INTRODUCTION

Understanding of the role of tandem repeats in DNA has grown significantly over the past 40 years. The discovery of DNA satellite in ***1961 by Kit***, prompted research into the properties of repetitive DNA. This eventually led to an understanding of the wide range of sizes and genomic locations of tandem repeats ***(Gelfand et al., 2006)***.

Tandemly repeated DNA is organized as multiple copies of a homologous DNA sequence of a certain size (repeat unit), that are arranged in a head to tail pattern to form tandem arrays. They are not evenly dispersed throughout the genome but they tend to cluster toward the telomeric end of chromosome. Thus they represent a distinct type of sequence organization shared by all sequenced genomes ***(Gelfand et al., 2007)***.

Tandem DNA in the human genome shows a wide range of repeat sizes and organization, ranging from microsatellites or short tandem repeat (STR) which are repeats of sequence less than 5 base pairs in length and minisatellites or variable number tandem repeats (VNTRs) or simple sequence repeat which contain longer blocks (10 to 80 bp) to megasatellites which contain up to several kb ***(Ames et al., 2008)***. STR and VNTRs are highly polymorphic with high heterozygozity rate in populations, not surprisingly, longer VNTR tend to show greater polymorphism and greater utily for genotyping ***(Warburton et al., 2008)***.

Individual repeats can be removed from or added to the VNTR via recombination or replication errors, leading to alleles with different numbers of repeats. Flanking the repeats are segments of non-repetitive sequence, allowing the VNTR blocks to be amplified by polymerase chain reaction (PCR) techniques and their size determined by gel electrophoresis ***(Morling, 2009)***.

Genetic markers are used for identity testing and paternity analysis depends on knowing the allele frequencies in the population. Although each VNTR locus is usually associated with a large number of different alleles in any population, each individual carries two alleles only—one on each of the two homologous chromosomes. Due to their high polymorphic content, VNTR constitute useful tools in population genetic studies in understanding population and ethnic migrations throughout history. They also constitute preferred systems for DNA “fingerprinting,” or determination of unique sets of genetic markers for individual identification. The direct practical applications of them includes criminal and forensic examinations, solving of immigration cases, and paternity testing, evaluating hematopoietic chimerism and in determining the origin of leukemic cells in patients with recurrent disease after BMT ***(Babushkina et al., 2011)***.

AIM OF THE WORK

The aim of this work is to standardize the PCR protocols of six VNTR loci (ApoB, 33.1 , 33.6, YNZ-22, HRAS and D1S80) in order to detect allelic poly-morphism among Egyptian population to detect the most polymorphic loci. These polymorphic loci may be used as fingerprint for genetic differentiation among Egyptian population in order to be used for follow up of chimeric status after bone marrow transplantation.

Chapter I

HUMAN GENETICS

Introduction:

"Everyone is different" Just as we all look different, our genomic sequences vary significantly. No two individuals can have an identical deoxyribonucleic acid (DNA) pattern except identical twins. Now recognizing differences in appearance between individuals is one thing; being able to genetically type them quickly. These differences in our genome are known as genetic variations and are caused by various types of mechanisms. A subset of these genetic variations is defined as genetic polymorphisms when they are observed at a frequency of >1% in a certain population ***(Nakamura, 2009).***

In ***1901***, ***Landsteiner*** discovered ABO blood grouping in humans. A set of blood group markers (red cell antigens and serum protein) were also used to identify victims and suspects in crime cases. As the polymorphism was very limited in ABO blood grouping, it was neither informative nor very suitable for excluding a person as a suspect in a criminal case. The advent of DNA-based markers has now revolutionized the field of forensic science, as it can measure an unlimited extent of polymorphisms sufficient to differentiate one individual from another ***(Varsha, 2006).***

DNA fingerprinting is a powerful technology that has revolutionized forensic science. DNA technology was first developed in England in ***1985*** by ***Jeffreys et al.*** This technology, so named because DNA is used for identification rather than latent physical fingerprints, gives a unique and specific profile similar to a thumb impression. DNA fingerprinting technology today has made it possible to identify the source of biological samples found at a crime scene and also to resolve disputes of paternity and other criminal cases ***(Richard et al., 2008)***.

DNA Structure:

The DNA is the hereditary material of living organisms. It is primarily found in the cell nucleus. DNA is organized into long threadlike structures called chromosomes. In addition to DNA, chromosomes contain associated proteins, predominantly histones, that organize the DNA into its native conformation. Human chromosomes can be classified by centromere position into three groups. If the centromere is central the chromosome is metacentric (or mediocentric). If the centromere is some what off-central the chromosome is submetacentric (submediocentric). And if the centromere is near one end the chromosome is acrocentric. A fourth type, telocentric, with the centromere at one end of the chromosome, does not occur in man. In man chromosomes 1,3,16,19 and 20 are metacentric or nearly so; the chromosomes number 13, 14,15 and 21,22,Y are acrocentric;

and the other chromosomes are submetacentric ***(Turnpenny and Ellard, 2007)*** **(Figure 1).**

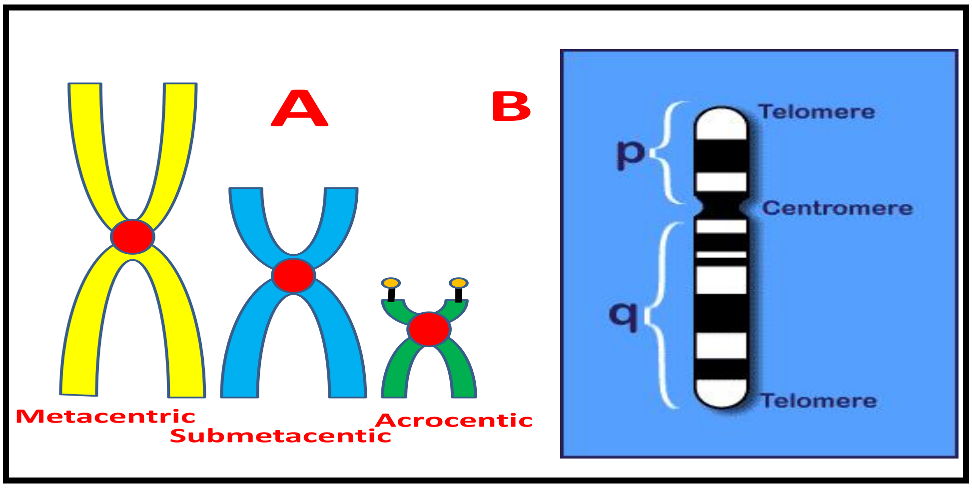

Figure (1): (A) Types of chromosomes, (B)The structure of the human chromosome ***(Turnpenny and Ellard, 2007).***

The functional DNA molecule is actually two strands of DNA intertwined, forming a double helix that is tightly packed in the cell nucleus.This structure was first proposed by ***Watson and Crick in 1953***. DNA is a polymer. Its individual units or building blocks are termed nucleotides. Each nucleotide consists of a five-carbon sugar (deoxyribose), a nitrogenous base, and a phosphate group. Only four types of nucleotides are found in DNA. The sole difference between them is which base is bound to the deoxyribose at carbon 1: adenine (A), guanine (G), cytosine (C), or thymine (T). The bases A and G, with two nitrogen rings, are called purines. Bases C and T have a single ring and are termed pyrimidines ***(Griffiths et al., 1992).*** Nucleotides are linked together by phosphodiester bonds to form a single stranded DNA. The chain has a polarity with the 5' phosphate in one end and the 3'hydroxy group at the other end. The DNA single strand can be represented in a simple way by its base sequence from 5' to 3' direction ***(Banfalvi, 2009)*** **(figure 2**).

The two DNA strands are held together by complement-tary base pairing between the purines and pyrimidines; that is, A pairs only with T, forming two hydrogen bonds, and C pairs only with G, forming three hydrogen bonds. Thus, regions of DNA rich in G-C are bonded together with more energy than regions of DNA rich in A-T pairs. Also, note that the two strands of the double helix run in opposite orientation; hence they are said to be antiparallel. One DNA backbone stretches from 5' phosphate to 3' hydroxyl, while the second runs from the 3' end to the 5' end ***(Lodish et al., 2000)***.

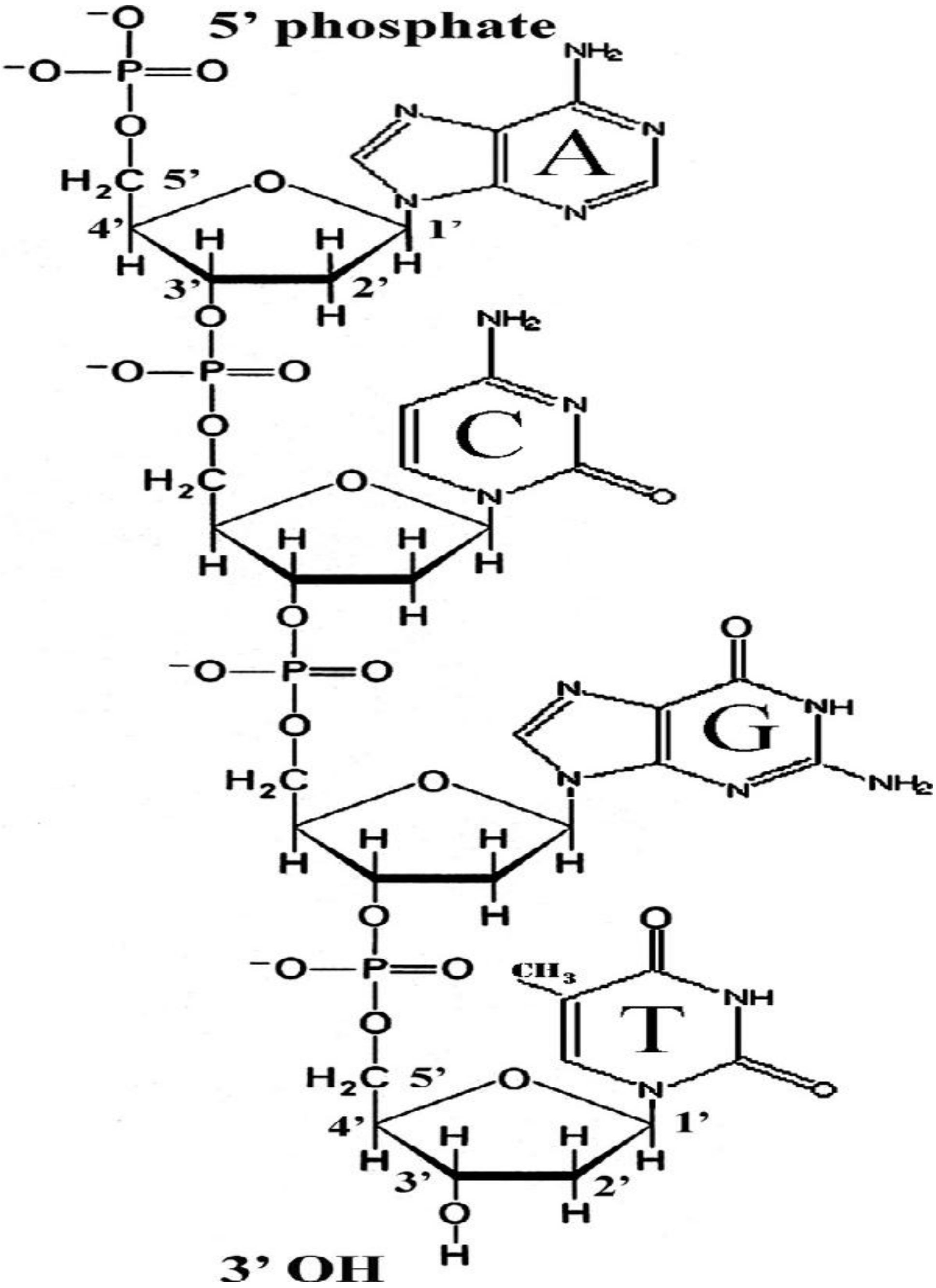

Figure (2): Primary structure of DNA ***(Banfalvi, 2009)***.

Types of DNA Sequences:

Only about 3-5% of the 3 billion basepairs that constitute the human genome codes for proteins. Scientists have begun to understand the significance of the sequences that comprise the remaining 95-97%. One interesting aspect of these noncoding sequences is the amount of interindividual variation that these sequences exhibit ***(Weidenhammer and Gregory, 2006).*** Remarkably, ~50% of the genome is composed of repetitive elements, and another ~45% is not conserved and does not belong to the functional categories ***(Antonarakis, 2010).*** The types of DNA sequence are summarized in the following **Table (1)**.

Table (1): Types of DNA sequence

A-Nuclear:

I-Genes and Gene Related Sequences:
1- Unique single copy genes:
2- Multigene families:
  (a) Classical gene families. (b) Gene superfamilies
3- Pseudogenes

II- Extragenic DNA(unique/ low copy number or moderate / highly repetitive):

1- **Dispersed DNA Repeats :**
a-Retrotransposons (class I elements)
  1-Long interspersed nuclear elements(LINEs)
  2- Short interspersed nuclear elements (SINEs)
  3- Retrovirus like (LTR transposons)
b- DNA transposons (class II elements)

2-Tandem repeats
a-Satellite
b-Minisatellite
  (i)Telomeric
  (ii) hypervariable
c-Microsatellite

B-Mitochondrial
I-Two rRNAgenes (ribosomal RNA)
II-022 tRNA (transfer RNAs)

(Turnpenny and Ellard, 2007)

DNA sequences are classified to nuclear and mitochondrial. Nuclear contain two major group: the first group contains genes and gene related sequences and extragenic DNA. Genes and gene related sequences contain unique single copy genes, multigene families and pseudogenes. Extragenic DNA contains dispersed DNA repeats and tandem repeats. Dispersed DNA repeats are two classes. The first class contains retrotransposons and the second class contains DNA transposons. Retrotransposons contains long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs) and retrovirus-like (LTR transposons). Tandem repeats contain satellite, minisatellite and microsatellite ***(Turnpenny and Ellard, 2007).***

In addition to nuclear DNA, the several thousands of mitochondria of each cell posses their own circular double strand DNA, mitochondrial DNA (mtDNA). Mitochondria have an extranuclear DNA genome, the sequence of which was first reported for humans in 1981. The human mtDNA is circular with 16,569 bp (as opposed to the linear ~3 billion bp in the nuclear DNA), and it exists in hundreds to thousands of copies in a single cell. MtDNA containing little repetitive DNA, and code for 37 genes, which include two types of ribosomal ribonucleic acid (rRNA), 22 transfer RNAs (tRNA) and 13 protein subunits for enzymes, such as cytochrome b and cytochrome oxidase, which are involved in the energy producing oxidative phosphorylation pathways. The likelihood of recovering mtDNA from very minute and degraded biological samples is greater than for nuclear DNA ***(Antonarakis, 2010)***.

In situations where nuclear (chromosomal) DNA typing is not an option (for example, insufficient quantities or too degraded), or an attempted typing using nuclear DNA markers is unsuccessful, mitochrondrial DNA typing can be used. MtDNA has been extracted from teeth, hair shafts, bone fragments, all of which fail to yield forensic results with nuclear DNA markers. The most important is the maternal pattern of inheritance of mtDNA as mitochondria of the fertilized zygote are inherited almost exclusively from the oocyte, leading to maternal pattern of inheritance that characterized many mitochondrial disorders ***(Weidenhammer and Gregory, 2006).***

I-Genes and Gene Related Sequences:

The gene is defined as a segment of DNA located on a chromosome, representing the fundamental physical and functional unit of inheritance, directing the synthesis of a protein, or associated with regulating function ***(Banfalvi, 2009).*** There are non-coding intervening sequences or introns, which separate the coding sequences or exons **(Figure 3)**. The number and size of introns in various genes in humans are extremely variable. Although there is a general trend that the larger the gene, the greater the number and size of the exons, however individual introns can be far larger than the coding sequences. Genes in the humans do not usually overlap being separated from each other by an average of 30kb ***(Turnpenny and Ellard, 2007).***

In overlapping genes a single stranded portion of DNA is coding for two separate proteins. Overlapping genes may evolve as a result of the extension of an open reading frame caused by a switch to an upstream initiation codon, substitutions in initiation or termination codons, and deletions and frame shifts that eliminate initiation or termination codons. Overlapping allows to produce more protein from a certain DNA sequence than if the genes were arranged linearly after each other. Such efficient packaging is unavoidable in circular single stranded viruses which would not have enough DNA to encode all their proteins if transcription took place linearly gene after gene (e.g. bacteriphage X174). Overlapping genes were found not only in viruses but also in prokaryotic and eukaryotic cells. For instance in the human mitochondrial DNA, there are two overlapping genes. Some algae with a small genome content (Guillardia, 5 × 105 base pairs) code for nearly 500 proteins, 44of them produced from overlapping genes ***(Banfalvi, 2009).***

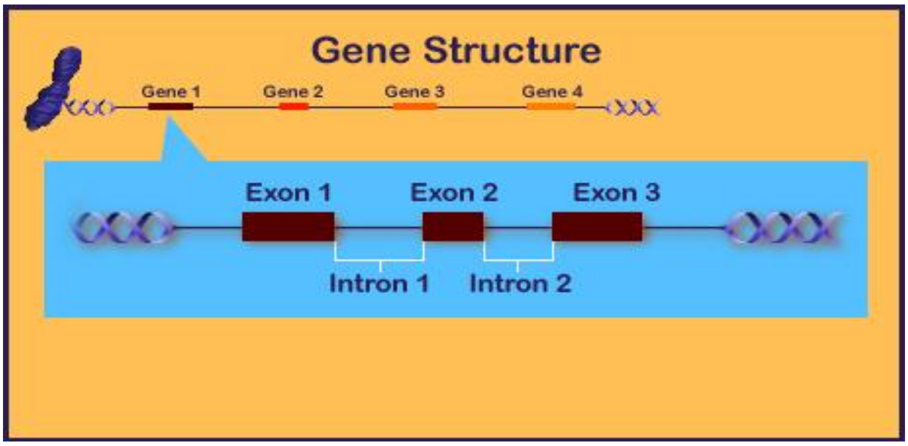

Figure (3): Structure of human gene ***(Turnpenny and Ellard, 2007).***

By **2003** the Human Genome Project came to an end, with a fairly complete listing of human genes which turned out to be less than 25,000, however by **2009** the human genome was expected to encode more than 100,000 genes ***(Banfalvi, 2009).*** Only about 3-5% of the human genome encodes expressed protein sequences. The vast majority of DNA about 95-97% in human DNA does not give rise to any proteins. The distribution of genes varies greatly between different chromosomes in certain parts of the chromosomes as in centromeric regions which contain few with the majority being located in subtelomeric regions ***(Weidenhammer and Gregory, 2006).***

Genes and gene related sequences are classified to unique single copy gene, multigene families and pseudogenes. The multigene families are further classified to classical gene families and gene superfamilies ***(Holsinger, 2008).***

1- Unique single copy genes:

The unique single copy genes code for polypeptides which are involved in or carry out a variety of cellular functions. These include enzymes, hormones, receptors and structural and regulatory proteins ***(Turnpenny and Ellard, 2007 & Feau et al., 2011).***

2- Multigene families:

DNA sequences in the nuclear diploid genome usually exist as two allelic copies (on paternal and maternal homologous chromosomes). In addition to this degree of repetition, about 40% of the human nuclear genome in both haploid and diploid cells is composed of sets of closely related DNA sequences. Those genes have similar functions, having arisen through gene duplication events with subsequent evolutionary divergence making up what is known as multigene families. Some are found physically close together in clusters, e.g. the α- and B-globin gene clusters on chromosomes 16 and 11 respectively, while others are widely dispersed throughout the genome occurring on different chromosomes. Multigene families can be split into two types, classical gene families and gene superfamilies ***(Strachan and Read, 1999 & Turnpenny and Ellard, 2007).***

(a) Classical gene families:

Classical gene families are distinguished by members which exhibit a high degree of sequence homology over most of the gene length or, at least, the coding DNA component, a feature which automatically identifies such sequences as being closely related evolutionarily as well as functionally. Examples of classic gene families include numerous copies of genes coding for various rRNAs gene families, which are clustered as tandem arrays, and the different tRNA gene families, which are dispersed in numerous clusters throughout the human genome. There is an extremely high degree of sequence similarity between family members ***(Dorus et al., 2003 & Turnpenny and Ellard, 2007).***