian haplotip, keupayaan mendiskriminasi dan kebarangkalian sepadan untuk data populasi yang terkumpul adalah masing-masing 0.9998, 0.9627 dan 0.0039. Jarak genetic berpasangan (R_{ST}) untuk set-set data Ghana dan populasi rujukan lain yang disimpan dalam *Y-STR Haplotype Reference Database* telah dianggarkan dan dipetakan menggunakan plot pengskalaan pelbagai dimensi (MDS). Guan dan Ewe berbeza secara signifikan berbanding Akans, Mole-Dagbon dan Ga-Dangme, tetapi kelima-lima subpopulasi ini terletak rapat dengan populasi Afrika yang lain dalam pemetaan data MDS. Jujukan seluruh kawasan kawalan mtDNA diperoleh menggunakan kaedah penjujukan *Sanger*. Dapatan kajian menunjukkan campuran jenis-jenis mtDNA yang diperoleh daripada populasi Afrika (97.59%) dan sebahagian kecil daripada populasi Asia (0.48%) dan Eropah (1.93%). Nilai kepelbagaian genetik yang tinggi (0.9994), kuasa diskriminasi (0.9991) dan nilai kebarangkalian padanan rawak yang rendah (0.054) menunjukkan bahawa analisis mtDNA untuk populasi ini boleh digunakan secara efektif dalam kes forensik. Pewarisan susur galur mtDNA yang paling kerap adalah L (97.59%) dan diikuti oleh U (1.45), N (0.48) dan X (0.48) dan subpopulasi Ghana berkait secara genetik dengan populasi Afrika Barat di dalam pemetaan data PCoA. Secara keseluruhan, kajian ini telah berjaya menjeniskan serta membangunkan dataset STR dan mtDNA untuk lima kumpulan subpopulasi di Ghana dan keputusan statistik menunjukkan bahawa kedua-dua penanda adalah berkesan untuk pemprofilan DNA forensik dan untuk mengkaji komposisi genetik di Ghana. Kajian pada masa hadapan harus menumpukan pada aspek lain dalam pemprofilan DNA seperti pengesahan kit pengekstrakan DNA dan kit genotip baru serta pembangunan data populasi bagi kumpulan populasi yang masih tidak dijeniskan di Ghana.

GENETIC PROFILING OF FIVE MAJOR GHANA POPULATION USING SHORT TANDEM REPEAT AND MITOCHONDRIAL DNA SEQUENCES FOR FORENSIC AND ANCESTRY STUDY PURPOSES

ABSTRACT

Population datasets of forensically relevant short tandem repeat (STR) and mitochondrial DNA (mtDNA) from a particular population group are needed before any DNA profile test results can be reliably submitted into evidence before the court. These population datasets are proven useful for calculating match probabilities in forensic or disputed paternity cases and for population genetic studies. However, autosomal and Y-chromosome STR and mtDNA sequence datasets for representative African populations, including Ghana, are lacking compared with European and Asian populations. Therefore, the present study is conducted to provide the first-ever STR and mtDNA population datasets that were screened from the same set of individuals representing the Akans, Ewe, Ga-Dangbe, Mole-Dagbon and Guang sub-populations in Ghana. Twenty-one autosomal STR loci in 515 genomic samples of these sub-population groups were typed using Investigator 24plex PCR amplification kit. The match probability of identity ranged from 1 in $1.30x10^{-25}$ to $6.28x10^{-26}$, and the combined power of discrimination ranged from 0.9999999999468 to 0.9999999999864. Genetic distances, phylogenetic tree and principal coordinate analysis (PCoA) revealed that the five populations are genetically closer to each other and neighbouring populations than distant localities. Unrelated males (n=268) were then genetically characterized for 23 Y-chromosome STR loci using Powerplex Y23 STR kit. The haplotype diversity, discriminating capacity and match probability for the pooled population data were 0.9998, 0.9627 and 0.0039, respectively. The pairwise

genetic distance (R_{ST}) for the Ghanaian datasets and other reference populations deposited in Y-STR Haplotype Reference Database were estimated and mapped using multidimensional scaling (MDS) plot. The Guan and Ewe were significantly different from the Akan, Mole-Dagbon and Ga-Dangme, but were all plotted closely with reference African populations in the MDS data mapping. The entire mtDNA control region in 207 unrelated individuals of the five major sub-populations of Ghana were obtained using Sanger sequencing method. Results showed an admixture of mtDNA types derived mainly from Africans (97.59%) and a minor proportion from Asians (0.48%) and Europeans (1.93%). A high value of genetic diversity (0.9994), power of discrimination (0.9991) and low value of random match probability (0.054) indicate that mtDNA analysis for this population can effectively be used for forensic casework. The most frequent mtDNA lineages were L (97.59%) and followed by U (1.45), N (0.48) and X (0.48), and Ghanaian sub-populations are closer to West African populations in PCoA data mapping. Overall, the present study has successfully typed and developed STR and mtDNA datasets for the five sub-population groups in Ghana. The statistical results showed that both markers are reliable for forensic DNA profiling purposes and for studying genetic makeup in Ghana. Future studies should focus on the other aspects of DNA profiling, such as validating newly DNA extraction and genotyping kits and developing population data (STR, mtDNA or other potentially relevant markers) for the remaining uncharacterized population groups in Ghana.

CHAPTER 1

INTRODUCTION

1.1 Introduction

In 1985, Sir Alec Jeffreys, an English geneticist, introduced the term DNA fingerprinting into forensic science, describing a technique used to examine the length variations of repeated deoxyribonucleic acid (DNA) sequences which later became known as variable number tandem repeats (VNTRs). In this technique, a restriction enzyme was used to cut at specific regions in the DNA. The use of restriction enzymes as part of the DNA fingerprinting technique earned the technique another name, restriction fragment length polymorphism (RFLP). The introduction of RFLPs in human identity greatly impacted forensic science, with several forensic laboratories and paternity testing laboratories conducting human identification testing based on this technique.

However, before this technology was made available, identification of biological evidence mostly depended upon the analysis of blood group marker systems, which had several limitations, including a limited supply of reagent and materials and not being suitable for highly degraded casework samples. For example, conventional blood typing methods depend on the availability of a good quality of blood or body fluid samples from the crime scene, while DNA fingerprint can be performed on the trace of nucleated biological materials such as saliva, hair root and shaft, nail and semen (Kuperus *et al.,* 2003; Dash *et al.,* 2020). Hence this technology has reduced the requirement for large amounts of biological materials and improved sensitivity in the analysis of crime scene samples which are mostly present in small amounts.

The term DNA profiling is preferred over DNA fingerprinting because the first describes a more advanced typing method for genotyping biological materials. Accordingly, each technique and marker system used in DNA profiling presents various applications, including in forensics and population genetics. For example, several autosomal short tandem repeat (biparental STR) loci can be simultaneously typed and widely adopted to individualise casework samples (Liu *et al.*,2013; Pajnič *et al.*, 2017; Ghiani *et al.*, 2019). In contrast, paternally inherited Y chromosome STRs (uniparental) and maternally inherited mitochondrial DNA (uniparental mtDNA) are commonly screened in sexual assault cases and the analysis of degraded samples (Holland, *et al.*, 1993; Prinz & Sansone, 2001; Cerri *et al.*, 2003; Purps *et al.*, 2015). Subsequently, this study will generate three unique datasets for the statistical evaluations of autosomal STR, Y- chromosome STR and mitochondrial DNA profiles (Nuclear markers) for forensic applications such as human identification, paternity, rape and population studies for the Ghanaian populace (Figure 1.1).

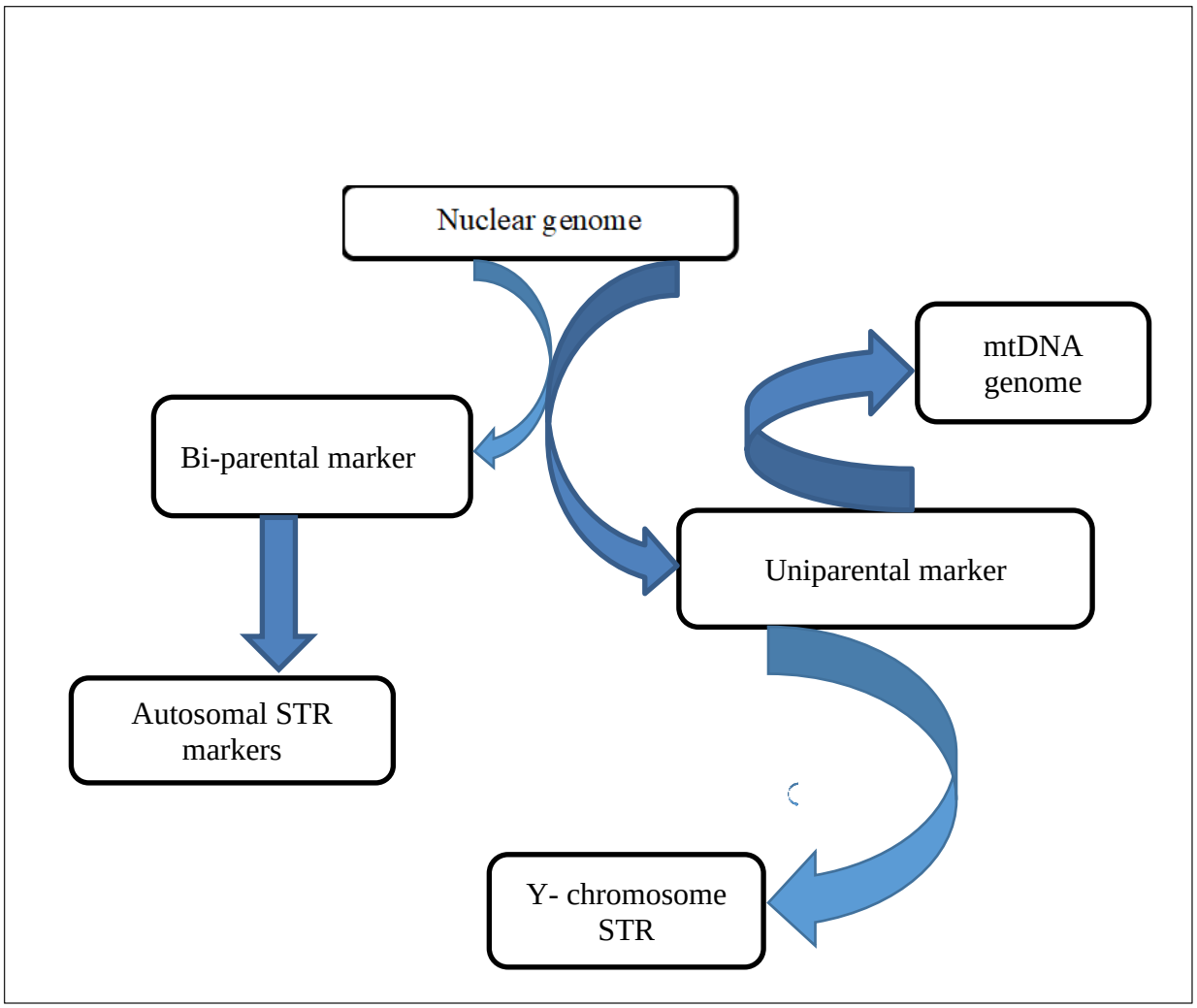

Figure 1.1 Relationship between the three genetic markers (autosomal STR, Y-chromosome STR and mtDNA) used in the study

1.2 Problem statement

DNA profiling is a method for sample individualization and is widely used for legal purposes. In the last three decades, many significant advancements have been made on DNA profiling, including genotyping protocols and the selection of genetic markers (refer to sub-section 2.3). Therefore, it is crucial to address technical and operational standard issues related to DNA profiling. This includes population databases for determining the probative value of DNA evidence forwarded to court developed in this study (Board, 2000; Gusmão *et al.,* 2006; Carracedo *et al*., 2010; SWGDAM, 2010; Roewer *et al.,* 2020).

In Ghana, population STR (both autosomal and Y chromosomes) and mtDNA data are available for a limited number of population groups (refer to sub-section 2.6 for a demographic profile in Ghana). These include previously reported studies on X-chromosome STR in Ewes (Poetsch *et al.,* 2011), mtDNA in Akans (Fendt *et al.,* 2012) and mtDNA and Y-chromosome and autosomal STR in the Bimoba clan in North-Eastern Ghana (Thiele *et al.,* 2008; Poetsch *et al.,* 2009). Without a complete set of STR and mtDNA population data from well-characterized population groups in Ghana, the rarity of any DNA profiles (e.g. 1 in million or 1 in billion) obtained either from the crime scene of disputed paternity testing cases cannot be statistically estimated. In this context, STR and mtDNA population data for assessing the probative value for Ghanaian specific population groups cannot be obtained or inferred from frequency data of other populations, including from other African-ancestry populations, as each population has a different genetic make-up that is shaped by local population history (e.g. admixture) and natural selection (Tishkoff *et al*., 2009; Tau *et al.,* 2017).

1.3 Hypothesis

i. Short tandem repeat and mitochondrial sequence data for the five majors populations of Ghana are useful for forensic analysis in Ghana.

ii. Short tandem repeat and mitochondrial sequence data for the five major populations of Ghana are useful for ancestry study purposes in Ghana

1.4 Significance of the study

As stated earlier (sub-section 1.2), population STR and mtDNA data are highly relevant to correctly assigning the weight of DNA profiles in the courtroom. Therefore, the present study was conducted to develop and establish for the first time a complete set of STR (autosomal and Y-chromosomes) and mtDNA population data for the five major Ghanaian sub-populations (Akans, Mole-Dagbons, Ewes, Ga-Dangbes and Guans). Previous study was limited to mtDNA involving a relatively smaller number of Akan (Asante Akyem) individuals (Fendt *et al.*, 2012). However, mtDNA data from this earlier study provide an interesting basis for comparisons with those generated from this study (refer to sub-section 6.5 – chapter result mtDNA).

In addition, the nature of genetic markers (i.e. STR and mtDNA) tested in some sub-populations of the Akans, Mole-Dagbons, Ewes, Ga-Dangbes and Guans are widely used for population genetics (Thiele *et al.*, 2008; Poetsch *et al.*, 2009; Nur Haslindawaty *et al*., 2010; Fendt *et al*., 2012; Sanches *et al*., 2014; Kareem *et al*., 2015; Tran *et al*., 2019). Therefore, autosomal, Y-chromosome STR and mtDNA population data generated for Akans, Mole-Dagbons, Ewes, Ga-Dangbes and Guans will be subjected to ancestral analyses and is anticipated to reveal ancestries and past episodes of settlement movements in Ghana. Furthermore, as part of quality assurance and data

sharing, genotyping results for Y-STRs and mtDNA will be submitted for quality control checks and acceptance by the International Y-chromosome haplotype database (YHRD) and the European mitochondrial DNA population database (EMPOP), respectively to increase the number of data available to the forensic community.

1.5 Objectives of the Study

1.5.1 General objective

To develop autosomal and Y-chromosome STR and mtDNA population datasets for the five major Ghanaian sub-populations.

1.5.2 Specific objectives

i. To genotype 21 autosomal and 23 Y chromosome STR loci in five Ghanaian sub-populations.

ii. To provide forensic parameters of 21 autosomal and 23 Y chromosome STR loci for the five Ghanaian sub-populations.

iii. To sequence control region (hypervariable regions I, II and III) of the mitochondrial DNA in the five Ghanaian sub-populations using Sanger Sequencing.

iv. To compute forensic parameters of the mtDNA control region for the five Ghanaian sub-populations.

v. To examine genetic relationships of the five Ghanaian sub-populations and other reference populations.

vi. To evaluate the current account of population origins and settlements in Ghana based on STR and mtDNA data of the five Ghanaian sub-populations.

CHAPTER 2

LITERATURE REVIEW

2.1 Human genome

The genome is the entire genetic code in an organism's cell. In humans, the genome is organized into 23 pairs of chromosomes in the cell nucleus and as a circular chromosome located in the matrix of the mitochondria (Butler, 2005; Chu and Giles, 1959). The nuclear genome size is approximately 3.2 billion base pairs (bp), whereas the mitochondrial genome is 16,569 bp (Butler, 2005; Holland and Lauc, 2014; Bosworth *et al.*, 2017). Here, nuclear and mitochondrial genomes will be described as the present survey involves molecular screening of short tandem repeats (STRs) located on autosomal and Y chromosomes and hypervariable regions on mitochondrial DNA (mtDNA).

About 5% of the genomes is called the coding region, which contains sequences that primarily determine the amino acid sequences of proteins. In contrast, the other 95% is referred to as a noncoding region with an unknown function (Lu *et al.*, 2015; Buckingham, 2012). The noncoding region of the genome is characterized by high mutation rates leading to increased genetic polymorphisms compared to the coding region that is genetically stable with minimal mutations. They are located throughout human chromosomes and significantly vary between individuals, groups or populations (Wolfe *et al.*, 1989; Di Iulio *et al.*, 2018). It is important to note that several coding regions like those that determine human leukocyte antigen, cytokine and blood group specificities are polymorphic. These together with those polymorphic motifs in the non-coding regions are widely adopted for mapping of diseases, population genetic studies and human identification (Sachidanandam *et al.*, 2001; International HapMap

Consortium, 2005; International HapMap Consortium, 2007; Sato *et al.,* 2010; Keshavarz *et al.,* 2019; Jinam *et al*., 2022). The organizational structure and variability in nuclear and mitochondrial genomes are described in the following sub-section.

2.1.1 Nuclear genome

The human nuclear genome is organized into 22 pairs of autosomes and one pair of sex chromosomes (Figure 2.1). The X and Y chromosomes make up the sex pair (Dash *et al.,* 2018). Scientists estimate that the human genome has about 20,000 to 25,000 protein-coding genes. However, the vast majority comprises non-coding DNA (genes only account for ~ 1.5% of the total sequence). Historically referred to as 'junk DNA', these non-coding regions are now recognised to serve other important functions (Gregory, 2005; Palazzo & Gregory, 2014). Examples include satellite DNA, telomeres, introns, ncRNA genes and gene regulatory sequences. DNA profiling is a technique by which individuals can be identified and compared via their respective DNA profiles (Butler, 2005). Within the non-coding regions of an individual's genome, there exists satellite DNA – long stretches of DNA made up of repeating elements called short tandem repeats (STRs). Tandem repeats can be excised using restriction enzymes and then separated with gel electrophoresis for comparison. Individuals will likely have different numbers of repeats at a given satellite DNA locus so that they will generate unique DNA profiles.

Longer repeats will generate larger fragments, while shorter repeats will generate smaller fragments (Ruitberg *et al.,* 2001, Butler, 2005; Hill *et al.,* 2009). STRs of the autosomes are the most popular STR markers in the field of forensic science with varying applications such as paternity testing, forensic DNA profiling (suspects and victims, missing persons) and population studies (Fan *et al.,* 2019; Li *et al.,* 2019; Srivastava *et al.,* 2019). Hence STR of the autosomes of individuals will be investigated

and complemented with paternal and maternal inherited Y-chromosome and mitochondria DNA to determine suitability of these autosomes and uniparental markers for forensic purposes and population genetic study in Ghana.

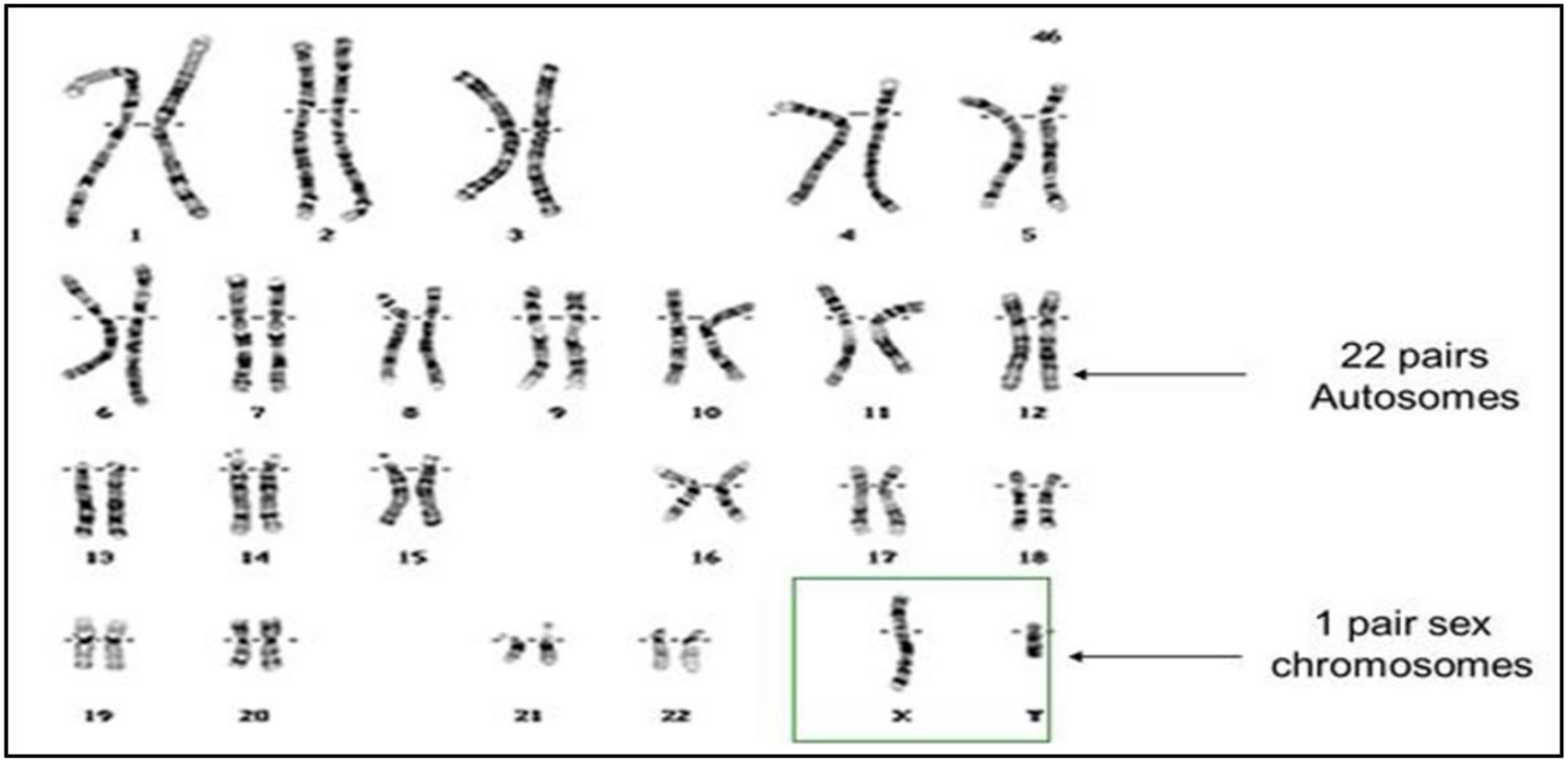

Figure 2.1 Karyotype of 22 pairs of autosomal chromosomes and one pair sex chromosome. This figure is edited from httpts://en.wikipedia.org/wiki/Autosome

2.1.1(a) DNA structure and organization

Nuclei acid, including DNA, comprises units of nucleotides of three parts: nucleobase, sugar, and phosphate (Figure 2.2). The nucleobase or 'bases' imparts the variation in each nucleotide unit, while the phosphate and sugar portions make the DNA molecule's backbone structure. The structure of DNA molecule was described to contain two complementary linear polynucleotide chains consisting of four types of monomeric nucleotides, namely adenine (A), cytosine (C), guanine (G) and thymine (T), over 60 years ago (Watson & Crick, 1953; Avery *et al.*, 1944; Chargaff *et al.,* 1951). Most importantly, the structure implied that the information in the DNA base sequence could possess the ability to be replicated into two identical copies under complementarity. This insight into the fundamental basis of genetics has underpinned the immense advances in genetic understanding and manipulation over the past 60 years. The various combinations of these four letters known as the nucleotides or 'bases' yield diverse variations among human beings. Humans have approximately 3 billion nucleotide positions in the genomic DNA; hence at four possibilities at each position, zillions of combinations are possible (Butler, 2009).

The deoxyribose is attached to the nitrogen of a base and the phosphate group to the deoxyribose. The individual nucleotides are held together by phosphodiester bonds. The two complementary linear polynucleotide strands are antiparallel: One strand is arranged 3 ' to 5 ' from left to right, while the other runs in the opposite direction, 5' to 3' from left to right. Base pairing of the strands involves forming hydrogen bonds that provide weak electrostatic attractions between electronegative atoms. The two sugar-phosphate backbones form the vertical double helix with the heterocyclic bases stacked horizontally in the centre (de Chadarevian & Kamminga, 2002). The two complementary DNA strands are coiled into a helical conformation stabilised

by chemical interactions and attractive stacking forces between the adjacent base pairs (Figure 2.3).

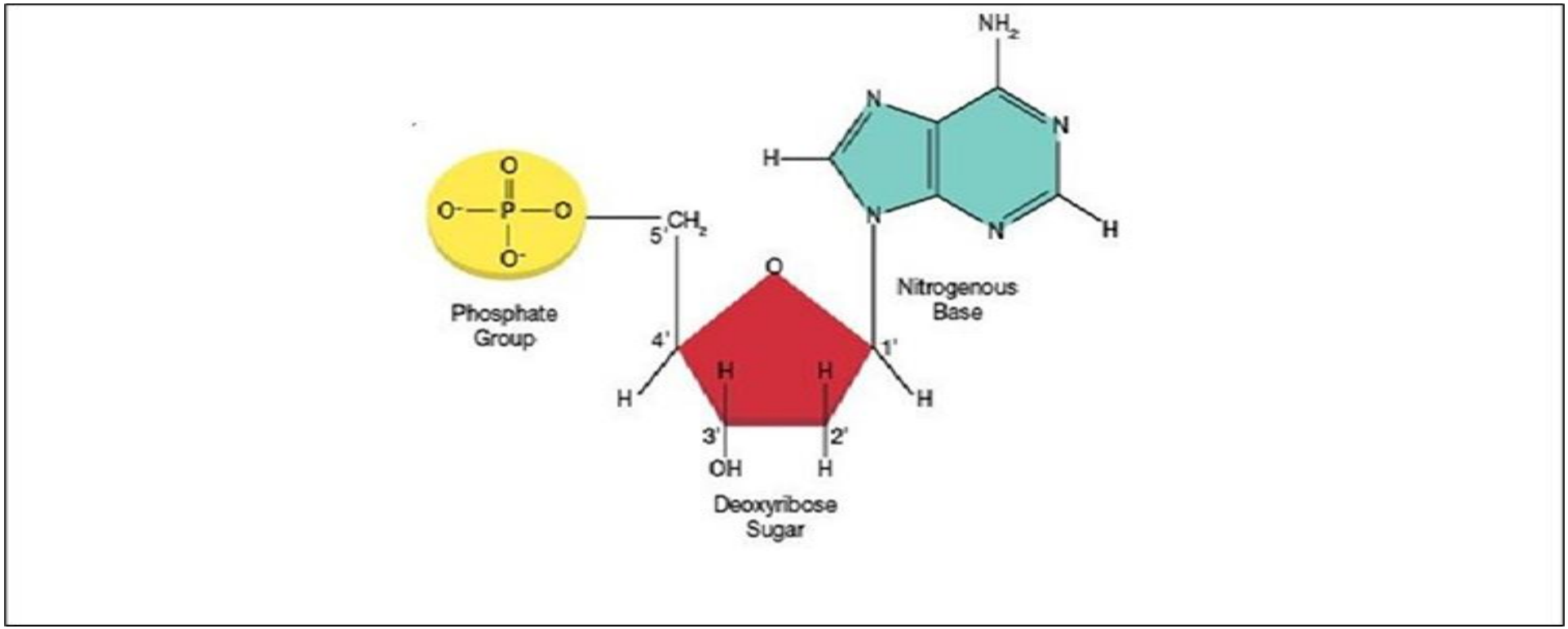

Figure 2.2 Parts of the nucleotide (this figure was adopted from https://www.pininterest.com.pin/29294515075199149 8/accessed 18th February 2022)

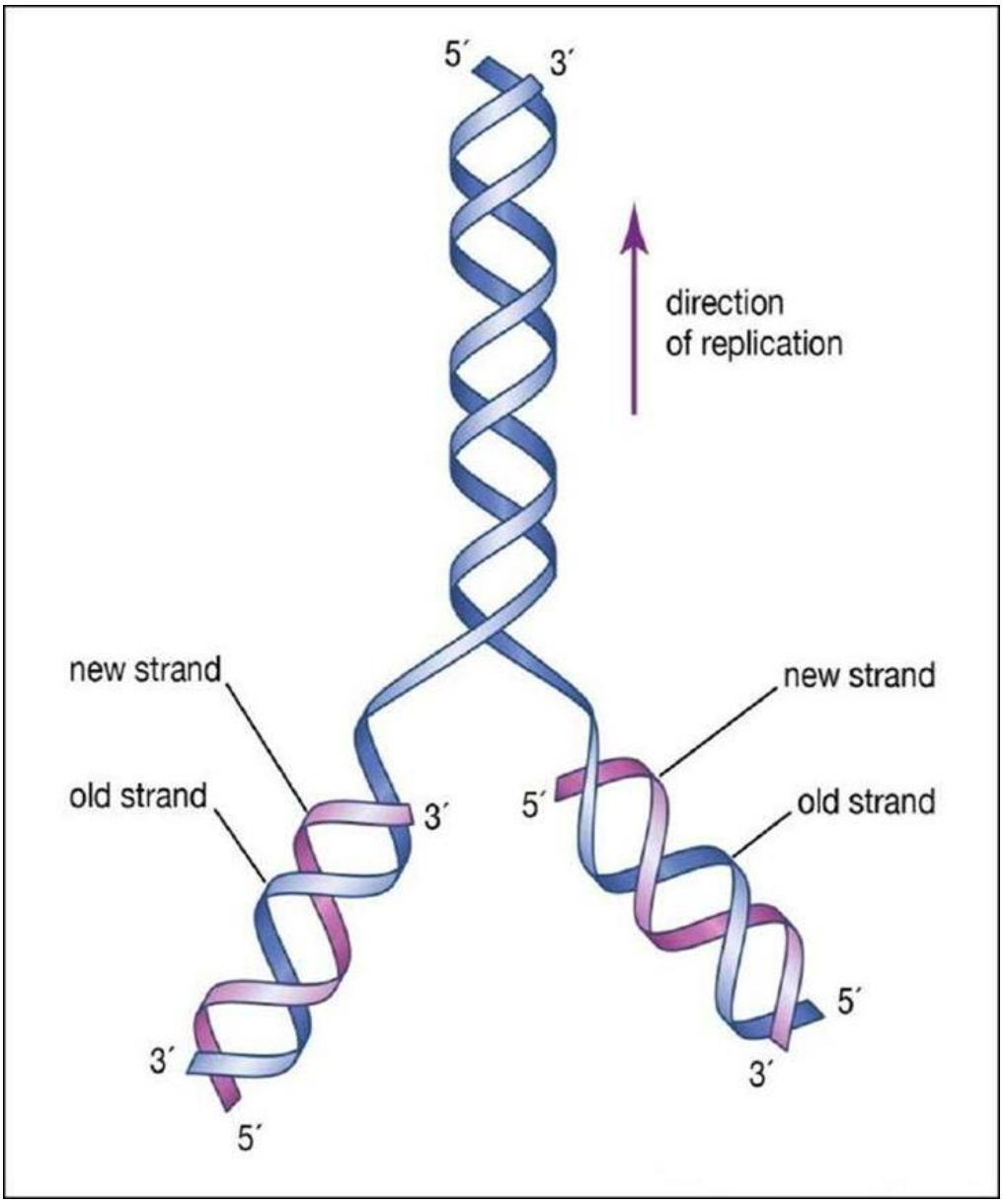

Figure 2.3 Double helix DNA structure (Roberts, Richard J. “nucleic acid”. Encyclopedia Britannica (this figure was adopted from https://www. britannica.com/science/nucleic-acid accessed 18th February 2022)

2.1.2 Mitochondrial genome

Mitochondria are double-membraned subcellular organelles present in high copy numbers (several hundred to several thousand) in the mammalian cell. They are the cell's powerhouse in which many crucial metabolic processes occur, such as oxidative phosphorylation (Wei, 1992; Holland & Parsons, 1999). Consequently, their unique genome (mtDNA) is distinct from the nuclear DNA. Each mitochondrion has more than two copies of the mtDNA; thus, there are, on average, thousands of copies of mtDNA in each somatic cell compared to the two copies of the nuclear genome (Wallace, 1997; Kobilinsky, 2005). The complete sequence of the human mtDNA was first obtained in 1981 (Anderson *et al.,* 1981). The revised version, where few corrections were made, is known as the Cambridge Reference Sequence (CRS) (Andrews *et al.,* 1999). This reference sequence is used to compare newly reported polymorphisms within mtDNA sequences. The mtDNA structure comprises two regions, namely the control and coding regions (Figure 2.4).

The control region (CR) is, also known as the noncoding region, includes the displacement-loop (D-loop), which is the most rapidly evolving part of mtDNA (Upholt and David 1977; Carracedo *et al.,* 2000; Herrnstadt *et al.,* 2002). This noncoding region is approximately 1,100 bp in length, situated between the mitochondrial proline and phenylalanine tRNA (tRNApro and tRNAphe). It contains the origin of replication for the heavy strand synthesis, both mitochondrial transcription promoters and served as the core site for mtDNA replication and transcription (Taanman, 1999; Chinnery, 2006; Kolesnikov & Gerasimov, 2012). The control region is highly variable and has an evolutionary rate nearly ten times higher than the gene coding region (Parsons and Coble, 2001). The variable sites in this control region consist of three hypervariable segments. Hypervariable segments I (HVS-I) range from position 16,024 to 16,365,

hypervariable segment II (HVS-II) extends from position 73-340 and hypervariable segment III (HVS-III) situated between position 438 to 574 (Greenberg *et al.,* 1983; Wilson *et al.,* 1993; Brandstätter *et al.,*2004).

Both HVS-I and HVS-II regions are well known as hotspots for base-pair substitutions (Wallace, 1995), presenting a high degree of discrimination for individuals of unrelated maternal lineages (Bär *et al.,* 2000; Just *et al.,* 2009). This feature is very useful in human identification and ancestral analysis (Gill *et al.,* 1994). HVS-III has a lower mutation rate than HVS-I, HVS-II and is mostly used as an additional polymorphic site to distinguish the same shared profiles of HVS-I and HVS-II individuals (Tamura and Nei, 1993; Meyer *et al.,* 1999). Hence, this study is designed to sequence hypervariable region I, II, III among the unrelated individuals in Ghana and determine their suitability for human identification and ancestral studies.

The CR is part of the mtDNA genome containing 37 genes that code for polypeptides essential for normal mitochondrial function. The 37 genes code for mRNAs of 13 polypeptides (Yusoff *et al.,* 2015) responsible for the oxidation phosphorylation process, two ribosomal RNAs (rRNA), and 22 transfer RNAs (tRNA) responsible for coding for polypeptides critical to the electron transport chain. The 13 polypeptides include seven subunits of complex I (ND1 to ND6), one subunit of complex III (cytochrome b), three subunits of complex IV (COX I, COX II and COX III) and two of complex V (ATPase 6 and 8).

The asymmetric distribution of guanine and cytosine permits the separation of mtDNA into "heavy" (H-strand) and "light" (L-strand) strands in alkaline density gradient centrifugation. The 2 rRNAs, 12 of the 13 polypeptides, and 14 of the 22 tRNAs are encoded by the heavy-strand genes (Anderson *et al*., 1981), whereas the light

strand codes for the remaining 8 tRNAs and ND6 polypeptide (Attardi and Schatz, 1988).

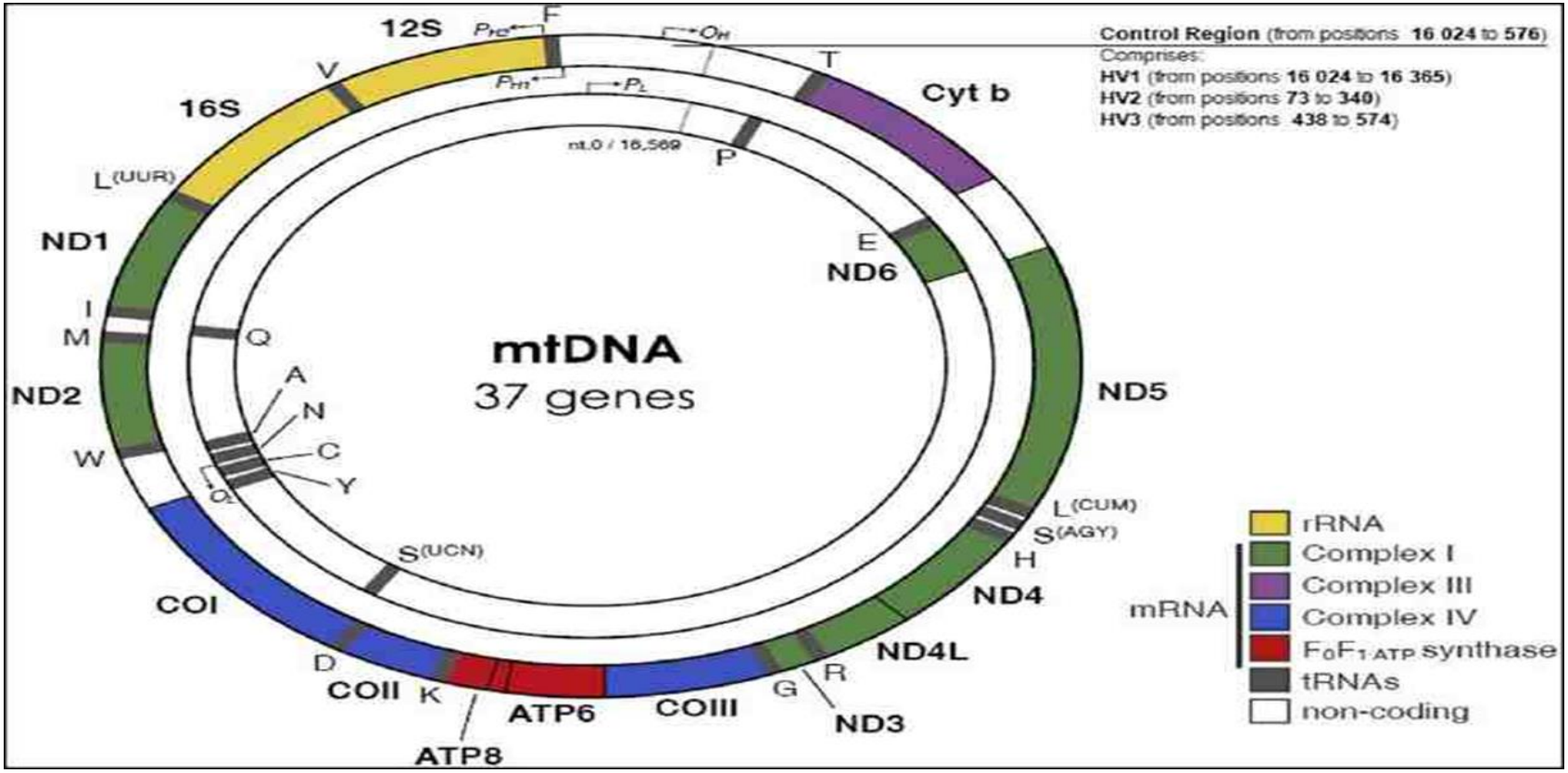

Figure 2.4 The human mitochondrial DNA genome with labelled control and coding (gene) regions. This figure is edited from Picard *et al.*, 2016

2.1.2(a) mtDNA heteroplasmy

Human mtDNA typically exists as monoclonal within an individual, but a condition called heteroplasmy does occur (Comas *et al.*, 1995). Heteroplasmy defines the presence of two or more types of mitochondrial DNA (mtDNA) within an individual. This condition is relatively rare and only about 10 to 20% of the population is found to exhibit this condition (Gibbons, 1998). The presence of heteroplasmy plays a role in the forensic interpretation of mtDNA matches between evidentiary materials and individuals or their maternal relatives, improving the probability of a match in forensic investigations (Melton, 2004; Lo *et al.*, 2005). In some cases, heteroplasmy can complicate the data interpretation of mtDNA analysis. A report had shown sequences from 12 hair shafts and a saliva sample from an individual differing at position 16,093. The saliva sample was homoplasmic at that particular site, while three of 12 hair shafts were heteroplasmic (C/T). If this were to occur in criminal cases, it would lead to false exclusion.

There are two forms of heteroplasmy; point heteroplasmy and length heteroplasmy. Point heteroplasmy occurs when an individual differs only at one nucleotide position in mtDNA molecules and length heteroplasmy is referred to as the variation of bases residing within a homopolymeric stretch (i.e., C stretches). Heteroplasmy at two sites in the same individual is also reported, a condition described as triplasmy (Tully *et al.*, 2000). Length heteroplasmy has been identified in mtDNA control regions HVS-I and HVSII due to the replication slippage mechanism (Bendall and Sykes, 1995). The variation in the length of homopolymeric C stretches begins at nucleotide position 16184 in HVS-I and 303 in HVS-II. The association between the presence of heteroplasmy and the length of the C tract at position 309 indicates that the

location of heteroplasmy to the C tract is causal. The localization of heteroplasmy to long repeated tracts and the association between heteroplasmy and longer tract length (such as at position 309) are heteroplasmic state is generated through slippage during DNA replication (Cavelier *et al.*, 2000; Alqaisi *et al.*, 2022). In particular, the C tract around position 309 appears to experience a "threshold" in stability. When expanding the tract beyond 7 repetitions with C, the stability decreases dramatically and strong segregation of heteroplasmy is seen between generations (Cavelier, *et al.*, 2000).

2.1.2(b) mtDNA haplogroup

Differences in human mtDNA are classified into haplogroups. A detailed catalogue of mtDNA haplogroups is recorded and deposited in a human mitochondrial genome database (MITOMAP), a database assigned to A to Z letters (Lott *et al.*, 2013). Each haplogroup was assigned in the order of its discovery and did not reflect the actual genetic relationships between populations. For example, those that are common to African populations are assigned as macro haplogroups L0, L1, L2, and L3 (Chen *et al.*, 2000; Salas *et al.*, 2002) while macrohaplogroup, L3 haplogroups M and N derived from Asian populations (Ballinger *et al.*, 1992; Torroni *et al.*, 1994; Wallace *et al.*, 1999). Equally, the M macrohaplogroup that gave rise to haplogroups that are frequent in the East Asian populations (A, B, C, D, G, and F) and Australasian populations (haplogroup S, P, and Q) were assigned based on the order of their discoveries, rather than the affinities between population groups (Schurr *et al.*, 1999; Torroni *et al.*, 1994). The world mtDNA human migration patterns inferred using mtDNA data and continent-specific haplogroups are shown in Figure 2.5.

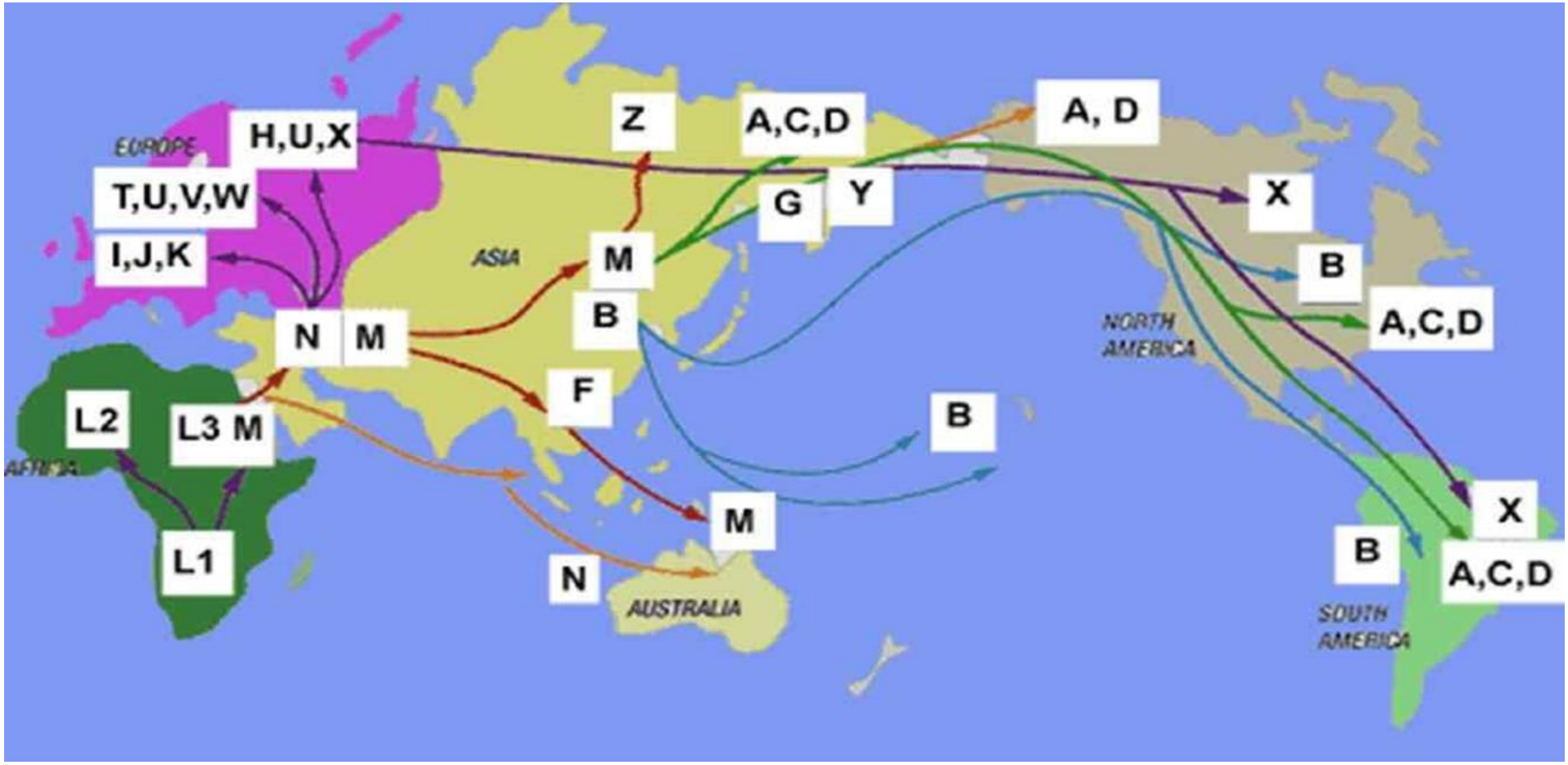

Figure 2.5 Directions of human migration patterns and continent specific haplogroups. This figure is edited from academic encyclopaedias (https://en-academic.com/pictures/enwiki/77/migration_map4)

2.2 Genetic markers in the genome

Polymorphic genetic markers, including single nucleotide polymorphisms, insertion/deletion polymorphisms, polymorphic repetitive sequences and structural and copy number variations and are described in sub-section 2.2.1 to 2.2.4.

2.2.1 Single nucleotide polymorphisms

A single nucleotide polymorphism (SNP) is defined as a single base change in genomic DNA where the sequence alternatives exist in normal individuals, and the least frequent allele has a frequency of at least 1%. The human DNA typically has one SNP for every 300 bases (Nelson *et al.,* 2004). Accordingly, there would be about 10 million SNPs for the entire genome's (3 billion bases). More than 60,000 SNPs are located in the coding regions of the genes (Sachidanandam *et al.,* 2001). The remainder constitutes the simplest and highest natural sequence changes in the noncoding region of our genome (Xu & Taylor, 2009). In the human population, more than 99% of the DNA is shared and SNPs account for 90% of the 1% that differs (Fairweather-Tait *et al.*, 2007). SNPs are used in a wide range of studies, including estimating the predisposition to a disease (Voronko *et al.*, 2008; Stefansson *et al.,* 2009), predicting specific genetic traits (Yip and Lange, 2011), predicting drug efficacy (Giacomini *et al.,* 2007) and tracking ancestral migration (Underhill and Kivisild, 2007). These earlier studies were conducted using SNPs located either in nuclear or mtDNA genomes.

2.2.2 Insertion/deletion polymorphism

Insertion deletion polymorphisms (INDELs) is a type of DNA variation in which a specific nucleotide sequence of varying lengths (i.e. less than 100 base pairs) is inserted or deleted (Teama, 2018). Indels are highly abundant in human genomes, second only to single nucleotide polymorphisms (SNP), and makeup 15–21% of human

polymorphisms (Mullaney *et al.,* 2010). The proportion of indels is 5.4-times between coding and noncoding regions (Chen *et al.,* 2009). Over the last few years, InDel polymorphisms have stirred considerable interest within the forensic community due to their short length, making them ideal markers for forensic analysis of degraded DNA as in SNPs (Weber *et al.,* 2002; Väli, *et al.,* 2008). The short amplicon size ranges, high multiplexing capability, and low mutation rate also makes them an attractive complement to mini-STRs. InDels can also be analyzed with the same simple end-labelled PCR primer methods as STRs, thus avoiding the multi-step protocols required of SNP typing single base extension assays and providing a more direct relationship between input DNA and peak height ratios (Fondevila *et al.,* 2011). Thus, InDel genotyping ought to be viewed as a serious contender for inclusion in the list of forensic markers.

2.2.3 Polymorphic repetitive sequences

Polymorphic repeat sequences comprise of DNA sequences that have multiple tandem copies of nucleotides (O'Dushlaine, 2005) localized largely in the centromere and telomere of the chromosome (Krynetskiy, 2017). Repeat sequences with greater than 100 bp, 10-100 bp and less than 10 bp are classified as macrosatellite, minisatellite (VTNR), and microsatellites (STRs), respectively (Jeffreys, 1985; Hua-Van, 2011; Ismail & Essawi, 2012). The less than 10bp have become the marker of choice as it able to predict much variability in the genome and with little starting material.

2.2.4 Structural and copy number variations (CNVs)

Structural variations are genomic alterations (e.g., an inversion) involving DNA segments of one kilobase (kb) or larger, while copy number polymorphisms are a duplication or deletion events involving >1 kb of DNA in comparison with a reference genome (Feuk, *et al.,* 2006; Redon, *et al.,* 2006). CNVs or a balanced structural

rearrangement are intermediate-sized structural variants of approximately 8–40 kb in size (Tuzun, *et al.*, 2005). CVNs are mapped mostly in genetic disease linkages and evolutionary studies (Stankiewicz and Lupski, 2002; Cheng *et al.*, 2005; Beckmann *et al.*, 2007).

2.3 History of forensic DNA profiling

DNA profiling is a process that enables the identification of individuals based on the unique patterns of DNA extracted from biological samples such as hair, blood, and semen (Butler, 2012). Different DNA profiling techniques exist, using restriction fragment length polymorphism (RFLP) or polymerase chain reaction (PCR). The availability of these techniques spans over time, with the RFLP technique being used to profile VNTR regions and PCR initially for VNTR but later used for STR profiling, respectively (Figure 2.6).