University of Glasgow
College of Medical, Veterinary and Life Sciences
Institute of Molecular, Cell and Systems Biology

Age-dependent microsatellite somatic mosaicism in humans

Saeed G. Alghamdi

A thesis presented in fulfillment of the requirements for the degree of Doctor of Philosophy

February 2014

Dedication

To My Wife and Our Progeny

Abstract

Several inherited diseases such as Huntington disease and myotonic dystrophy type 1 are associated with the expansion of repeats. A high level of age-dependent instability has been observed in the expanded alleles present in the germline and soma. The present study covers the investigation of other non-disease-associated expanded microsatellites in order to explore their variability, which in turn could be used to predict age in the general population.

First, using Tandem Repeats Finder, 23 pure loci with alleles >50 repeats in at least one of two reference genomes were identified. Several loci among those show high levels of variation in the general population with a considerable proportion of large expanded alleles. SP-PCR analysis revealed that these loci showed a relatively low level of somatic instability and mostly only small length changes.

The SP-PCR approach is laborious and time consuming; these disadvantages could potentially be overcome by the use of new technologies based on high throughput analysis. In this study a target enrichment sequencing approach using custom bait and Illumina Paired-End Sequencing using Illumina GA IIx platform to capture and sequence sequences in question was used. 25,539 pure (100% match) mono, di, and trinucleotides with unique flanking sequences and copy number ≥5 were investigated. Similarly, telomeric regions, DNA repair genes, the *DMPK* region and SNPs were captured using custom designed baits. The experiment generated millions of reads for each sample, ranging from 5.5 to 9.5 million reads. Subsequently, CLC Genomics Workbench and Bowtie 2 were used to map reads to the HG19 reference genome and to count the number of reads aligned to each DNA sequence. The effect of age at sampling on the generated reads was also investigated. The data showed a significant correlation between unique sequence reads and age at sampling.

Microsatellites, including mono, di and tri nucleotides, were successfully captured using custom baits based on unique flanking sequences. The microsatellite total reads were found to be correlated with age at sampling. Genotyping was found to be successful for di, tri and tetra nucleotide loci however PCR slippage was commonly found in the dinucleotide repeats. By contrast due to the extensive PCR slippage observed in mononucleotides, reliable genotyping was difficult to achieve.

Unlike microsatellites, no significant correlation was observed between DNA sequences of repair genes or the autosome gene read counts with age at sampling. Thereby, this feature could be exploited to normalise the read counts such as of telomeres.

Thousands of human SNPs were successfully captured using baits having a single mismatch. Baits were designed to ensure an equal capture for both alleles that could exist at a single SNP. The overall coverage of SNPs was noticeably good and used to generate reliable genotype.

Age-dependent telomere shortening was estimated. The relative telomere length showed a significant correlation with age at sampling. The data show four major variants of telomeric repeats, TTAGGG, TTGGGG, TGAGGG and TAAGGG with a shortening rate of 60 bp/year.

In summary, the primary aim of this project was not achieved due to the low level somatic instability of microsatellite and low-resolution power of traditional SP-PCR. However, the success with the NGS observed in this study, indicate a reasonable potential of using microsatellites to estimate ageing in humans.

Acknowledgement

It is a huge undertaking for me to transform myself from a police officer to a PhD student of genetics. However, that transition could only become possible because of the support and encouragement I relished during the course of my PhD studies. First, I would like to thank my supervisor, Professor Darren Monckton, for helping me all the way through from the designing the experiments to the critical analysis of the results. I am also indebted to Prof. Darren for his much needed feedback to improve the write up of this thesis and sieving out the scientific errors helping me all the way through. I am also grateful to Dr. Pawel Herzyk, Julie and for their support in the next generation sequencing experiment. All colleagues of Lab 412, especially Dr Anneli Cooper, Dr Sarah Cumming, Dr Catherine Higham, Berit Adam, Eloise Larson and Josephine deserve a resounding thanks for providing a healthy and friendly environment, which facilitated the progress of my work. I am also grateful to Dr. Graham Hamilton and two of MRes students Carlos Del Ojo Elias and John Cole for their support in bioinformatics analyses.

My deepest gratitude is due to my family for their patience and support and enduring my absence from home affairs, during the course of this study. Without their always available moral support, it is difficult for me to imagine such endeavour.

Declaration

The research presented in this thesis is my own original work, unless otherwise stated, and has not been submitted for any other degree

Saeed G. Alghamdi

Table of contents

List of figures

List of tables

List of abbreviations

AT	Annealing temperature
bp	Base pair
CODIS	Combined DNA index system
DM1	Myotonic dystrophy type 1
E. coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
EtBr	Ethidium bromide
FDP	Forensic DNA phenotyping
L	Litre
M	Molar
m	Milli (10^{-3})
Mb	Megabase
min	Minutes
MMR	Mismatch repair
MW	Molecular weight
n	Nano (10^{-9})
OXPHOS	Oxidative phosphorylation
PCR	Polymerase chain reaction
ROS	Reactive oxygen species
SDS	Sodium dodecyl sulphate
sec	Seconds
SNP	Single nucleotide polymorphism
SP-PCR	Small pool PCR
STR	Short tandem repeat
TRDB	Tandem repeat database
TRF	Tandem Repeat Finder
Tris	Tris (hydroxymethyl) amino methane
UTR	Untranslated region
UV	Ultraviolet
VNTRs	Variable number tandem repeats
µ	Micro (10^{-6})
°C	Degrees Celsius

1 Introduction

Ageing is a process in which an organism's condition declines, the fertility rate decreases and the risk of death increases (Kirkwood, 2005). The ability to measure this ageing process and accurately determine the biological age of an individual has important practical applications in the field of forensic science (Alkass *et al.*, 2010). Establishing the approximate age of an unidentified person of interest from available biological material may provide important leads in criminal investigations, disaster victim identification, or missing person cases (Alkass *et al.*, 2010).

Current age estimation for both living and dead subjects is dependent on several anatomical and morphological features such as odontology or skeletal analysis (Alkass *et al.*, 2010). However, using anatomical and morphological methods are expensive, laborious, and require expert examiners (Alkass *et al.*, 2010). Most importantly, the sample material requirements, in terms of quality, quantity and tissue type for these analyses, often precludes the suitability of such techniques for the limited biological material recovered from most crime scenes. The ability to measure human ageing using molecular techniques from samples that are devoid of anatomical and morphological information, such as bloodstains, thus has the potential to be enormously valuable (Alvarez *et al.,* 2006).

In most crimes, a biological sample can be collected from the crime scene from which genetic material (DNA) can be extracted for analysis. DNA carries genetic information in humans, which with the exception of identical twins, is unique for each individual using standard forensic DNA profiling. Recently, using ultra-deep next generation sequencing to identify very rare mutations that arise early after the human blastocyst has divided into two, the origin of twins, can successfully distinguished monozygotic twins as such mutations will be carried on into somatic tissue and the germline (Weber-Lehmann *et al*., 2014).

If the DNA profile can be determined, this profile can then be compared to that of a direct suspect, or in the absence of a suspect it can be checked for a match against a national DNA forensic database. If no match is identified, the profile can be stored in a forensic DNA database indefinitely for future comparisons to any DNA profiles added to the database over time (Corte-Real, 2004).

Since its introduction the genetic profiling of biological samples has evolved to become one of the most powerful tools in forensic science. However, standard DNA profiles provide no phenotypic information beyond the sex of the individual.

The development of new techniques to accurately predict the biological age and other phenotypic characteristics from DNA, could yield new tools to aid crime scene investigators develop a picture of unknown DNA profiles at crime scenes and narrow the pool of potential suspects associated with a criminal investigation.

1.1 Ageing theories

Individuals of the species *Homo sapiens* have a long lifespan, as with most other mammalian species, and take a long time to reach sexual maturity (Kanungo, 2005). The reproductive stage occupies a significant period of the life span, and is followed by a post-reproductive stage in both sex. In most cases, the rate of reproduction is inversely proportional to the productive period and longevity. Ageing is a complicated process and encompasses a variety of mechanisms, of varying importance throughout the species lifespan. In order to use DNA to estimate the age of an individual, the DNA changes associated with the ageing process need to be thoroughly studied and understood. Ageing also affects organ function as human organ function relies on how their cells function. Ageing cells may have impaired function. When ageing cells die and are not replaced, the cell number is reduced. Low cell numbers affect organ function. Bones and joints become less dense and become weaker and more prone to break. Their calcium content is reduced as the amount of absorbed calcium from nutrients decreases (Beers *et al*., 2004).

Muscle mass and muscle strength also tend to decrease. Muscle mass reduction is due in part to the fact that muscles are physically used less and begins to reduce in size (Beers *et al.*, 2004). The levels of growth hormone and testosterone also decrease, which reduces muscle development. Ageing skin becomes thinner, less elastic, drier, and finely wrinkled. The production of collagen, which provides strength to skin, and elastin which provides flexibility to skin, is decreased in ageing skin. The fatty layer under the skin becomes thinner. It protects and supports skin, and helps conserve body heat. A thin fat layer reduces tolerance for cold and promotes wrinkling (Beers *et. al*., 2004). In the elderly, heart and blood vessels become stiffer. The heart and arteries are less elastic and fail to expand to accommodate the blood. Consequently, risk of high blood pressure tends to increase (Beers *et. al*., 2004).

Several overlapping theories, each supported by scientific evidence, have attempted to explain age-dependent DNA changes (Brierley *et al.,* 1997). These DNA-age-dependent changes could have a practical application to forensic science, permitting age determination from the samples collected from crime scenes. Some of the theories that attempt to describe ageing at the genetic level are discussed in the following sections.

1.1.1 Somatic mutation theory

Somatic mutation theory is the first theory that deals with the cause of ageing at the genomic level. Whole body irradiation of rats leading to the early onset of death compared to controls was demonstrated over 50 years ago (Kanungo, 2005). Later, considerable similarities have been shown in the process of ageing and death, between the irradiated rodents and humans to those of normal individuals, suggesting that radiation accelerates the process of ageing. However, an increased incidence of neoplasia was noted in the former (Sacher, 1956; Henshaw, 1957 and Warren, 1956). Based on these observations, the somatic mutation theory of ageing was proposed by Szilard (1959). The theory postulates that over time, random mutations tend to accumulate in the somatic cells and this increasing mutation load in the cell, decreases the production of functional proteins. Consequently, this leads to cell death and the gradual decline of functional somatic cells and overall functionality of the organism.

Sporadic cancers represent 95% of the total burden of cancers. In the last 50 years, the somatic mutation theory of oncogenesis has been developed to explain the relationship between mutation and cancer. According to this theory, over time the accumulation of mutations in a single cell results in aberrations in controlled proliferation (cancer), thereby suggesting that cancers are derived from a single cell, and are hence monoclonal. The theory is well substantiated by the discovery of the oncogenes that regulate proliferation and mutation in those leads to the dysregulation of cell division and cancers (Soto and Sonnenschein, 2004).

Several proteins encoded by the human genome have anti-cancerous roles (Sherr, 2004). The tumour suppressor genes act by repressing the genes that activate the cell cycle, encoding DNA damage repair proteins and cell adhesion proteins. Mechanistically, the tumour suppressor genes manifest their effects both by single- and double-hit mechanism (Knudson, 1971). Amongst several tumour suppressor genes, p53 is most well studied tumour suppressor gene that induces apoptosis thus preventing the proliferation of aberrant cell growth (Baker *et al*., 1990). Homozygous loss of p53 was observed in 65% of colon

cancers (Kapiteijn *et al*., 2001), 30-50% of breast cancers (Megha *et al*., 2002) and 50% of lung cancers (Olivier *et al*., 2010).

1.1.2 Error theory

In 1963, Orgel suggested that errors in information transfer steps, such as transcription and translation may result in the accumulation of defective proteins, resulting in ageing. For instance, the insertion of nucleotides in the messenger RNA during transcription may result in a frame shift in the permutation of the amino acids. Such a change may lead to a structurally and/or functionally impaired protein.

In 1970, Orgel modified his theory by adding that errors may not always accumulate, since successive generation of protein synthesis machinery is distinct. Error theory and somatic mutation are similar and difficult to distinguish (Holliday and Tarrant, 1972). Kanungo and Gandhi (1972) found that the activity of malate dehydrogenase is the same in both young and old rats and displayed no age-related difference.

1.1.3 Dysdifferentiation hypothesis

According to Cutler, cells' gradual drifting from their proper differentiation state with time results in ageing (Cutler, 1984; Dean *et al.,* 1985; Kator *et al.,* 1985). Controlled differentiation leads to creation of the organism. Proper differentiation is maintained to the extent in time that is necessary to ensure the evolutionary success of the organism; then ageing begins to set in as a result of the slow random loss of this differentiation. At the molecular level, several regulatory mechanisms such as methylation of genomic DNA (at the promoter region) and posttranslational modifications of the associated proteins such as histones are involved in epigenetic gene regulation (Jaenisch and Bird, 2003). These modifications alter the compactness of DNA in the nucleosome, thereby allowing and/or inhibiting the transcriptional machinery to function by preventing their association with DNA. This in turn results in the differentiation of tissue in terms of structure and function (Spivakov and Fisher, 2007). During aging, the loss of efficiency of these epigenetic factors results in the loss of differentiation and consequently their function (Teschendorff *et al.*, 2013).

1.1.4 The disposable soma model

Ageing is a biological process but not a disease, and may be explained by the disposable soma theory. Kirkwood asked how best an organism should distribute its metabolic resources, primarily energy, to ensure its survival from one day to the next (maintenance), and breeding to produce offspring to secure the existence of its genes when the organism dies (reproduction). Reproduction is essential for survival of the species while maintenance is needed to reproduce. However, all species are not immune to the hazards of the environment, such as predators, starvation and disease (Kirkwood, 1993). These hazards would limit the average survival time, even if ageing did not occur. Maintenance is just needed to ensure that the body (soma) remains in steady condition long enough to reproduce until death occurs, in most cases from accidental causes. Larger investment in maintenance is not always a good way to ensure survival (Kirkwood, 1993). It has a disadvantage because maintenance reduces resources that are essential and better used for reproduction. The theory Kirkwood conceived determined that an individual should invest no more resources in the maintenance of somatic tissues than are necessary for the soma to survive to reproduction (Kirkwood, 1993).

Longevity in mammals results from the relationship between the rates of growth and reproduction and an increase in the accuracy of synthesis of macromolecules (Aubert and Lansdorp, 2008). Natural selection will favour longevity mechanisms that supply the best cooperation between investments in somatic maintenance and in reproduction. Consequently, organisms tend to limit investments in the maintenance and repair of somatic cells and tissues. Maintenance systems are costly, but an important part of the metabolic process. Therefore, it is concluded that intrinsic ageing of somatic cells is caused by accumulation of unrepaired cellular damage that may result in weakness, disease, and death. This idea leads to a number of other testable hypotheses (Kirkwood *et al.,* 2000). First, ageing is not programmed, as there are no active mechanisms that can cause death. As an alternative, existence is dependent upon genes that control the levels of somatic maintenance and repair ”longevity assurance”. Secondly, extrinsic hazards such as predators and disease determine longevity of different species by optimizing the level of somatic maintenance according to the degree of hazard the species is susceptible to. The species can adapt to the environment surrounding it. As hazards reduce, there may be a selective advantage to increase maintenance. This can explain the greater longevity of animals that can fly such as birds and bats when compared with those that must stay on the ground. Thirdly, many maintenance mechanisms contribute toward ageing and longevity,

and are thus likely to be controlled by multiple genetic processes. The rate at which damage accumulates or is repaired affects the levels at which individual genes are set. Finally, the ageing process is subject to both intrinsic and extrinsic random factors (Kirkwood *et al.,* 2000). The role of chance in ageing is proved clearly by the difference in lifespan for genetically identical populations such as inbred mice when maintained in homogeneous environments (Kirkwood *et al.*, 2000). The relative investment is keeping the germline in good shape. Such errors do not accumulate in germ cells. In addition, a"quality mechanism" ensures that ova and embryos minimise errors during oogenesis and early development.

In summary, the disposable soma theory stipulated that organisms should optimize the share of metabolic resources between somatic maintenance, growth and reproduction in such a way that the resources directed to maintenance are adequate to keep the soma in good condition throughout the natural lifespan in the wild environment, but less than would be required for indefinite somatic survival. The disposable soma theory thus fills the gap between mechanistic and evolutionary theories of ageing by proposing that ageing results from the progressive accumulation of molecular and cellular damage, as a direct consequence of evolved limitations in the genetic settings of maintenance and repair functions (Drenos and Kirkwood, 2005).

1.1.5 Mitochondrial mutations accumulate with age

A typical somatic human cell contains 100-10,000 copies of mitochondrial DNA of average 16.5 kbp size (Figure 1.1)(Scheffler, 2007; John *et al.*, 2010). Interestingly these numbers remains consistent with age as well (Miller *et al.,* 2003). The mutation rate in the mitochondrial DNA (mtDNA) is 10 times higher than that in nuclear DNA (Alexeyev, 2009). The mitochondria tend to accumulate mutations throughout the lifetime, which possibly leads to the impairment of oxidative phosphorylation and reduced energy production, resulting in disease (Sukernik *et al.,* 2002). This increased rate of mutation is mainly a function of the lack of efficient DNA repair machinery. These diseases are likely to be age related and tissue specific, as the accumulation of mutations increases with age and thus tissues with higher metabolic rate tend to be the most affected. For example, a deletion of 7,436 bp was identified in myocardium mtDNA in all subjects of a study aged over 70, with the percentage of mitochondria affected increasing with age (Hattori *et al.,* 1991). Moreover, in a more recent study, accumulation of mitochondrial mutation has been found associated with muscle fibre loss (Herbst *et al.*, 2007).

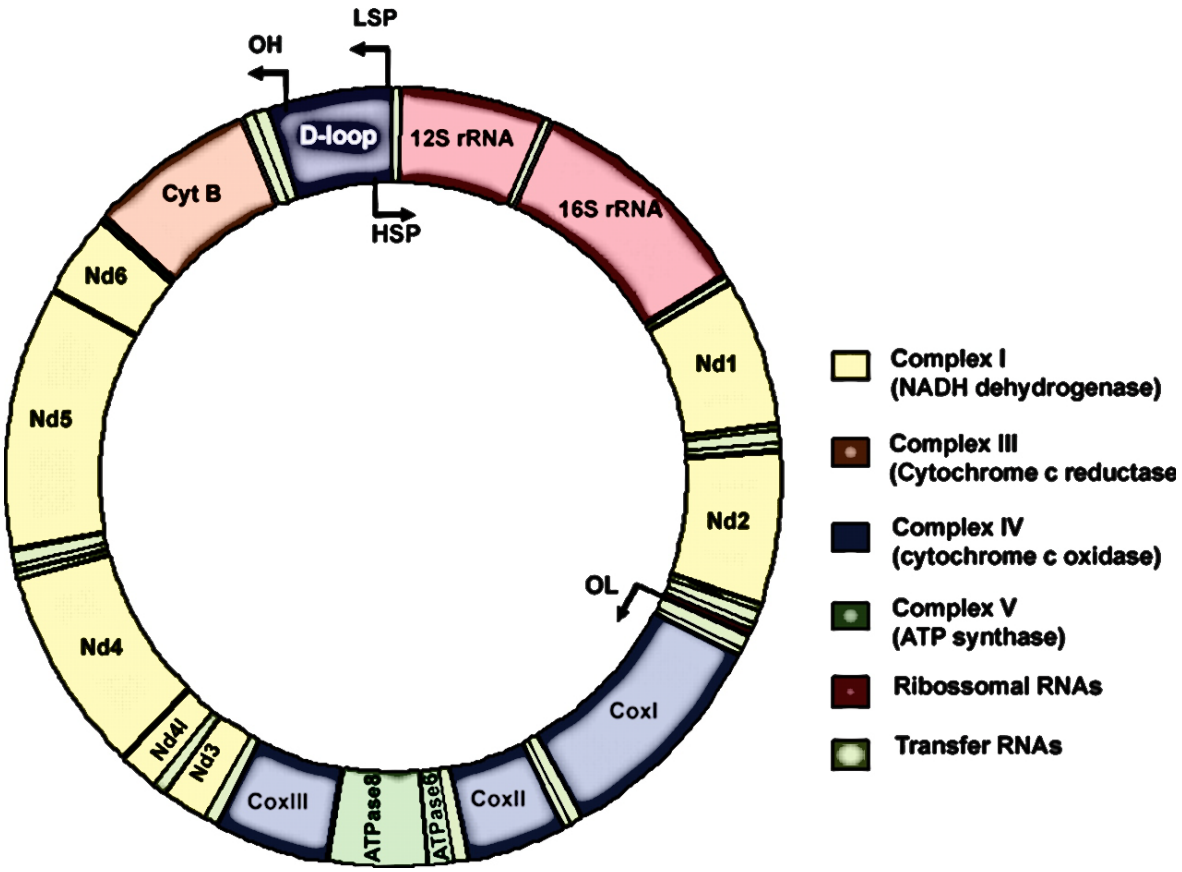


Figure 1.1:The mtDNA structure (adapted from John *et al.*, 2010).

Mitochondria are subcellular organelles situated in the cytoplasm of eukaryotic cells. They have a range of functions such as energy generation via the citric acid cycle, β-oxidation, the urea cycle, and calcium storage and haem synthesis. However, among these the most important function is the generation of ATP. ATP is generated by oxidative phosphorylation process (OXPHOS) (Greaves and Turnbull, 2009). Mitochondria are unique; they contain the only extra-nuclear source of DNA in animals. MtDNA is a circular, double-stranded DNA molecule, in which the genes, unlike those of the nuclear genome, contain no introns. A few non-coding bases may be located between mtDNA genes which in most cases lack the termination codons (Greaves and Turnbull, 2009). Energy demand is proportional with the number of mtDNA copies present in different cell types (Greaves and Turnbull, 2009). The multi-copy character of mtDNA means that any new mutations occurring in the mitochondrial genome cannot cause a biochemical phenotype until they reach a minimal threshold level (Greaves and Turnbull, 2009). This is because the mutant mtDNA co-exists with the wild-type DNA in a condition known as heteroplasmy (Larsson and Clayton, 1995). Typically, 50–60% of mtDNA needs to be mutant for a defect to be observed (Shoubridge, 1994). MtDNA mutations are caused due to the leakage of reactive oxygen species (ROS) during the OXPHOS process since the

oxidative phosphorylation system is located close by on the inner mitochondrial membrane making mtDNA susceptible to damage, lack of histones and mtDNA exist as single-stranded most of the time because of its mode of replication. These make the mtDNA mutation frequency much higher than that of the nuclear DNA. The mitochondrial theory of ageing was proposed subsequent to a study that proved the majority of ROS present in the cell are produced from the mitochondria. Mitochondria have a critical role in cellular metabolism, but are prone to accumulate high levels of DNA damage. It has been suggested that the accumulation of mtDNA damage may contribute to ageing (Greaves and Turnbull, 2009). In a cycle of positive feedback, mutations result in damaging cellular OXPHOS, which increases ROS production and results in further damage to mtDNA. This process will finally lead to cell death. In addition to mtDNA mutations resulting from oxidative damage, mtDNA mutation can also occur throughout life due to errors of the mtDNA polymerase (Greaves and Turnbull, 2009).

Mutations that accumulate or occur during ageing could damage one or more of the respiratory chain components. However, as mentioned above, multiple copies of the mitochondrial genome within a single cell prevent the expression of the functional defect unless the mutation expands to reach the threshold that leads to biochemical deficiency (Greaves and Turnbull, 2009).

In the ageing process, accumulation of mtDNA mutations leads to a gradual decline in mitochondrial function within somatic cells. Some ageing-associated studies on mtDNA mutations have revealed that not only the generation of ATP decreases but also there is increased production of ROS such as superoxide anions and hydrogen peroxide in the mitochondria of ageing tissues. In human somatic cells, more than 90% of the oxygen is consumed by mitochondria and about 1–2% of the O_2 is transformed to superoxide anions through the electron transport chain (Ma *et al.,* 2009). In tissues and cultured cells from elderly subjects, an increase of oxidative stress can alter the expression and activities of antioxidant enzymes and increase oxidative damage to DNA, RNA, proteins, and lipids. At low concentrations, ROS can serve as a signalling molecule in the regulation of cell proliferation and other cellular functions. Antioxidant enzymes provide a defence system used by the cell to keep the ROS at a level that prevents oxidative stress, which is elicited by aerobic metabolism. This defence mechanism can dispose of both endogenous and exogenous ROS to minimize their damaging effects. Briefly, when ageing alters this defence system, the decline in mitochondrial function and increase in ROS production have been suggested to be the major contributory factors in ageing (Ma *et al*, 2009).

Moreover, failures of the mtDNA repair mechanisms may contribute to the accumulation of both oxidative damage and mutation of mtDNA during ageing.

Point mutations, deletions and duplication of mtDNA are all found to accumulate in a variety of tissues during ageing. Most of these mtDNA mutations arise in the mid-thirties and accumulate with age in post-mitotic tissues of humans (Michikawa *et al.*, 1999). Although mtDNA damage clearly accumulates, some have challenged the direct relevance to ageing, since the proportion of mutant mtDNA is predicted to be too low to induce an important impact on the function of mitochondria in aged tissues. The observed or detectable mutations may be just a part of the ageing-associated alterations of mtDNA (Ma *et al.,* 2009). However, most previous studies used whole tissues to screen for mtDNA mutations rather than using individual cells. Due to the uneven distribution of the mtDNA molecules and clonality in certain cells, the mosaic pattern of respiratory chain deficiency during ageing may appears. Furthermore, point mutations are known to be accumulated in the D-loop region of mtDNA with ageing in human tissues and cultured skin fibroblasts (Ma *et al.,* 2009). The D-loop accumulates more mutations in comparison to the rest of the mtDNA, which may be due to the relaxed selective pressure due to the absence of critical genes in the region (Larizza *et al.*, 2002).

In addition to cell death, it has been recognised that mutation accumulation in mtDNA sub-threshold stress can lead to cell growth retardation and activate cellular senescence in dividing cell lines such as skin fibroblasts, melanocytes and endothelial cells, subsequent to exposure to oxidative agents such as H_2O_2 (Dumont *et al.*, 2000; Toussaint *et al.,* 2002). More than one hundred distinct mtDNA mutations have been found in ageing human tissues (Ma *et al.*, 2009). Some individuals may be more resistant to oxidative stress than others, and thereby able to slow down the ageing process (Beckman and Ames, 1998).

Studying how mutations accumulate in mitochondria should be based on analysis of individual cells rather than homogenised tissue. Khrapko *et al.* (1999) carried out such a study, on human cardiac tissue and Kopsidas *et al.* (1998) for skeletal muscle. Both have detected much higher levels of mutant mtDNA copies, which could reflect tissue differences. Subsequently, estimation of age cannot be applied to subjects of an advanced age and who have heart disease or other disorders (Meissner *et al.*, 1999).

Mitochondrial DNA varies considerably between cell types and with ageing with reference to the number of mutations and copies (Kajander *et al*, 2002). This suggests that with the advent of proper assays such attributes of the mitochondria could be used for age

estimation. Earlier studies (Meissner *et al.*, 1999; vonWurmb-Schwark *et al.*, 2002), used the common deletion in mtDNA extracted from skeletal muscle cells to estimate the age of samples. However, no statistical significance was observed due to large variation between similarly aged individuals. However, mitochondrial DNA testing is laborious, time consuming, expensive, and is highly sensitive to substrate and laboratory contamination (Friedman, 1999). Moreover, maternal inheritance (Giles *et al.*, 1980) and high probability of heteroplasmy (Lodish *et al.*, 2000) may question the reliability of the usage of mtDNA for the age estimation. For instance prior knowledge about mutations and/or heteroplasmy is a pre-requisite in order to design baits/primers for empirical studies.

1.1.6 Age-dependent telomere shortening

In humans, the terminal end of the chromosomes are referred to as telomeres which are known to contain DNA repeats (TTAGGG). The word telomere is derived from the Greek word *telos* and *meros* which respectively mean end part. Without telomeres chromosomes would fuse and frequently break during mitosis. Telomeres include unique DNA sequences that are essential to secure chromosome ends, to prevent chromosome fusing during mitosis, to offer chromosome stability, and to ensure accurate segregation of genetic material during cell division (Aubert and Lansdorp, 2008).

In humans, the telomere sequence is 5′-TTAGGG-3′, although variant forms such as TTGGGG and TGAGGG exist sub terminally (Lindsey *et al.,* 1991). The telomere lengths vary from 3 to 20 kb (Harley *et al.,* 1990). Among different species, copies of this basic repeat unit in telomeres are variable, from chromosome to chromosome within a species and even on the same chromosome at different stages of the life cycle (Lewin, 2004). Previously, it was thought that the telomere region was more prone to damage as no nucleosomes were present within the telomeric region (Rattner, 1995). However, recent studies showed about 80% of telomeric DNA is organised in tightly packed nucleosomes separated by 10-20 bp of linker DNA in higher eukaryotes (Pisano *et al.,* 2008).

Human somatic cells have 92 telomeres (Lewis, 1998). Subtelomeric/telomere associated repeats are a complex set of repeats located just internal to the telomeric repeats. Their sequences are not conserved in eukaryotes and their function is unknown (Strachan *et al.,* 1999). No protein-encoding genes are found in telomere sequences (Bekaert *et al.,* 2004).

During DNA replication, the ends of chromosomes possess unique problems because DNA polymerases can only elongate from a free 3′ hydroxyl group. A back stitching mechanism used by the replication machinery builds the lagging strand. Along the lagging strand

template, RNA primers provide the 3′ hydroxyl group at regular intervals. The leading strand elongates in the 5′-3′direction continuously all the way to the end of the template. Even if a final RNA primer was built at the very end of the chromosome, the lagging strand stops short of the end and still would not be complete. As the final RNA primer will provide the free hydroxyl group to synthesize the DNA, but the RNA primer later needs to be removed. The chromosome would progressively shorten during each replication cycle because of this inability to replicate the ends.

Telomere length declines with each cell division, providing a marker for cellular ageing. Telomere lengths are found to be variable within an individual and may be associated with the variability in reproductive ageing. This length variability may be due to differences in telomere length at formation, activity of telomerase during early development, and the cell division rate and telomere loss with each cell division. Some studies postulate X-linked inheritance of telomere lengths (Nawrot *et al.* 2004), while others suggested paternal inheritance of telomere lengths (Njajou *et al.* 2007; Nordfjäll *et al.* 2005). Many control populations showed high variability in telomere length at any given age and the rate of telomere length decline with age (Hanna *et al.*, 2009). Due to the inter-individual differences in telomere length at birth, telomere length is highly variable (Okuda *et al.*, 2002) and suggesting the rate of telomere shortening thereafter. There is also evidence that telomere shortening rate is faster during early childhood than in later life (Zeichner *et al.*, 1999). So, an individual with faster age-dependent telomere shortening may not always have shorter telomeres. Unlike somatic telomeres, sperm telomeres are longer and not shortened with age (Harley *et al.*, 1990; Harley, 1991; Ozturk *et al.*, 2014). In contrast to the somatic cells, in the germ line, telomerase remains active which in turn compensates for the shortening of telomeres due to the end replication problem (Ozturk *et al.*, 2014). This results in the increase in the length of telomeres in the germ cells. See further details in Chapter 5.

1.1.7 Insulin synthesis and life expectancy

By virtue of several animal model studies in *C. elegans* (Apfeld *et al.*, 2004), *Drosophila* (Gimenez *et al.*, 2013) and mouse (Flurkey *et al.*, 2001), the role of insulin growth factors (IGFs) on the longevity of the organisms is now widely established. Mutants of IGFs showed 20%-70% increase in the average life span. Multiple mechanisms have been proposed to account for this increase such as: reduction in the insulin level, increased sensitivity for insulin, change in the carbohydrate and lipid metabolism, reduction in ROS production and delay in the onset of diseases with association to ageing (reviewed in

Bartke, 2005). Humans ageing over hundred years show exceptional insulin sensitivity (reviewed in Bartke, 2005) and low incidence of diabetes (reviewed in Bartke, 2005). This is in contrast to normal individuals from the same population which exhibit progressive increase in insulin resistance during ageing. It was also observed reduction in the amount of food intake "caloric restriction" increases longevity in healthy animals suggest it may be linked to reduced GH/IGF-I signalling (reviewed in Bartke, 2005).

1.2 The importance of age estimation

In recent years, the need for molecular approaches for forensic age estimation has increased at an exponential pace, as there are many situations in which accurate age determination is needed:

1.2.1 Crime

In most crimes, a biological sample (mostly blood) can be collected from the crime scene, but there is often no direct suspect with which to compare it. In this case if the DNA profile is determined, it can be checked for a match in the database. If no match is identified, the profile is stored in the database for further comparisons. Standard DNA profiles provide no phenotypic information beyond the sex of the individual. In the last few years, many studies have shown the possibility of predicting the race and some of the physical features of suspects and victims from the DNA that is collected from crime scenes. DNA contains information about many human traits, such as eye colour, hair colour, and height (Allen *et al.*, 2010). Forensic DNA phenotyping (FDP) is a method where the identity from traits of an unknown suspect can be predicted by analysing crime-scene DNA. This prediction could include external characteristics, behavioural features, geographic origins, and even the surname in some situations (Koops and Schellekens, 2007; Kayser and Schneider, 2009). Phenotyping can be done by either of two approaches. The first approach is indirect phenotyping, where the genetic ancestry of the person predicts the external characteristics such as skin colour and hair colour. In the direct phenotyping approach, an external characteristic is directly determined from the gene responsible for it. Indirect phenotyping has been used in a few real criminal cases and it has shown positive results (Frudakis, 2010). In contrast, the direct method is feasible, but not routinely conducted (Graham, 2008). In FDP, many traits need to be predicted to reveal the identity of a suspect. Each trait requires confirmation by using different informative markers. In 2002-2003, a panel of 71 SNPs was used to predict the race of a Louisiana serial killer (Frudakis, 2010).

In some crimes of particular interest, mass screening can be carried out. In 1986 in England, the forensic DNA profiling was first used to identify Colin Pitchfork in the rape and murder of two young girls (Wambaugh, 2011). In an attempt to identify a suspect, nearly 4,000 men from three neighbouring villages with age range between 13 and 34 were tested. This screening process excluded the original suspect (Friedman, 1999). As all adult males in three villages were asked to volunteer and provide blood or saliva samples. Colin Pitchfork, the murderer, arranged with a friend to give blood in his name. However, this friend was later overheard talking about the switch and that he had given his sample under Colin Pitchfork's name (Wambaugh, 1989). In similar crime cases, if the age of the person leaving the biological sample could be determined this would help limit the screening process to a specific age group. This will minimize the number of suspects and the number of samples to be tested. This also will decrease the cost and anxiety to the public. Moreover, this also will help police to minimize the search circle for suspects and criminal records to certain age groups or certain criminal style. For instance DNA profiling with an ability to provide investigators clues regarding the age of the individual (perpetrator/victim) from which DNA belongs to could be extremely useful especially considering the possible age difference, such analysis holds high value for cases involving kidnapping of young children.

In some criminal cases, the suspect does not have valid identification documents and age determination is important to determine if a person has legal responsibility. Suspects under 14-years old are not usually susceptible to criminal legal responsibility. Suspects between 14 and 18 are subjected to special criminal standards when they accused of a criminal offence. In most countries suspects over 18, are considered to have full legal responsibility. In the future, some countries such as Spain, suspects over 18, but under 21, may be subjected to the standards now also applied to criminals under 18 (Schmeling *et al.*, 2003).

1.2.2 Immigration

Immigrants without valid identification documents and who do not know their age are suspected of making false statements with regard to their age and whose true age is of legal relevance in criminal, civil or asylum proceedings. Illegal immigrants under 18 may be placed under the guardianship of the authorities (Garamendi *et al*, 2005). In this connection, precise estimation (not range) could only be useful.

1.2.3 Civil law

In developing countries such as Saudi Arabia, birth certificates were only routinely issued from 1980 and only a few people have birth certificates before that date. As a consequence, there is the possibility that individuals may make false statements regarding their age to meet the job age requirements and/or applying for driving licenses, and other similar age-dependent issues.

Job retirement is also age-dependent, and people who lie when they applied for the job or assigned a specific date by the government may try to misinform their age to extend their job career.

1.3 Age estimation

Current age estimation for both living and dead subjects is dependent on several anatomical and morphological characteristics. Samples in which these criteria can be measured are rarely to be found in crime scenes. For instance, bones recovered from human remains after being buried or burned may undergo degradation. Samples recovered from a crime scene are also affected by weather and time, and need to be collected as soon as possible for accurate age estimation.

1.3.1 Age estimation for living subjects

Previously, estimation of age of living subjects has been studied (Alvarez *et al.,* 2006; Tsuji *et al.,* 2002; von Wurmb-Schwark *et al.,* 2002).

Age estimation for living individuals is based on physical assessments, such as an X-ray of the left hand and the clavicles, and dental architecture. The outcomes of all these methods are grouped together to estimate the age. These investigations need to be carried out by a group of experts and cooperation between experts is required. This includes a forensic physician, an X-ray examination of the left hand by a radiologist, dental status and analysis of an orthopantomogram by a dentist. The physical examination includes measures such as height and weight and visible signs of sexual maturity (Schmeling *et al.*, 2003). The criteria obtained from X-ray images of the hand must be evaluated by analysing morphologic maturation status of all epiphyseal cartilages in the hand and size of the sesamoid bone of the metacarpophalangeal joint of the thumb. These images must be compared with standard images for the relevant age and sex using, for example, a radiographic atlas (Schmeling *et al.*, 2003).

When using teeth for age determination, the main criteria investigated include eruption and calcification stages of the permanent teeth and the size of the dental pulp cavity. It is also important to describe the average number of decayed, missing and filled teeth. Given that in the early childhood, teeth undergo significant numbers of morphological/biochemical changes, thereby they could be used for age determination. However, after reaching maturity, such alteration in tooth composition is significantly minimised hence age estimation using adult teeth is rather difficult (Panchbhai, 2011). Finally, the forensic physician will summarize all the results obtained from the various examinations to determine the final age estimation.

These procedures are laborious and time consuming. They also include radiation hazard since the examined person is exposed to X-ray in most examination phases. X-rays are of particular concern for pregnant women if they are subjected to this investigation. However, magnetic resonance imaging (MRI) may be used to minimize the radiation exposure, but it is an expensive tool and often not available (Schmeling *et al.*, 2003).

1.3.2 Age estimation at death

A number of studies have investigated the estimation of age at the time of death (Meissner *et al.,* 2006; Meissner *et al.,* 1999; Takasaki *et al.,* 2003).

Anatomical features are the most frequently used methods for determination of age at death. In 1920, Todd described the first proper method of age determination. Todd examined a sample of 306 males of known age at death and he developed a ten-phase system for age determination ranging from 18 to 50 + related to changes of the pubic face (Meindl and Lovejoy, 1985).

Epiphyseal closure (in the limb bones) undergoes alteration till the age of 25, hence could be used for the estimation of age up till then. Similarly, pubic symphysis also keeps joining up until the age of 50 therefore could be considered reliable for the estimation of individual below age 50. Schranz (1959) used both the amount and arrangement of cancellous bone in the proximal end of the humerus as methods for determining age. The amount of cancellous bone (soft bones) decreases with increasing age. However, the reliability of this method alone to estimate age is undetermined. Moreover, the method can only be applied to specimens where the proximal end of the humerus is present and intact (Kerley, 1965). Another anatomical feature used to estimate the age at death is suture closure in the plates of the skull. The method involves reading and scoring of ten-suture closer regions (Meindl and Lovejoy, 1985).

Anatomical features using a single skeletal indicator to estimate age at death are unlikely to reflect factors that accumulate with chronological age accurately but each feature can provide valuable information to the age estimation (Meindl and Lovejoy, 1985).

Moreover, age estimation from the skeletal remains requires collaborative efforts from different areas of expertise namely odontologist, anatomist and possibly palaeontologist and archaeologist. Soft tissues (blood, other body fluids and occasionally other boneless tissues such as brain or skin) are routinely found at the crime scenes therefore age estimation method could not implemented in this connection. Thus, these samples demand assessment using modern molecular biology or molecular genetics tools (Ren *et al.,* 2009).

1.3.3 Age estimation at death using biochemical and radioisotope methods.

Accurate age estimation at death using anatomical features remains a major problem. Numerous methods were developed to estimate age at death from the skeleton. These techniques rely upon degenerative changes of bones that show significant variation in the timing of their occurrence between individuals. Age estimations resulting from using these techniques have large margins of error and may underestimate the age of older individuals and overestimate the age of younger individuals. Therefore, reliable methods need to be developed which are less susceptible to such problems (Griffin *et al.*, 2009).

Biochemical methods based on amino acid racemization of aspartic acid and asparagine acid in teeth can provide more accurate estimates of age. This technique is based upon the fact that two different enantiomeric forms, L and D of amino acids can be exist, which are mirror images. In the human body, all the amino acids are initially synthesised in the L form, but this conformation is thermodynamically unstable. In teeth, a spontaneous racemization reaction will occur until an equilibrium of L and D enantiomers is formed and there is no tissue turnover with time. This process is strongly temperature dependent (Griffin *et al.*, 2009). This method is able to overcome some of the limitations associated with anatomical features, especially if followed by special morphologic dental and skeletal methods (Griffin *et al.*, 2009). However, teeth are needed to be available to conduct these measurements.

Measuring the half-life of radioisotopes is another approach to determine age at death. The reason behind using radiocarbon analysis is that aboveground nuclear weapons testing during the cold war (1955-1963) caused an extreme increase in global levels of carbon-14

(^{14}C). The change in the C-14 level has been carefully recorded over time (Alkass *et al.*, 2010). However, the subject must have been alive at the time of exposure to such radiation. This method with the use of other isotopes such as elements with longer half-lives, has been commonly used to estimate the age of fossils (Alkass *et al.*, 2010).

Sample availability is an important issue to determine which method needs to be used. Mostly, the human body is made of nucleated cells. Different types of these cells can be recovered from a crime scene. Theoretically, a DNA profile can be generated from any nucleated cell containing double-stranded DNA when ideal conditions are present (Kloosterman and Kersbergen, 2003). In addition, hair is a biological material that can be naturally shed or fall out due to a violent act. For better resolution, it is convenient to have hair samples with intact roots (follicles), as samples in such condition generally generate a reliable DNA profile (Goodwin *et al.,* 2011). Briefly, blood (white blood cells), sperm cells, epithelial cells and hair follicles are common nucleated cell types that are recovered from crime scenes. Bloodstains are the most common sample collected from crime scenes (Van Oorschot and Jones, 1997). Any method used to estimate age for healthy individual or at death must be accurate, simple, rapid, and reproducible (Ren *et al.* 2009).

In summary, using anatomical and morphological methods is expensive and laborious and needs expert examiners. Using molecular methods has the potential to be more helpful as biological samples can be recovered from many crime scenes. In forensic medicine, age determination for human remains and living individuals is an important issue. Besides DNA profiling, age estimation is not a widely used method for individual identification. However, this may be due to lack of a reliable method.

1.3.4 Estimating human age from T-cell DNA rearrangements

Specific receptors are used by T lymphocytes use, called T-cell receptors (TCRs), to recognise foreign antigens. To create a broad repertoire of TCR molecules, immature T lymphocyte experiences unique somatic rearrangements in its TCR loci during intra-thymic development. In this rearrangement process, the signal joint TCR excision circles (sjTRECs) DNA sequences in the TCR loci are deleted and circularised into episomal DNA molecules such as the **δ**Rec-**ψ**Jα sjTREC (Figure 1.2) (Zubakov *et al*., 2010).

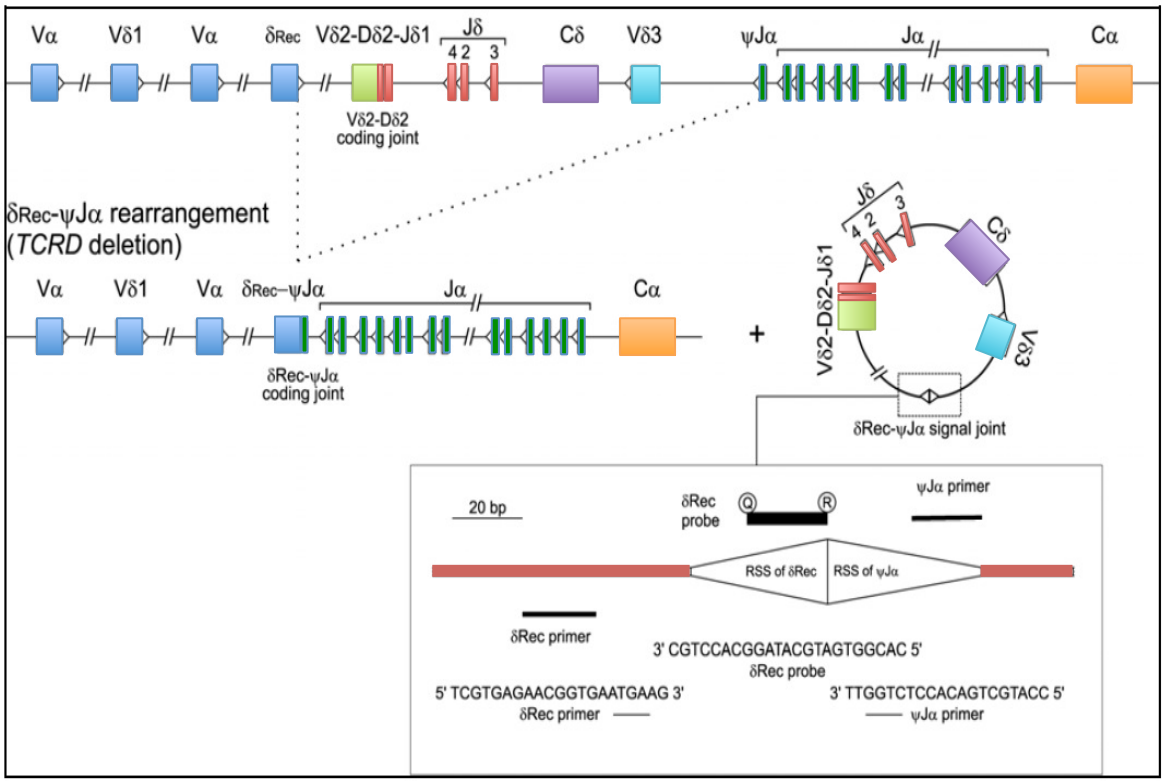


Figure 1.2: Schematic model of recombination processes leading to sjTREC formation (adapted from Zubakov *et al.*, 2010).

Approximately 70% of all newly-formed mature TCRαβ^{+} T lymphocytes encompass this sjTREC. There is linear decline in sjTREC number with increasing human age, reflecting the replacement of thymus by adipose tissue in life-long process that starts shortly after birth and consequent loss of thymic function (Zubakov *et al*., 2010). This biological phenomenon can be used for estimating human individual age accurately and reliably. A TaqMan qPCR approach was used to quantify sjTREC levels, normalised to the single-copy albumin gene to account for the amount of input DNA Zubakov *et al*., 2010. This approach, employing small amplicon sizes of 140 bp for sjTREC and 118 bp for albumin, can be successfully applied to aged blood stains and requires small amounts of DNA to analyse from 50 ng of DNA. Using blood does not require the availability of invasive samples such as bones or teeth unlike odontological or skeletal approaches for age estimation or some biochemical methods, and thus expands the availability of biological age estimations for practical applications. However, the use of this method is limited to blood samples and body parts containing blood and is not possible for other body parts or fluids, such as semen or saliva, that do not contain T cells in quantities required for sjTREC detection(Zubakov *et al*., 2010). Peripheral sjTREC level found to be significant as an age indicator and the standard error (SE) of estimation was found to be 10.47 years (Ou *et al*., 2011).

1.4 Microsatellites

Microsatellites are defined as simple DNA sequence repeats composed of 1–6 nucleotides with mostly homogeneous array that range 10–1000bp (Figure 1.3). They were discovered independently in 1989 by two research groups, Litt and Luty (1989) and Weber and May (1989). Microsatellites occupy nearly 3% of the human genome and are found on all human chromosomes. Microsatellites are also sometimes called short tandem repeats (STRs) and polymorphic microsatellites are called short tandem repeat polymorphisms (STRPs) (Kidd *et al*., 2004). STRs are commonly present in noncoding DNA and are relatively rare in protein-coding regions (Li *et al*., 2002). Dinucleotides are the most frequent STRs presents in many species (Wang *et al*., 1994; Schug *et al*., 1998). Trinucleotides are the most frequent in coding regions (Li *et al*., 2002).

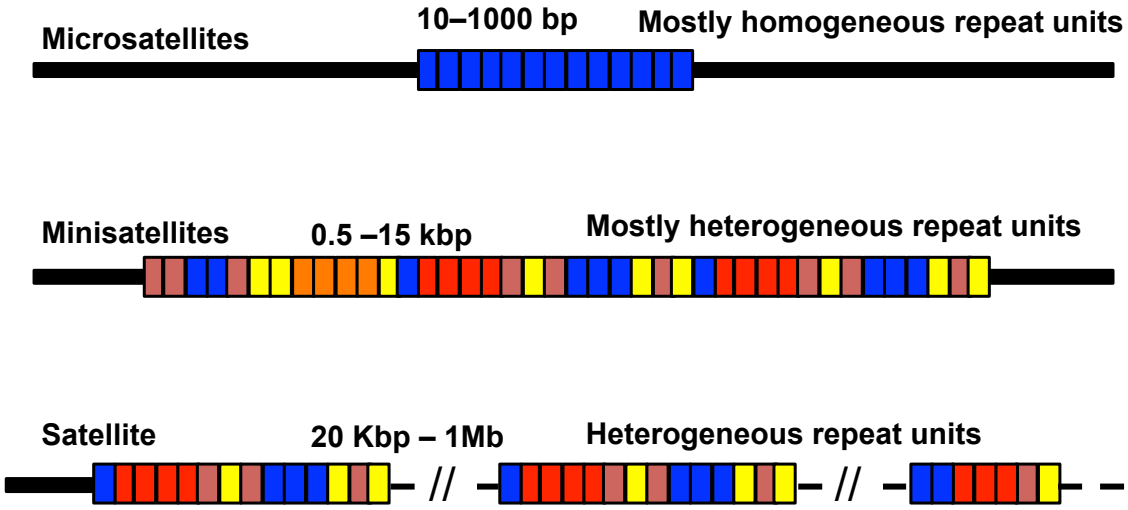


Figure 1.3: Microsatellites, minisatellites and satellites criteria.

Frequently, the number of repeats varies from one individual to another providing useful tools for individual identification, relationship testing and as markers in medical genetic studies (Jobling and Gill, 2004). Polymorphisms are genetic variants that occur in more than 1 % of the population. They are common enough to be considered normal variation in the DNA. They can be used in molecular evolutionary studies and population genetics (Meyer *et al.*, 1995; Brinkmann *et al.*, 1996b). Although the majority are believed to be selectively neutral, these repeats can be associated with both morphological changes and human disease (Pearson *et al.,* 2005). Microsatellites are characterised by their high mutation rates and levels of polymorphism depending on repeat number, length, and motif size (Galindo *et al.*, 2009). Variation in the general population is generated by germline

mutation events that change the number of repeats. In addition to variation in the germline, some simple sequence repeats are also unstable in somatic tissues and the number of variants increases with age (Higham and Monckton, 2013). They have also provided the basis for DNA profiling and for population genetic studies (Collins *et al.*, 2003). Microsatellite polymorphisms also underlie some fragile sites and certain neurodegenerative disorders are attributed to repeat sequence instability such as Huntington disease and myotonic dystrophy (Collins *et al*, 2003).

The different applications of microsatellite are:

1.4.1 Linkage analysis

Linkage analysis is an evaluation of the degree of co-inheritance of phenotypes and/or genetic polymorphisms. DNA sequence variations found between homologous chromosomes can be used to create linkage maps, which in turn can be used to map genetic disease genes. Linkage mapping of genetic disease genes requires informative markers that are distributed uniformly throughout the genome. In addition to informativeness and genomic distribution, the genotyping method should be simple and robust (Weber, 1990), characteristics, which are fulfilled by microsatellites. The first human linkage map conducted by Keller *et al.*, (1987) was based on the mode of inheritance of 403 polymorphic loci. However, given the cumbersome efforts required for minimal data acquisition, linkage mapping with microsatellites has now been replaced by SNP analysis (Frazer *et al.*, 2007).

1.4.2 DNA profiling

Human identification and relationship testing (DNA profiling) are the main aims of analysing forensic DNA samples. In 1984, Dr. Alec Jeffreys used DNA in forensics for the first time (Jeffreys *et al.,* 1985). He discovered variable number tandem repeats (VNTRs) in intronic regions of DNA and they were specific to an individual. In addition, he developed DNA profiling techniques to detect the VNTRs in humans (Butler, 2005).

Minisatellites or VNTRs are 10-60 bp repetitive stretches of DNA with heterogeneous array that range 0.5–30 kb, by contrast microsatellites are the 1-6 bp repetitive sequences present in the DNA (Figure 1.3). Both minisatellites and microsatellites are used for genomic analysis. Given the large size of minisatellites, they are highly variable and generally found to be informative for DNA finger-printing. However, due to the large size of the repetitive sequence it is often found difficult to amplify those regions by PCR,

therefore analysis of VNTR requires considerable amounts of sample and Southern blotting. In addition, comparison of samples using minisatellites is generally very difficult because of the substantial differences and relatively less accurate size estimation. Conversely, microsatellites are relatively less informative but due to the short stretches of repetitive sequence they can be conveniently amplified and small samples can provide an adequate quantity of DNA for the subsequent analysis. Moreover, microsatellite analysis provides precise measurement of the region down to single nucleotide resolution.

DNA profiling using VNTRs includes the extraction of DNA from samples linked to forensic case. Restriction enzymes used to fragment the DNA around the flanking region of VNTRs. Then, the fragments are resolved according to size 500 bp to 30 Kbp and blotted by Southern hybridisation (Jeffreys *et al.,* 1985).

In 1985, DNA fingerprinting of a family of African origin was presented to the United Kingdom Home Office to resolve an immigration issue (Jeffreys *et al.*, 1985). Since then, the UK Home Office and other law enforcement agencies considered it appropriate to be used to resolve civil and criminal investigation. In 1986, DNA was used to resolve a criminal case for the first time and this technology helped in the identification of Colin Pitchfork as the killer of two schoolgirls in Leicestershire (Friedman, 1999).

DNA profiling has come to be the conventional tool for the analysis of biological material collected from a crime scene and replaced serological method such as blood grouping. The most common samples recovered from a crime scene are from blood, semen, bones, teeth, hair, saliva, urine, or faeces, all of which are potential sources of DNA (Solomon *et al.,* 2008). Small amounts of DNA can also be extracted from nucleated cells left behind on cigarette butts, licked envelopes or postage stamps, fingerprints, chewing gum, wristwatches, earwax, debris from under fingernails, and toothbrushes (Butler, 2005).

However, the original Southern blot technique was laborious, and required relatively large amounts of undegraded DNA (~3 μg). An important technical advance occurred in 1986, following Kary Mullis' development of the polymerase chain reaction (PCR) method that revolutionised DNA profiling. PCR is a potent molecular technique allowing *in vitro* enzymatic amplification of a given DNA sequence from small amounts of starting material (Mullis, 1990). Short tandem repeated sequences represent a rich source of highly polymorphic markers for DNA profiling. Currently, short tandem repeat (STR) loci have replaced minisatellites to produce a DNA profile of an individual. STRs such as trinucleotide and tetranucleotide repeats are highly polymorphic in humans, have higher

distinguishing power than SNPs and can be easily amplified using PCR (Edwards *et al.*, 1991). Hundreds of SNPs are required to achieve distinguishing power that can be used in forensic identification because they are biallelic. Using biallelic SNPs will be problematic when investigating mixed samples recovered from crime scenes.

Currently, commercially developed kits are now available and routinely used to solve human forensic casework. They include autosomal STR multiplexes where single-tube PCRs can amplify multiple loci (reviewed in) (Jobling and Gill, 2004). Different numbers of STR loci are used. For example, in the United States of America and Saudi Arabia, forensic laboratories routinely use 16 STR loci while the United Kingdom and much of Europe use 13 STR loci with the XY homologous amelogenin gene (Butler 2006). The amelogenin varies between X and Y chromosome hence is generally used to type the sex (Wurmb-Schwark *et al.*, 2007). In Australia 9 STR loci are used in forensic laboratories with the sex-determining loci.

Genotyping using STRs is able to discriminate among individuals with a high degree of success. DNA profiling currently has the power of discriminating one in a trillion among unrelated individuals (Butler, 2005). The likelihood of two people having the same DNA profile is predicted to be $<3 \times 10^{-11}$ to 5×10^{-19} (reviewed in (Jobling and Gill, 2004).

In 1994, the UK was the first country to build a National DNA Database, a searchable record of DNA profiles, (NDNAD) (Home Office, 2013), quickly followed by the USA's combined DNA index system (CODIS) (Investigation, 2013). In 2001, Australia established their National Criminal Investigation DNA Database (NCIDD) (http://www.crimtrac.gov.au/our_services/BiometricServices.html). Similar databases and supporting legislation have been introduced in many other countries.

1.5 Microsatellite variants that cause disease

Expansion of repeat length beyond the range observed in the general population of a subset of microsatellites causes a range of inherited diseases. Microsatellite expansions are associated with at least 40 inherited human disorders (Gomes-Pereira and Monckton, 2006). The location of the expanded repeat within the affected gene is important as it may influence the pathogenicity of the expansion. Most of these disorders involve expansion of a trinucleotide repeat. Most of the medical conditions that result from repeat expansions are characterised by genetic anticipation. Anticipation is a clinical observation defined as an increase in the severity of the disease and a decrease in the age-of-onset in successive generations (Höweler *et al.*, 1989). It is explained by germline instability that is biased

towards repeat expansion and it is widely accepted that the greater the length of the repeats, the earlier will be the onset of the associated disease. The possibility of anticipation in DM1 was first reported in 1918 shortly after the recognition of the disease. It had been observed very early on that adult onset of disease was associated initially with cataracts as the only symptom in the older relatives, while the next generation were more severely affected, indicating anticipation of disease (Fleisher, 1918). Sherman proposed a phenomenon analogous to anticipation in fragile X syndrome, where, increased penetrance was found in the subsequent generations. Importantly, carrier males (males with mutant X chromosome with respect to the pedigree position) appeared to be the source of transmission. Typically, the daughters of carrier males inherited the mutated X chromosome from their fathers and consequently transmitted it onto her affected sons. This phenomenon has been referred to as the Sherman paradox (Sherman *et al.*, 1984; Sherman *et al.*, 1985). In 1991 and 1992, the molecular basis of the anticipation of the DM and fragile X syndrome was uncovered with the discovery of the expansion of CGG and CTG repeats as a contributing factor (Fu *et al.*, 1991; Verkerk *et al.*, 1991; Aslanidis *et al.*, 1992; Brook *et al.*, 1992; Buxton *et al.*, 1992; Harley *et al.*, 1992). In either case the repeat magnitude increases in the generation and has to reach to a threshold to show its pathogenic manifestations. Finally, the anticipation phenomenon was recognised as one of the characteristic features of the diseases, which is the function of dynamic mutation. The sex of the mutant allele inherited from the transmitting parents could differently influence the intergenerational stability. In certain disease such as fragile X syndrome, the repeats are more unstable and susceptible to long expansion when inherited from the mother (Nolin *et al.*, 1996). By comparison, in cases of DM1, short expansions were found to be more unstable when transferred from the male parent, while long expansions are more or less exclusively transferred from females and are associated with the congenital onset of DM1 (Brunner *et al.*, 1993a; Harley *et al.*, 1993; Lavedan *et al.,* 1993; Ashizawa *et al.*, 1994b; Jansen *et al.*, 1994). In the case of HD, repeats inherited from the father are more unstable in comparison to those inherited from the mother (Kremer *et al.*, 1995).

Even with the intergenerational repeat instability that links all diseases described earlier, the level of somatic instability varies. In diseases such as DM, high levels of somatic mosaicism have been observed (Martorell *et al*., 1998). However in HD, peripheral tissues show limited range of somatic mosaicism, nevertheless the magnitude of mosaicism could be higher in the central nervous tissues (Kennedy *et al*., 2000). Somatic instability of expanded simple repeats was firstly shown by the diffuse hybridisation signals observed when using Southern blotting to analyse restriction fragments of genomic DNA (Aslanidis

et al., 1992; Buxton *et al.,* 1992; Fu *et al.,* 1992; Harley *et al.,* 1992). These heterogeneous signals were later revealed to consist of unresolved alleles of different-sized CAG·CTG repeat arrays. Disorders associated with somatic mosaicism are likely to be age-dependent, and highly tissue-specific (Wong *et al*., 1995).

1.5.1 Myotonic dystrophy type 1 (DM1)

Myotonic dystrophy type 1 (DM1) is one of the most dramatic examples of simple repeat instability in the soma. In adults, DM1 is the most common form of muscular dystrophy and is caused by an expansion of a $(CTG)_n$ repeat within the 3′-UTR of the *DMPK* gene (19q13.3). Healthy individuals have between 5 and 35 CTG repeats (Harper, 2001). DM1 is an interesting disease to investigate the dynamics of somatic mosaicism as it is a relatively high occurrence disorder with frequency estimated to be 1 in 8,000 in the population and many patients inherit reasonably large expansions (60–1,000 CTG repeats) (Harper, 2001). DM1 somatic instability has been documented in a variety of human tissues, including peripheral blood lymphocytes. DM1 CTG repeat expansion is proven to be age-dependent, with older patients having longer average DM1 repeat lengths and broader ranges of variation observed (Gomes-Pereira and Monckton, 2006). Follow up assessment of the evolution of somatic mosaicism over time in the same patient has shown an ongoing increase in allele length (Gomes-Pereira and Monckton, 2006).

1.6 Somatic mosaicism

Experimentally, the expanded alleles in DM1 and other repeat-associated diseases appeared as a diffuse smear by Southern blot when obtained by the restriction digest of the genomic DNA (Aslanidis *et al.*, 1992; Buxton *et al.*, 1992; Fu *et al.*, 1992; Harley *et al.*, 1992). A technique to resolve this smear by characterizing the repeat length variation between individual cells was developed in 1995 using the small-pool PCR (SP-PCR) method (Monckton *et al.*, 1995b) and used to confirm that somatic instability of the repeats is age and size dependent (Wong *et al.*, 1995). The rate of repeat instability varies in different diseases and in different tissues. For instance, in DM, it has been found that expanded alleles are noticeably larger in the skeletal muscles as compared to lymphocytes (Anvret *et al.*, 1993; Ashizawa *et al.*, 1993; Thornton *et al.*, 1994; Monckton *et al.*, 1995b). Furthermore, the expanded repeats changes with the age of the individual. Although no heterogeneity has been detected in the congenital cases, increase in the level of heterogeneity has been shown as the age at sampling increases (Wong *et al.*, 1995; Martorell *et al.*, 1998). Collectively, it is acceptable to suggest that somatic mosaicism is

dependent on the age of the individual, and the tissue in which it has been explored, and biased towards expansion.

A number of transgenic mouse models and cell lines have been generated to model the somatic mosaicism (Scherzinger *et al.,* 1997; Lia *et al.,* 1998; Fortune *et al.,* 2000; Kennedy and Shelbourne, 2000; Gomes-Pereria *et al.,* 2001). Similar to the DM1 patients, in DM1 transgenic mouse models, average repeat length was found to be always higher in muscles than in lymphocytes (Fortune *et al.,* 2000). Consistently, in case of Huntington (HD) knock-in mouse models, animal with 72>CAG<80 repeats striatum has shown more repeat instability as compared to peripheral tissues. However this difference vanished in mice with 150 repeats. These observations provide profound insights into the involvement of somatic mosaicism in the disease progression and severity given that phenotypically, in HD the striatum neurons are the most affected in early stages (Kennedy and Shelbourne, 2000; Kennedy *et al.*, 2003). Overall, repeat dynamics in the mouse model mimics the features (age dependency, expansion bias and tissue specificity) found in repeat-associated diseases in humans.

1.7 DNA Repair

Genomic DNA is constantly exposed to a variety of genotoxic stresses such as UV light, ionising radiation, and oxidative stress, as well as chemical agents. To counter such stresses, DNA repair systems continuously scan the genomic DNA to identify faults and subsequently repair them. DNA repair is widely known to be involved in several cellular process like cell cycle check point control, transcription, DNA methylation, maintenance of genomic stability, proliferation and apoptosis (Bellacosa, 2001; Schofield and Hsieh, 2003).

Widely conserved from bacteria to humans, DNA mismatch repair (MMR) is one of the most important DNA repair mechanisms. Fundamentally, MMR is involved in correcting the DNA replication errors such as base-base mismatches and insertion-deletion loops, due to the error prone nature of DNA polymerase (Schofield and Hsieh, 2003). Cells lacking the MMR system are often unable to repair insertion-deletion loops of repetitive DNA sequences resulting in the gain or loss of short DNA repeats during replication, a phenomenon termed microsatellite instability (Harfe and Jinks-Robertson, 2000). This observation suggests that MMR presents itself as a main contributing mechanism to curtail the mutation rate and guarding DNA from errors appearing during DNA replication (Schofield and Hsieh, 2003).

The molecular pathway involved in the MMR pathway has been considerably unravelled in *Escherichia coli* (*E.coli),* which for the sake of comprehensibility has been divided into three steps namely initiation, excision and resynthesis. These steps are performed by an array of proteins encoded by genes including *MutS*, *MutL*, *MutH*, and *MutU*. Initiation involves recognition of base-base mismatches or insertion-deletion loops by a *MutS* protein. Subsequently, ATP-mediated binding of *MutS* and *MutL* and activation of *MutH* (an endonuclease) produces a nick in the 5′ or 3′ direction of the nascent strand of hemimethylated DNA. Then a DNA helicase, *MutU,* removes the damaged DNA from the newly synthesised DNA strand. Finally DNA polymerase III and DNA ligase fill the gap in the DNA strand (Bellacosa, 2001; Schofield and Hsieh, 2003). The human genome encodes five *MutS* homologues (*MSH*), out of which *MSH2, MSH3* and *MSH6* are known for their involvement in MMR pathway where as *MSH4* and *MSH5* contribute to meiotic recombination. *MSH2* is an essential monomer component of two different heterodimers: *MutSα* and *MutSβ* which are respectively composed of *MSH2* and *MSH6* and *MSH2* and *MSH3*. *MutS*α binds to single base mismatches and single base insertion-deletion loops while *MutSβ* exclusively binds to the insertion-deletion loops up to 12 bp (Bellacosa, 2001; Peltomaki, 2001; Schofield and Hsieh, 2003). Similar to *MutS*, four homologues of *MutL* are organised into heterodimers: *MutLα* which interacts with the single base mismatches and also to single and large base insertion/deletion loops composed of monomers of *MLH1* and *PMS2* (Harfe and Jinks-Robertson, 2000; and Aquilina and Bignami, 2001). Interaction between *MLH1* and *MLH2* constitute a heterodimer *MutLβ* that binds to the insertion-deletion loops up to 12 bp (Bellacosa, 2001; Peltomaki, 2001a; Schofield and Hsieh, 2003).

Unlike in bacteria, MMR has the ability to discriminate between the DNA template and newly synthesised strand in vertebrates. Strand recognition is mediated by free 3′ or 5′ ends of the daughter strand. This repair mechanism would not occur without initiation by the mismatch repair enzymes. The cell's ability to repair DNA is impaired if a cell is unable to faithfully correct the error in the newly synthesised DNA sequence, and if the affected cell also does not undergo apoptosis, then mutations may accumulate, eventually resulting in tumour formation.

1.7.1 Microsatellite instability in cancer

Genome integrity is central to cell survival and disturbance in it (instability) is a common trait observed in many cancers. Microsatellite instability, which is the frequent contraction and expansion of short tandem repeats in the genome, is widely associated with deficiency in the mismatch repair (MMR) machinery in humans. This impairment in DNA repair often contributes to oncogenesis (reviewed in Wu *et al.*, 2011). Instability in microsatellites has been associated with the early prognosis of gastric, pancreatic and colon cancer (Lawes *et al.*, 2003). The maintenance of microsatellite repeat integrity is the function of the MMR system, that repairs insertion deletion loops on the nascent DNA strands. Several genes are directly involved in the MMR mechanism, of these association has been established for *hMSH2* and *hMLH1*, which are found to be mutated in 40% and 50% (respectively) of cases of Lynch syndrome (hereditary nonpolyposis colorectal cancer (HNPCC), a form of colorectal cancer (Peltomaki, 2005). Recently another gene, *hMRE11*, has been directly associated with the MMR machinery. As *hMRE11*-*hMLH1* complex is the first molecule recruited at DNA double-strand breaks, it plays a pivotal role in DNA repair machinery (Vo *et al.*, 2005). Interestingly, 25% of the missense mutations found in *hMLH1* in Lynch syndrome are unable to form a complex with *hMRE11* (Zhao *et al.*, 2008), implying the association of *hMRE11* as well with the microsatellite instability and colorectal cancer. Some other studies also show association of abnormal expression of *hMRE11* with human cancers (Giannini *et al.*, 2004; Ottini *et al.*, 2004). Nevertheless, the generalisation of these findings to other human cancers awaits further investigation.

1.8 Mechanisms of unstable repeat expansion

Based on research observations generated from studies in *E.coli* and *Saccharomyces cerevisiae,* a number of models explaining repeat expansion have been proposed, segregated into two classes: one involves replication slippage, while the other includes recombination and inappropriate DNA MMR. Intriguingly, in contrast to humans, where expansion is generally noticed, among the simple model organisms contraction is more commonly observed (Wells *et al.*, 2005; Lahue and Slater, 2003). Another model which explains triplet repeat expansion mechanism is based on MMR knock-out transgenic mice models (Gomes-Pereira *et al.*, 2004b).

1.8.1 DNA polymerase slippage model

The body of evidence suggests expansions and deletions during DNA replication, repair and/or recombination occur concomitantly (Wells *et al.*, 2005). The first evidence of this connection emerged from studies conducted on non-mammalian systems, which suggested that mutation of unstable CAG•CTG were the likely result of replication slippage during the cell cycle (Richards and Sutherland, 1994; Wells *et al.*, 1998).

In addition to MMR, replication slippage is another important mechanism that may be involved in repeat instability, especially in sequences like CTG•CAG, GAA•TTC and CGG•CCG repeats (Wells *et al.*, 2005). In a simple replication slippage model, the nascent strand is separated from the template DNA strand during replication, which allows the daughter strand to slip relative to the parent strand. However, subsequent reannealing of the nascent strand to the template DNA renders misalignment on the repeat region resulting in the mutation. This raises the possibility of backward slipping of the newly formed DNA strand and re-initiation of replication at a new point. Consequently, a corresponding repeat region is added to the DNA strand given the secondary structure formed by the added repeat is not repaired by the MMR before the initiation of second round of DNA replication (Figure 1.4). Conversely, forward slippage of the nascent strand results in the loss of repeat (deletion/contraction) (Figure 1.4) (Richard and Sutherland, 1994). This proposal is further strengthened by observations obtained from the bacterial and yeast empirical studies which suggest that unusual structures formed by the repeat region may lead to the changes in the repeat number during DNA replication (Ohshima and Wells, 1997; Pelletier *et al.*, 2003).

It has been demonstrated that trinucleotide repeats tend to form stable secondary structures that differ structurally from typical double-stranded slipped-DNAs conformation (Pearson and Sinden, 1996). Moreover, these secondary structures contribute in the expansion process by acting as a mutagenic intermediate that impairs normal processes like replication or transcription (Callahan *et al.*, 2003; Wells *et al.*, 2005). Among those secondary structures included are hairpins (Petruska *et al.*, 1996; Darlow and Leach, 1998), triplex (Bidichandani *et al.*, 1998; Gacy and McMurray, 1998) and slipped-stranded structures (Pearson and Sinden 1996; Pearson *et al.* 1997; Pearson and Sinden, 1998). Moreover, studies in the same model organisms also suggests that contraction and expansion of repeat tracts occurred predominantly in the lagging strands of nascent and template DNA strands respectively (Freudenreich *et al.*, 1997; Miret *et al.*, 1997). The present model provides suitable explanation for the small changes in the repeat regions,

however, multiple slippage events would be required in a single replication round to account for the large repeat expansions. Alternatively, if the repeats are present in the Okazaki fragments, which are prone to slippage, expansion and contraction rate of repeats would be noticeably higher (Sarkar *et al.*, 1998).

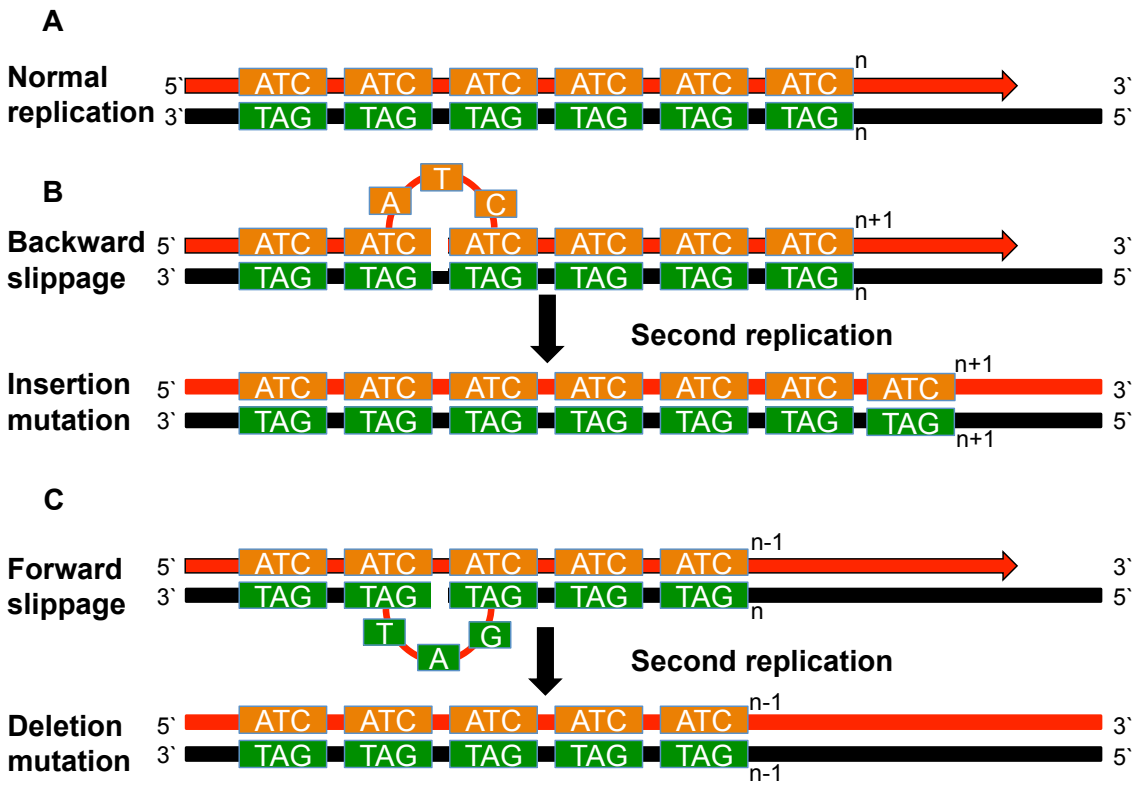


Figure 1.4: Replication slippage. Mispairing involves only one repeat. (a) Normal replication. (b) Backward slippage, resulting in the insertion mutation. (c) Forward slippage, resulting in the deletion mutation (adapted from Molecular Biology Web Book).

Moreover, studies in the same model organisms also suggests that contraction and expansion of repeat tracts occurred predominantly in the lagging strands of nascent and template DNA strands respectively (Freudenreich *et al.*, 1997; Miret *et al.*, 1997). The present model provides suitable explanation for the small changes in the repeat regions, however, multiple slippage events would be required in a single replication round to account for the large repeat expansions. Alternatively, if the repeats are present in the Okazaki fragments, which are prone to slippage, expansion and contraction rate of repeats would be noticeably higher (Sarkar *et al.*, 1998). Additionally, it has been proposed that if the Okazaki fragments are not processed by the flap endonuclease (*FEN1*), gap filling and ligation, this in turn could lead to the overall large expansion of repeats resulting in genetic instability (Callahan *et al.*, 2003). Nevertheless, direct evidence of the replication-mediated repeat expansion in mammals is still lacking in *in vivo* mammalian models (Pearson *et al.*, 2005; Wells *et al.* 2005).

The data collected from prokaryotic models favour a replication slippage mechanism, but there is very little direct support for this model in eukaryotic cells. Without a doubt, the absence of connection between proliferation and somatic instability, and the need for efficient MMR machinery to create expansions, argues against the replication slippage model (Gomes-Pereira and Monckton, 2006). Moreover, the replication slippage model cannot explain the accumulation of the large length changes observed over time in post-mitotic tissues such as muscle and brain (Gomes-Pereira and Monckton, 2006).

1.8.2 Recombination-dependent mechanisms of expansion

Another proposed underlying mechanism of repeat instability is recombination. Homologous recombination events between flanking regions of repeats are relatively rare events, but limited data are available in this regard for other model organisms (Pearson *et al.*, 2005; Wells *et al.*, 2005). Recombination events within the repeat tracts by intrachromosomal exchange between sister chromatids was thought to contribute in the germline expansions at the DM1, DM2, FRAXA, SCA8, SCA10 and FRDA loci (Pearson *et al.*, 2005). However, empirical support for this assumption has been found only in bacteria and yeast, which suggests that expansions of CTG repeats might occur by unequal crossing over or gene conversion (Jakupciak and Wells 1999; Jakupciak and Wells, 2000b; Jankowski *et al.,* 2000; Wells *et al.* 2005). Gene conversion processes found in bacteria and yeast may also by associated with both somatic and germline instability of triplet repeats in humans (Jakupciak and Wells, 1999, 2000a).

Recombination requires DNA double-strand breaks (DSB), which halts replication at the unprocessed Okazaki fragments or at the repeat regions (Ohshima and Wells, 1997; Samadashwily *et al.,* 1997; Freudenreich *et al.*, 1998). Earlier, it has been demonstrated that the repair of DSB within CTG•CAG repeats alters the number of repeat units (Marcadier and Pearson, 2003). Though most of these studies have been conducted in bacteria (Sarkar *et al.,* 1998; Jakupciak and Wells, 2000a; Napierala *et al.,* 2002) and yeast (Freudenreich *et al.,* 1998; Richard *et al.,* 1999; Jankowski *et al.,* 2000; Richard *et al.,* 2000), the exact degree and role of recombination in repeat expansion in humans is largely elusive (Pearson *et al.,* 2005; Wells *et al.,* 2005).

1.8.3 Inappropriate DNA MMR

Over the last few years a considerable quantity of data have been generated supporting the involvement of MMR in the repeat instability in both the germ line and somatic tissues in mice (Manley *et al.,* 1999; Kovtun and McMurray, 2001; van den Broek *et al.,* 2002; Savouret *et al.,* 2003; Gomes-Pereira *et al.,* 2004b; Savouret *et al.,* 2004; Foiry *et al.,* 2006), which is potentially independent of the cell cycle and biased toward expansion (Gomes-Pereira *et al.,* 2004b). In the models based on inappropriate MMR repair, the repeat sequences do not form secondary structure during replication. However, at some stage of the cell cycle the DNA helix opens to become a single-stranded molecule and during reannealing they form slipped DNA structures where the mismatched regions loop-out to 1-3 repeats (Figure 1.5). Following this, interplay of *MutSβ* and MutLα or MutLβ incorporate the loop-outs into the DNA strands resulting in the expansion of repeats. Multiple occurrences of such events could result in the large expansion of repeats observed (Gomes-Pereira *et al.,* 2004b).

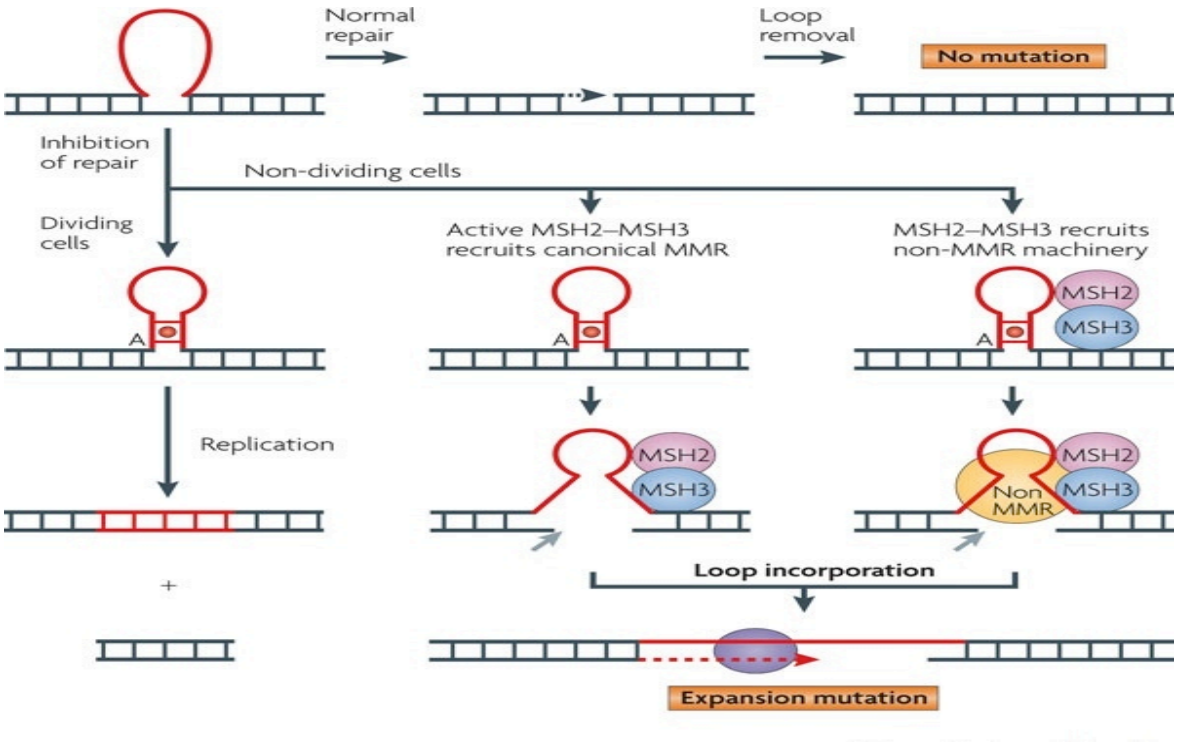


Figure 1.5: Loop incorporation into DNA duplex. Normally, mismatch repair (MMR) machinery removes small loops in DNA (red) during replication). Failure to remove the loops can lead to expansion. In dividing cells, the uncorrected loop (shown here as a hairpin with an A mismatch (red circle) is copied into DNA during the next round of replication (adapted from McMurray, 2010).

1.9 Modifiers of repeat dynamics

Several factors mediate the intricate mechanism of repeat dynamics in simple tandem repeats, which are primarily divided into two groups namely *cis*-acting and *trans*-acting modifiers. *Cis*-acting modifiers affect the repeat dynamics as they are physically linked to the repeat sequence or region present in the proximity of the sequence, whereas *trans*-acting modifiers may be on another chromosome, but interact with repeats to show their effect (Richards, 2001; Cleary and Pearson, 2003; Pearson *et al.*, 2005).

Genetic factors that induce DNA instability are inherited and affect the individual genotype. Anticipation, which characterises triplet repeat disorders, is a good example as the progeny inherit expanded or contracted repeats. Gene-specific *cis*-elements and *trans*-acting DNA proteins determine the patterns of inherited and tissue-specific instability. *Cis*-elements that might affect instability can be either internal, including repeat sequence, tract length and purity, or external, including flanking sequence elements, nucleosomes, CpG methylation and replication origins close to the repeat tract (Pearson *et al.*, 2005).

Several *cis*-acting factors have been found to be associated with mediating the instability of triplet repeats. Some of the major factors include sequence and number of repeats, the presence or absence of interrupting repeat units and the sequence composition of the flanking sequences. Only certain triplets of repeats are unstable, amongst which $(CNG)_n$ are the most common motif recognised for repeat instability. The underlying mechanism for repeat instability may be the formation of unusual DNA structures which consequently may impair replication and repair resulting in genetic instability (Chen *et al.,* 1998; Pearson and Sinden, 1998; McMurray, 1999; Sinden *et al.,* 2002). In general, shorter and interrupted alleles are more stable than longer and pure stretches of repeats. The microsatellites mutation rate associated with motif length, copy number, array length, distance from exons, sex and age, but not with recombination rate, distance from telomeres, and parental heterozygosity (Ashley and Warren, 1995: Sun *et.al*, 2012). It has been demonstrated that attaining a particular threshold of the repeat length is a pre-requisite to confer their instability. It has also been found that the GC content of the flanking DNA to repeats is positively correlated to their stability *i.e*. the higher the GC content of the flanking region the more instability is observed in the repeats (Brock *et al.*, 1999). It is proposed that GC content alters the chromatin structure and/or methylation status of genomic DNA which in turn facilitates the repeat mediated genetic instability (Brock *et al.*, 1999).

Copy number and repeat length are also associated with repeat instability. It has been demonstrated that the repeat array needs to exceed a stability threshold to become unstable (Pearson *et.al*, 2005). In myotonic dystrophy (DM1) this instability threshold is~37 CTG repeats in the 3′UTR locus (Pearson et.al, 2005). Moreover, recent studies have also pointed toward repeat purity as another crucial element affecting repeat instability. Pure repetitive sequence, consisting of 100% match or no interruptions, can potently affect instability. An interrupted repeat tract, identified in some DM patients, appears to stabilize the repeats, delay the age of onset, and decrease the disease severity (Braida *et.al*, 2010). Interruptions can significantly increase repeat sequence stability for example, the presence of a single CAT interruption can result in somatically stable CAG in SCA1 alleles that are as large as 39 repeats, whereas alleles that have 40 'pure' repeats are unstable (Pearson *et al.*, 2005).

1.9.1 Trans-acting modifiers of genetic instability

Trans-acting modifiers of genetic instability include the sex of the transmitting parent, tissue specificities and age of the individual. Other proposed *trans*-acting modifiers include protein molecules involved in some housekeeping functions such as replication, DNA repair and recombination. There are a number of candidate proteins in this regard including *MSH2, MSH3* and *PMS2*, which are involved in mismatch repair. It has been demonstrated that absence of these proteins significantly increases the genomic instability in mouse models (Wind *et al.*, 1995; Hegan *et al.*, 2006; Abuin *et al.*, 2000).

1.10 Age-dependent somatic mosaicism of microsatellites

One major shortcoming of the small pool PCR method is the extent of amplification of the DNA required to assess less frequent mutations, which could be laborious and extensively time consuming. In order to measure mutation with 2% frequency, an average of 100 alleles needs to be amplified from single DNA molecules. Therefore, Coolbaugh-Murphy *et al.* (2005) developed robotic technology augmented by advanced DNA sequencing apparatus; the precision and accuracy could be improved with considerable time saving.

They used this robotic apparatus to examine six microsatellites to measure mutation frequencies in different age groups. The Peripheral Blood Lymphocytes DNA samples were collected from 17 normal blood donors. The selected samples were segregated into three groups, among which six were 20-30 y/o, five were 35-50y/o and six were from 60-70 y/o (Coolbaugh-Murphy *et al.*, 2005). The result obtained showed a highly significant

increase in the mutation frequency with age (< 0.01 for 20–30 y; < 0.02 for 35–50 y/o; < 0.04 for 60–70 y/o).

The study demonstrated an increase in the mutation frequency in normal somatic cells with age suggesting a linear relationship between somatic mosaicism with age (Coolbaugh-Murphy *et al.*, 2005). However, a limited number of microsatellites were investigated using SP-PCR and there is still very little known about somatic instability at most loci. This approach is technically demanding and has not been widely replicated.

1.11 Project hypothesis

Simple-sequence DNA repeats are present throughout the human genome. Frequently, the number of repeats varies from one individual to another providing useful tools for individual identification, relationship testing and as markers in medical genetic studies. Variation in the general population is generated by germ line mutation events that change the number of repeats. In addition to variation in the germ line, some simple sequence repeats are also unstable in somatic tissues and the number of mutations increases with age (Coolbaugh-Murphy *et al.*, 2005). The magnitude of somatic mosaicism increases with age therefore it is conceivable that older people exhibit high somatic mosaicism in comparison to younger individuals. However, at most loci the frequency of mutations in the soma is very low but many loci were not investigated. In this project we will seek to identify simple sequence repeats in the human genome that have relatively high rates of mutations in somatic tissues. Such markers may have significant application in ageing biological samples (*e.g*. scene of crime samples) and as tools in ageing research and addressing fundamental problems in DNA repair and individual specific variation in mutational dynamics.

This study aims to estimate human age from DNA extracted from biological samples collected from crime scenes (mainly blood) by analysing the natural instability (somatic mosaicism) of the nuclear DNA in the general population.

In summary, the aim of the project is to :

1. Identify microsatellites with relatively high somatic mutation rates in the general population.
2. Investigate the level of somatic instability in the general population and the age - dependent progression of repeat instability.
3. Define a panel of microsatellites that can be used to predict age of human tissues.
4. Investigate new technologies (since small pool PCR is laborious), such as high throughput DNA sequencing.

2 Materials and methods

2.1 Materials

Standard suppliers, such as: Fisher Scientific, Invitrogen, Sigma-Aldrich, provided chemicals, molecular biology reagents and plasticware unless otherwise stated.

2.1.1 Oligonucleotides

The oligonucleotide primers used in this project were designed using the CLC Genomic workbench 3.0 and were purchased from Sigma-Aldrich. The primers' sequences are listed in Table 2.1.

Table 2.1: Oligonucleotides

Target	FWD. Sequence (5'-3')	REV. Sequence (5'-3')	AT (°C)
19-AT-431	AGGGAAGGGGCAGTGTCAAA	GTGGGGGAAGGAGTGAGAGT	54.6
X-AGAAT-759	TGGGCAACAGGAGCAAAACTC	ATCCCAGCCCTTCCTAGCCA	60
17-TCTT-376	CACCTCAGCCCCTCACAATG	CAGCCTGGGCAAAAAGAGCA	58.3
14-TTTC-814	ATCCAGAGGACACAGCAATCC	GAAGGTGGAGCTTGCAGTGAG	60
7-AT-232	TCAGAAATGTGGAGGGGCATGT	GAGGCACTCCGTACTACAGGGG	57.4
2-CA-369	GGCGCGGGGAGGTTTGCAGA	CCGCCTGCTCTTTGTCCTTT	58.4
21-TCCCT-409	TGACAGACCTCCAGTGTT	GGTGGCACGCATCTGTAGTC	53.6
10-CA-994	CACCATGCCTGGCTAATTTCTC	AGCATCACCTCATTCTGCCG	54.4
1-TTC-102	AATGAGCCAAAAAATATACA	CTTCCTGGGCCCTATTTCAG	57
15-CA-519	AGCCACAACCCCTTCTCTAGG	AGGCTGAGGCAGGTGAATTG	60
19-CA-424	GGGCTGGGGTGTAGTTAGTC	GTGGTGCAGTGGAGGTATCTG	55
5-AGA-126	GGGGAGAGAGAGCGAGAGAAAG	CTGCCCCCTCAGCTCAATC	57
2-CA-751	CCCAAACAACAAACAGCTCCCC	CAGCCACAGTCCAGGCAACA	60
X-AT-535	GCCTACCAAGGGTTGTAACT	GTCAGGCATAGTGCACATGC	----
2-CA-113	ACCGGCCCTCGATTTTCTTC	GTGTGAACCACTGTGCCCAG	60
20-CA-417	GCCACCTTCCAACATCAAGCAA	CCTGATCCCCACCCCAATGTTA	60
7-TTTC-132	GCACAAGTGTGTAATTGCTGG	GGGCGACAGAGTGAGACCCTG	58.4
16-TTTC-505	CAGGCGACAAGGGATCAAAT	CTGCGCTTCTCCAATCAATCCT	58.4
11-TC-107	CAATCCGGAAGGCACAGGT	GGGCTGTTTGGGGGAACAT	54.6
1-CTCCCT-151	GTCTCATGAGACATGAGAGA	AGGCAGAGAGGCTCCTCATA	53.6
2-AGGAA-603	TGGGCAACAGGAGCAAAACTC	ATCCCAGCCCTTCCTAGCCA	60
2-CA-181	CCCTTCCATTTCCAAGATACTC	CTGTCTCTCATCACACGTCCTC	60
13-AAG-102-i	CCTGCCGTTCCAGTTGCCA	GCTCATGGCCGCTTTTGC	57.4
13-AAG-102-o	CCTGCCGTTCCAGTTGCCA	CAGAAAGGCTCATGGAGGTG	58
13-AAG-102-ins	CTGTTAGTCATAGTACCCCAAG	CAGAAAGGCTCATGGAGGTG	59.5

2.2 Techniques

All kits and reagents were used as per the manufacturer's instructions, unless otherwise stated.

2.2.1 Polymerase Chain Reaction (PCR)

2.2.1.1 Standard PCR

Genomic DNA (10-20 ng) was amplified in a 10 µl reaction, with 1 µM of each primer, 1U of *Taq* DNA polymerase (Sigma) and 1X PCR buffer (45 mM Tris•HCl pH 8.8,11 mM ammonium sulphate, 4.5 mM $MgCl_2$, 0.048% (v/v) 2-mercaptoethanol, 4.4 µM EDTA, 1 mM dATP, 1 mM dCTP, 1 mM dGTP, 1 mM dTTP, 0.113 mg/ml BSA) (Jeffreys *et al*., 1990). The reactions were performed in a thermal cycler (Biometra and PeqSTAR-96 Universal Gradient) cycled through 30 rounds of (96^{o}C for 45 sec, annealing temperature (AT) for 50 sec, 70^{o}C for 1 min) and 70^{o}C for 10 min.

2.2.1.2 Small-pool PCR analysis

A small-pool polymerase chain reaction was used to detect and quantify repeat length variations of microsatellites (Monckton, *et al.*, 1995). DNA samples were serially diluted in 1xTE and 0.1 µM of forward primer. DNA samples were diluted down to different ranges of concentrations, equivalent to 1 ng/µl, 500 pg/µl, 100 pg /µl, and 10 pg/µl, equivalent to 167 DNA molecules to 1 DNA molecule. PCR was then carried out using DNA concentrations appropriate to the question (see the Results for more details).

A standard PCR was for 30 cycles and 1 µl of diluted DNA was amplified in a final volume of 7 µl containing 0.2 µM of each primer, 0.35 U of *Taq* DNA polymerase (Sigma) in 1X PCR buffer.

2.2.2 Electrophoresis

2.2.2.1 Agarose gels

For bulk PCR products (10 ng input of DNA, 1.5% (w/v) agarose gels (30 X 15 cm) were used to resolve the DNA fragments. Agarose gels (~300 ml) were run in 0.5 X TBE buffer (45 mM Tris, 45mM boric acid, 1 mM EDTA pH 8) with 0.5 µM of ethidium bromide and run at a constant voltage (100-130 V/cm). A UV trans-illuminator (UVP Image Store 7500 system wavelength), which has 302/365 nm wavelengths, was used to visualise the DNA fragments and photograph them. Molecular Imaging Software 1D 5.0.6.20 (Carestream)

was used to size the bands by comparing them to molecular weight markers (1 kb+ ladder) supplied by Life Technologies.

For SP-PCR, 3 µl of 5X Orange G loading dye (0.2% (w/v) Orange G, 15% (w/v) Ficoll, 1X TBE) was added to the 7 µl of PCR, and 5 µl of the mixture was loaded onto 1.5% (w/v) agarose gels (20 X 40 cm) in an 0.5 X TBE buffer to resolve the DNA fragments. Firstly, 300 V was applied for 15 minutes, then the power was reduced to 180-240 V for 19 hours at 4^{o}C. The molecular weight marker used was a 1 kb+ ladder (60 ng/ml 1 kb ladder, 1X DNA loading dye in 1X TBE).

2.2.3 DNA transfer from the gel onto nylon membranes by Southern blot

To transfer DNA from gels to nylon membranes, Southern "squash" blots were performed. Firstly, the gels were immersed in distilled water. The rinsed gel was then washed with gentle shaking in the following solutions: 10 minutes in a depurinating solution (0.25 M HCl), 30 minutes in a denaturing solution (0.5 M NaOH, 1.5 M NaCl) and 30 minutes in a neutralising solution (1.5 M NaCl, 0.5M Tris pH 7.5). The gels were rinsed in distilled water between washes. For the purpose of blotting, cling film was placed on the bench followed by one wet sheet of gel blotting filter paper. Then, the gel was inverted and the membrane (Hybond-N form GE Healthcare) put on the top side face gel was marked with pencil, on two sheets of gel blotting filter paper. Finally, a thick layer of paper towels, a glass plate and a weight (~1 kg) were placed on top. The nylon membranes were wetted first in distilled water and then in a neutralising solution and the gel blotting papers were wetted in a neutralising solution. No bubbles were allowed in any of the layers, and were removed carefully using a clean glass pipette. The blots were left overnight to ensure that the DNA transferred to the membrane by capillary action. The nylon membranes were incubated at 80°C for ~2 hrs for drying. The dried membranes were UV crosslinked using a Stratagene UV crosslinker 2400 at 1,200 J/m^2 to fix the DNA.

2.2.4 Preparation of DNA probes

Locus-specific probes were generated using both forward and reverse primers of each pattern, in order to amplify (20 ng) of genomic DNA. The PCR products were loaded onto 1.5% (w/v) agarose gel in 0.5X TBE buffer and resolved for 2~3 hours at 120 V. A gel slice containing the desired probe fragments was purified using the QIAquick gel extraction kit from Qiagen. The concentration of purified DNA was measured with a Nanodrop ND-100 Spectrophotometer and (20-30 ng) and the DNA marker (2.5 ng) was radiolabelled with 1.85 MBq [α- ^{32}P] dCTP (3,000 Ci/nmol) using ready-to-go DNA

labelling beads (GE Healthcare), incubating the reaction at 37°C for 30 min. Then, the probe was denatured for 5 mins at 100°C and cooled on ice prior to adding to the hybridisation buffer.

2.2.5 Southern hybridisation

The membranes were wetted in distilled water, rolled up and placed in hybridisation glass bottles (with the DNA side facing inwards) with 5ml of hybridisation buffer (7% (w/v) SDS, 1 mM EDTA, 0.5 M Na_2HPO_4) and rotated for ~1 hour at 65°C in a rotating oven. The denatured probes were added to the bottle containing a fresh 5 ml of hybridisation buffer. The membranes were hybridised overnight in a rotating oven at 65°C. Later, the blots were washed with a high-stringency washing solution (0.2% (w/v) SDS, 0.2X SSC): 2x 30 min at 65°C and 1x 30 min at room temperature. The membranes were dried at 80°C, exposed to X-ray film at room temperature and then developed using an X-Ograph Compact X2 system (X-Ograph Ltd.). Kodak supplied both the fixing and developing solutions.

2.2.6 Sanger sequencing

Gel-purified PCR fragment and appropriate sequencing primers were sent for sequencing to Source BioScience UK Limited, 1 Orchard Place Nottingham Business Park, Nottingham, NG8 6PX, UK. The sequences were visualised and analysed using Codon code Aligner V4.1.1

2.2.7 Next generation sequencing

One of the project's aims was to investigate new technologies, since small-pool PCR is laborious to use. High throughput DNA sequencing was used to capture a set of target regions such as microsatellites and telomere regions (which are discussed later in the study). In order to use this approach, custom probes (baits) were used to capture the target sequences.

2.2.8 Probe (Bait) design

Different strategies were used to design baits depending on the properties of the target sequences. Some of these design methods are novel and both the sequencing approach and bait designing are novel too. Each of these strategies will be explained in detail later in this thesis, in Chapter Four.

All the baits were designed using free eArray Agilent designing tools (currently known as SureDesign) provided by Agilent Technologies. The total size of the captured sequences

covered by these baits was 1.5-2.9 Mb. Two SureSelect^{XT} Custom kits were ordered. The total number of baits was 57,646, and 49,833 of these were unique (*i.e*. there was a single copy of each).

2.2.9 NGS library preparation

Human DNA samples (3 μg) were fragmented using a Bioruptor UCD 200. The fragmented DNA and the finished libraries were run on an Agilent 2100 Bioanalyser on high sensitivity DNA chips to assess the quality and ensure the desired average length of 200-300 bp. The Illumina GA IIx platform, and the Agilent SureSelect^{XT} Target Enrichment System for Illumina Paired-End Sequencing Library were used to generate the sequencing reads.

The DNA fragmentation, library preparation, and sequencing were carried by the Polyomics facilities at the University of Glasgow.

2.3 Data analysis software

Prism 6.0 was used for the linear regression analysis. For other relatively simple statistical calculations and charts, Microsoft Excel was used.

2.3.1 Online bioinformatics resources

2.3.1.1 Tandem repeat database (TRDB)

TRDB (https://tandem.bu.edu/cgi-bin/trdb/trdb.exe?taskid=1) was used to identify repetitive patterns in DNA sequences.

2.3.1.2 RepeatMasker

RepeatMasker® (http://www.repeatmasker.org) was used to screen DNA sequences for interspersed repeats and to identify unique DNA sequences.

2.3.1.3 Human BLAST search

The human BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was used to identify any similarities between the sequences present in the database.

2.3.2 Software

2.3.2.1 CLC Genomic workbench

The CLC Genomics Workbench (http://www.clcbio.com) was used to design the primers and for the analysis of the data generated from the NGS experiment.

2.3.2.2 Bowtie 2

Bowtie 2.0 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) was used to count and map the readings to the human genome reference (HG19). The Bioinformatics unit at Polyomics at the University of Glasgow carried out this process.

2.3.2.3 Integrative Genomics Viewer (IGV)

The sequences were also visualised and analysed using IGV. This free software is capable of showing the bait, reference and sample to be tested on a single viewing page (Thorvaldsdóttir *et al.*, 2013).

3 Investigation of microsatellite somatic instability in the general population

Tandem repeat sequences occupy ~3% of the human genome (Lander *et al.,* 2001). Often, these sequences are polymorphic in copy number (Litt *et al*., 1989). High levels of variability make these sequences ideal genetic markers and they are commonly used for linkage analysis, DNA profiling and relationship testing (Gulcher *et al*., 2012; Thompson *et al*., 2012). Their variability is based on a high rate of *de novo* germline mutation that generates alleles of different length (Weber and Wong, 1993). In the soma, microsatellites are similarly unstable and *de novo* mutation can be detected in tumours caused by impaired DNA mismatch repair (Geiersbach and Samowitz, 2011). Similarly, in normal tissues, age-related *de novo* somatic mutations were identified, but with very low frequencies which are difficult to measure and required sophisticated techniques (Coolbaugh-Murphy *et al*, 2005).

Usually, repeat length variations have no phenotypic effect. However, some expanded microsatellites cause inherited human disorders such as HD and DM (discussed in chapter 1). In these conditions, repeat copy numbers have expanded more than the general population observed allele length (<40 repeats) and may reach >1,000 repeats in some disorders. In such conditions, the alleles become highly unstable and the germline mutation frequencies are nearly 100% (Gomes-Pereira and Monckton, 2006). These mutations favour expansion and anticipation is observed in many of these conditions. They are also somatically unstable. The process is age dependent, tissue specific and favours expansion too. The progress and disease severity are thought to be proportional to the expansion rate (Morales *et al*, 2012; Swami *et al*, 2009).

The study of somatic instability is limited by the availability of the samples from patients with rare diseases and the relative instability of non-expanded alleles in the general population. Thus, in this project the aim is to identify large expanded microsatellites in the general population to test their somatic instability.

3.1 Results

The first step in investigating the level of somatic instability of microsatellites in the human genome is to identify microsatellite loci that are large and expanded in the general population. For this process the publicly available and easily accessible Tandem Repeats Database (TRDB) was chosen (Gelfand *et al*, 2007). The database screens any given genome for all tandem repeats ranging from 1 to 2,000 nucleotides using the Tandem Repeat Finder program (Benson, 1999).

In the Human Genome Project, a clone-based approach was used. This approach used overlapping large segments of genomic DNA. These segments are called and generally contained 40,000 – 200,000 nucleotide bases. Then, clones are into bacteria vectors, to generate large amounts of DNA for sequencing. Individual clones were sequenced by fragmenting them into reads, sequencing the reads and then assembling the reads to generate the full clone sequences, which are called components. Mapping methods were used to determine overlapped components, and then used these data to organise the components into a linear arrangement. Assembling the sequence is continued into larger units as long as there were overlaps between individual components. The resulting "top-level" sequences that have no gaps and represent a continuous region of a chromosome are called contigs or scaffolds. To generate chromosome models, scaffolds that belonged to individual chromosomes are ordered and joined. For the human genome, three reference sequences were available; the human genome reference sequence (HG19), the J. Craig Venter genome sequence and the Celera genome sequence, each of which reports a different number of tandem repeats for each chromosome using the TRDB default parameters (Table 3.1). The total number of tandem repeats identified for HG19, Venter, and Celera were 942,700, 863,896, and 903,560 respectively (Table 3.1) which individually represent ~4 % of the human genome size (Table 3.1). Chromosomes 1 and 2 have the largest number of tandem repeats in all three genomes with 72,306 and 73,515 repeats respectively in HG19 (Table 3.1). Chromosome Y and chromosome 21 showed the lowest number of repeats with 13,088 and 13,393 repeats respectively. In most cases the number of repeats identified is proportional to the length of the respective chromosome (Table 3.1).

To assess the possibility of over / under representation of microsatellites across all the chromosomes, a chi square test was performed. The observed repeats were retrieved from TRDB while the expected number of microsatellites was calculated assuming an equal distribution of microsatellites on all chromosomes proportional to their length. Our

significant Chi-square threshold with 23 degree of freedom when P value = 0.05 has a value of 35.2 in Chi-squared table. This means that when we calculate the Chi-squared, and given the final values is >35.2, the null hypothesis will be rejected. The sum of Chi-square values showed significant deviation (P value = < 0.0001) in both human genome references. For instance in chromosome number 15, the observed repeats were found as 25,186 and 22,893 for HG19 and Venter reference sequences. However, the expected number of repeats is 31,078 and 28,480 with HG19 and Venter respectively (Table 3.2). The Chi-square values showed significant deviation indicated that number of repeat observed in the both human genome references were under represented in chromosome 15 (Table 3.2). Inversely, the number of repeats observed in chromosome 19 were over represented given the observed repeats were 30,512 and 29,263 for HG19 and Venter reference sequences. The expected repeats were 21,660 and 19,850 respectively. Significant Chi-square values obtained implied that number of repeats observed in both the human genome references were over represented in chromosome 19.

Table 3.1: Number of tandem repeats identified on each chromosome for the three human reference genome sequences using the TRDB (Boby *et al.,* 2005 and Benson, 1999)

Number of repeats (HG19)	Chromosome	Number of repeats (Venter)	Number of repeats (Celera)	Chromosome size (Mb)
72,306	1	67,980	68,993	249
73,515	2	70,414	70,531	237
56,100	3	54,872	55,028	192
56,029	4	54,614	54,548	183
53,215	5	51,359	51,209	174
50,801	6	49,919	52,945	165
54,365	7	52,411	51,558	153
45,473	8	44,178	44,430	135
38,534	9	34,282	34,468	132
44,387	10	42,564	42,628	132
40,687	11	40,115	40,019	132
43,560	12	42,407	43,957	123
29,387	13	28,856	29,339	108
27,552	14	26,381	26,846	105
25,186	15	22,893	22,544	99
33,498	16	32,063	29,059	84
31,050	17	29,581	29,607	81
23,571	18	23,170	22,906	75
30,512	19	29,263	30,293	69
21,540	20	20,895	21,432	63
13,393	21	12,353	12,588	54
15,255	22	14,650	15,202	57
49,696	X	11,433	49,521	141
13,088	Y	7,242	3,909	60
942,700	**Total**	**863,896**	**903,560**	**3,003**
122,223,156	**Array Length (bp)**	**110,149,776**	**109,118,348**	**3,003,000,000**
4.01	**(%)**	**3.67**	**3.63**	**100**

Table 3.2 Comparison of the expected and observed values of the total repeats found across different chromosomes.

		HG19			Venter		
Chromosome	**Chromosome size (Mb)**	**Observed repeats**	**Expected repeats**	**Chi-square**	**Observed repeats**	**Expected repeats**	**Chi-square**
1	249	72,306	78,166	439	67,980	71,632	186
2	237	73,515	74,399	11	70,414	68,180	73
3	192	56,100	60,273	289	54,872	55,234	2
4	183	56,029	57,447	35	54,614	52,645	74
5	174	53,215	54,622	36	51,359	50,056	34
6	165	50,801	51,797	19	49,919	47,467	127
7	153	54,365	48,030	836	52,411	44,015	1,602
8	135	45,473	42,379	226	44,178	38,836	735
9	132	38,534	41,437	203	34,282	37,973	359
10	132	44,387	41,437	210	42,564	37,973	555
11	132	40,687	41,437	14	40,115	37,973	121
12	123	43,560	38,612	634	42,407	35,384	1,394
13	108	29,387	33,903	602	28,856	31,069	158
14	105	27,552	32,962	888	26,381	30,206	484
15	99	25,186	31,078	1,117	22,893	28,480	1,096
16	84	33,498	26,369	1,927	32,063	24,165	2,581
17	81	31,050	25,427	1,243	29,581	23,302	1,692
18	75	23,571	23,544	0	23,170	21,576	118
19	69	30,512	21,660	3,617	29,263	19,850	4,464
20	63	21,540	19,777	157	20,895	18,124	424
21	54	13,393	16,952	747	12,353	15,535	652
22	57	15,255	17,893	389	14,650	16,398	186
X	141	49,696	44,263	667	11,433	40,563	20,919
Y	60	13,088	18,835	1,754	7,242	17,261	5,815
Total	**3,003**	**942,700**	**ΣChi-square**	**16,060**	**863,896**	**ΣChi-square**	**43,850**
			P value	**<0.0001**		**P value**	**<0.0001**

3.1.1 Expanded microsatellite loci in the human genome

To select microsatellites suitable for this study from the original pool of available tandem repeats, criteria that affect repeat instability were defined. It is known that pattern size, increased repeat length (copy number), and greater repeat purity influence repeat instability (Brinkmann *et al*., 1998). In this study, repeat pattern size was selected to be ≥ 2 and ≤10 bp, the repeat copy number was set to be ≥50 copies, and only pure repeats were considered (*i.e*. match 100%). The human genome reference sequence (HG19) and the J. Craig Venter genome sequence were searched to identify microsatellites of interest.

Initially the TRDB revealed that HG19 contains the larger number of tandem repeats, with 942,700 compared to 863,896 from the Venter reference sequence. Applying the selected criteria to the TRDB derived from both the HG19 and Venter reference sequences dramatically reduced the number of valid candidate microsatellites (Figure 3.1).

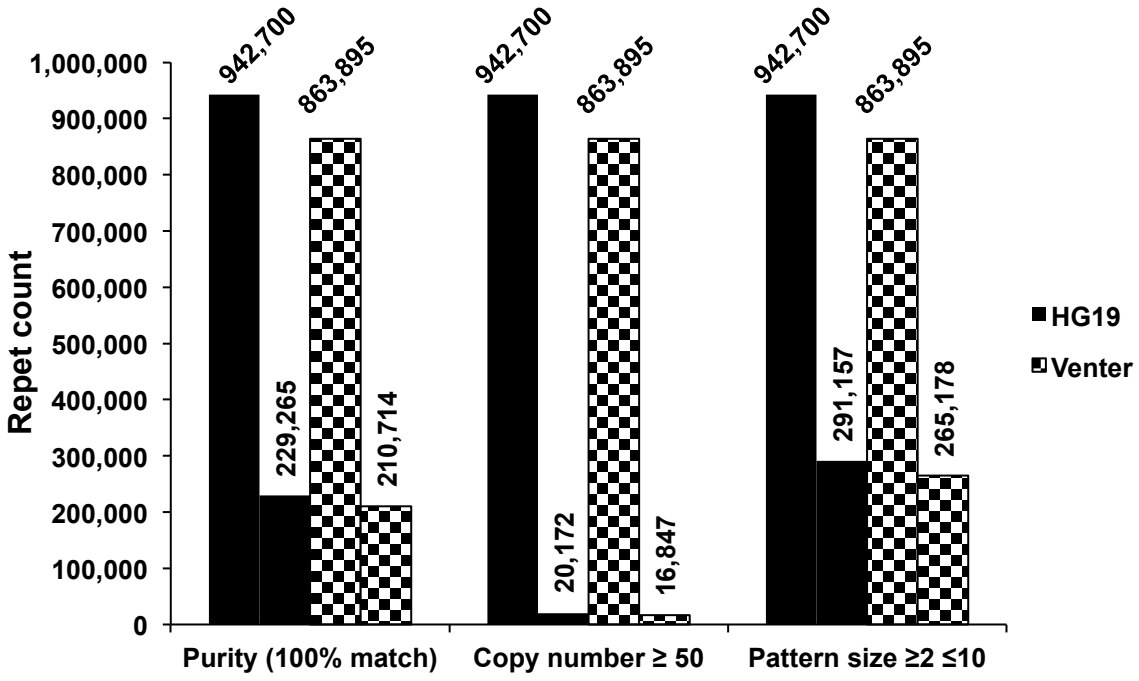


Figure 3.1: The effect of selected criteria on repeat count as identified using TRDB.

The values did not show any noticeable variation in different reference sequence with the set criteria. However, the result showed a variable decline in repeat count with each parameter used (Figure 3.1). Filtering by both repeat pattern size and repeat purity (100%) demonstrate a similar effect, with a reduction in eligible repeat counts of 70% and 75% respectively. This indicates that pure repeats (*i.e*. with no interruptions) represent only 25%

of the tandem repeats in both the HG19 and Venter reference sequence. The greatest effect on microsatellite number was observed when the copy number ≥50 filter was applied. Applying these criteria massively reduced the number of eligible repeats from 942,700 to 20,172 in the HG19 set and from 863, 895 to 16,847 in the Venter set, a reduction of 98% (Table 3.3).

Table 3.3: Eligible tandem repeats from the human genome using different selection criteria

Criteria	HG19	% of total	Venter 2007	% of total
Total number of repeats	**942,700**	100	**863,895**	100
Purity (100% match)	229,265	24.8	210,714	24.4
Copy number ≥ 50	20,172	2	16,847	2
Pattern size ≥ 2 ≤10	291,157	30.9	265,178	30.7
Pattern size ≥ 2 ≤10 + Purity (100% match)	101,209	10.7	93,195	10.8
Purity + Pattern size ≥ 2 ≤10 + Copy number ≥ 50	**15**	0.0016	**8**	0.001

After considering the effect of each individual parameter, the search criteria were applied collectively and the findings are described in Table 3.3. For HG19, the TRDB showed 942,700 microsatellites, of these 291,157 microsatellites exhibit pattern size ≥ 2 ≤10. The number of microsatellites decreases to 101,209 when 100% purity (no interruption) is used for further sorting. Finally, applying the copy number ≥ 50 filter has significant effect and results in only 15 microsatellites which meet all the parameters required for this study. Similarly, when the J. Craig Venter genome was analysed, TRDB identified 863,895 repeats of which 265,178 microsatellites contain a pattern size ≥ 2 ≤10. The number of microsatellites decreased to 93,195 when 100% purity (no interruption) was used for further sorting. The third filter of copy number ≥ 50 significantly reduced the final microsatellite count to a total of 8 microsatellites consistent with the criteria used. Overall, rather surprisingly, only 8 (Venter) and 15 (HG19) microsatellites were identified from an initial pool of 863,895 - 94,2700 tandem repeats, which fit with selection criteria (Figure 3.2). Interestingly, none of the microsatellites identified in the two human reference genomes overlapped, meaning that in total 23 different microsatellites were identified for further investigation.

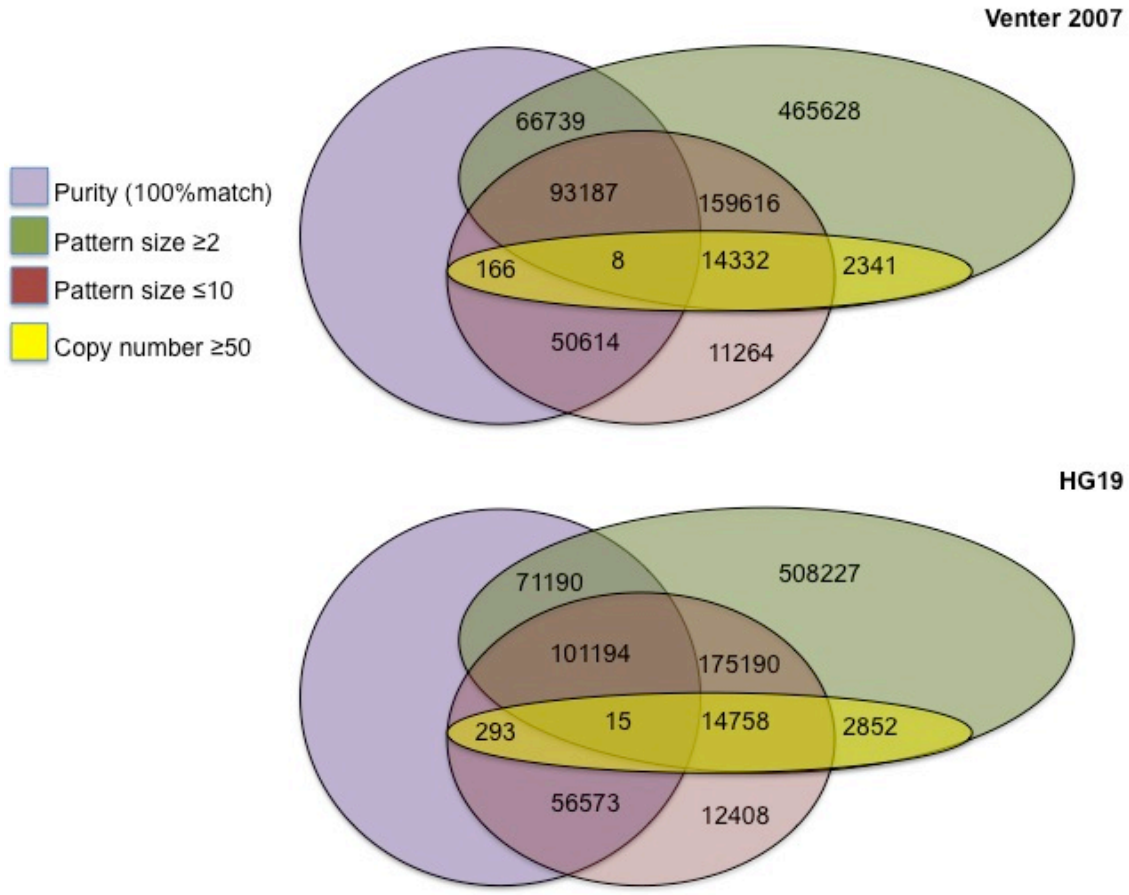


Figure 3.2: Venn diagram showing the effect of individual criteria on repeats count in different reference sequences as indicated.

3.1.1.1 Expanded microsatellites identified in TRDB

In the Venter data, 8 expanded microsatellites were obtained after applying the selection criteria (Table 3.5). Five di-nucleotide microsatellites were identified all composed of a CA pattern. Three out of 5 were located on chromosome 2, while the two others were located on chromosomes 15 and 20. Two microsatellites identified were penta-nucleotides and were located on chromosome 2 and chromosome X, with sequence patterns AGGAA and AGAAT respectively. Only one microsatellite identified was a tetra-nucleotide and was located on chromosome 14 with the pattern, TTTC. None of the 8 microsatellites identified from Venter database were tri or hexanucleotide.

In the other reference genome, HG19, where 15 microsatellites were identified, 6 microsatellites were di-nucleotides (located on chromosome 2, 7, 10, 11, 19, and X) with pattern AC, AT and TC (Table 3.5). Two microsatellites were penta and hexa-nucleotides are located on chromosome 21 and chromosome 1 with a TCCCT and CTCCCT pattern respectively. Furthermore, 3 microsatellites identified were tetra-nucleotides (located on

chromosomes 7, 16, and 17), and all have the TTTC pattern sequence. In contrast to Venter, HG19 has 3 microsatellites identified as tri-nucleotides (located on chromosomes 1, 5, and 13) with the GAA sequence pattern. Briefly, dinucleotide repeats are the most frequent microsatellites and the most common sequence is AC•GT. No pure GC dinucleotide or pure CAG•CTG were identified with copy number >50. For the sake of clarity, a marker name was assigned, which comprises chromosome number, repeat motif and the first 3 numbers of their genomic location (Table 3.5).

To further refine the analysis, expanded microsatellites were compared for their observed and expected values for their over/under representation. The expected expanded repeats were calculated by divided the total expanded repeat by total haploid genome size of human in Mbp, the product subsequently was multiplied by the size of the respective chromosome. The calculated expected expanded microsatellite values were ≤ 5 and Chi-square analysis cannot be used to find how significant the deviation between observed and expected expanded microsatellites. Therefore, Poisson distribution was conducted. Under the Poisson distribution, the expected number of expanded microsatellites represent the average number of microsatellites in a given chromosome size and used to predict the probability of seeing the observed expanded microsatellites in that given average. For most cases, the proportional were very high ranging between 27% and 70% indicated we would observe these numbers of expanded microsatellites giving this average frequency based on chromosome size. However, the data showed that expanded repeats are overestimated in chromosome 2 with relatively low probability of frequency 2.7% (Table 3.4). With reference to chromosome 2, irrespective of the reference sequences examined, compared to expect expanded repeats, 1.8, the observed number of repeats was found to be 5 (Table 3.4). Over representation of expanded microsatellite could possibly be due to the multiple mapping of same microsatellite particularly in chromosome 2. To investigate this further a dot plot of each microsatellite with each other microsatellite was generated Figure 3.3). A window size of 9 bp and similarity of 70% was selected to minimize the noise. The data show that these microsatellites are different and showing low sequence similarity (Figure 3.3).

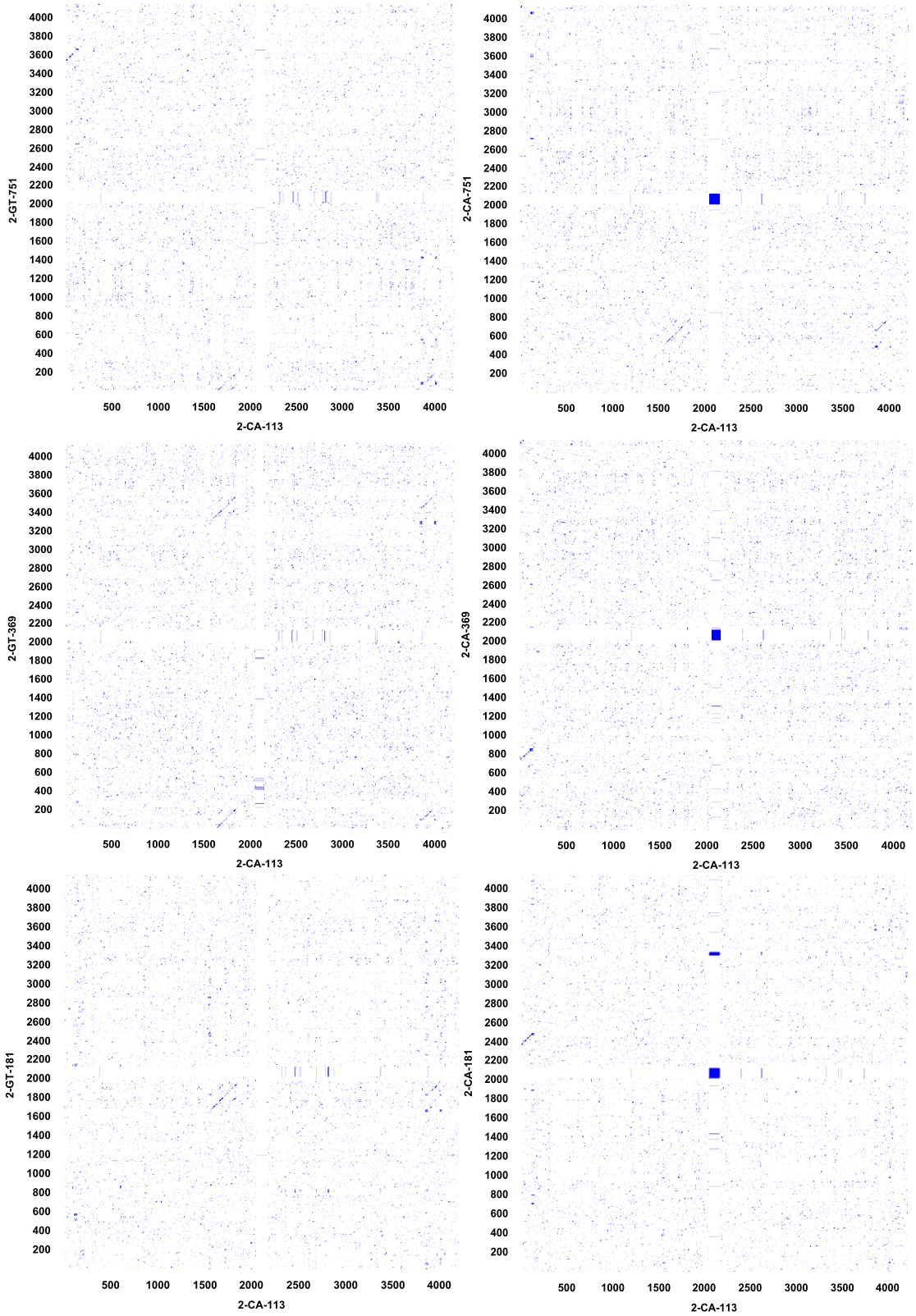


Figure 3.3: Dot plot analysis showing the sequence comparison of two dinucleotides as indicated. The graph indicates low sequence similarity. The central dark blue square represents the presence of identical tandem repeats.

Table 3.4: Comparison of the expected and observed values of the expanded repeats found across different chromosomes.

Chromosome	Chromosome size (Mb)	Observed expanded microsatellite	Expected expanded microsatellite	Poisson distribution
1	249	2	1.91	0.27
2	237	5	1.82	0.027
3	192	0	1.47	0.23
4	183	0	1.40	0.247
5	174	1	1.33	0.352
6	165	0	1.26	0.284
7	153	2	1.17	0.212
8	135	0	1.03	0.357
9	132	0	1.01	0.364
10	132	1	1.01	0.368
11	132	1	1.01	0.368
12	123	0	0.94	0.391
13	108	1	0.83	0.362
14	105	1	0.80	0.359
15	99	1	0.76	0.355
16	84	1	0.64	0.337
17	81	1	0.62	0.334
18	75	0	0.57	0.566
19	69	2	0.53	0.083
20	63	1	0.48	0.297
21	54	1	0.41	0.272
22	57	0	0.44	0.644
X	141	2	1.08	0.198
Y	60	0	0.46	0.631
Total	**3003**	**23**		

Table 3.5: Expanded microsatellites identified in TRDB

No.	Name	Chromosome location	Indices	Pattern	Copy number	TRDB
1	19-AT-431	19q13.1	43167387--43167883	AT	248	H*
2	X-AGAAT-759	Xp22.3	7592280--7592828	AGAAT	109	V*
3	17-TCTT-376	17q12	37655127--37655460	TCTT	83	H
4	14-TTTC-814	14q33	81443573--81443835	TTTC	65	V
5	7-AT-232	7p15	23271418--23271541	AT	62	H
6	2-CA-369	2p25	3698911--3699032	GT	61	H
7	21-TCCCT-409	21q22	40955745--40956050	TCCCT	61	H
8	10-CA-994	10q24	99438638--99438757	GT	60	H
9	1-TTC-102	1p13	102123411-102123592	TTC	60	H
10	15-CA-519	15q22	51983268--51983383	GT	58	V
11	19-CA-424	19p13.3	4247580--4247695	GT	58	H
12	5-AGA-126	5q23	126583011--126583179	AGA	56	H
13	2-CA-751	2p12	75118984--75119094	CA	55	V
14	X-AT-535	Xp11.2	53513547--53513656	AT	55	H
15	2-CA-113	2p15	113760073--113760180	GT	54	V
16	20-CA-417	20q13.1	41780725--41780831	GT	53	V
17	7-TTTC-132	7p21	13242596--13242810	TTTC	53	H
18	16-TTTC-505	16q12.1	50509578--50509792	TTTC	53	H
19	11-TC-107	11q23	107461059--107461165	TC	53	H
20	1-CTCCCT-151	1q21	151571757--151572077	CTCCCT	53	H
21	2-AGGAA-603	2p15	60357664--60357926	AGGAA	52	V
22	2-CA-181	2q32	181818655--181818759	AC	52	V
23	13-AAG-102	13q33	102813925--102814076	AAG	50	H

*H:HG19 *V:Venter

3.1.1.2 Human BLAST search

DNA sequences of the selected 23 microsatellites were retrieved from TRDB. Each DNA sequence containing the microsatellite repetitive sequence region and 2,000 bp of flanking sequence from both sides of the repeat were obtained. Then, human genome BLAST searches (http://blast.ncbi.nlm.nih.gov/Blast.cgi) were performed to confirm the TRDB findings. Obtained microsatellite DNA sequences (query sequence) were investigated against publicly available human genomic sequences. The BLAST search results returned the best possible alignments that include both our query and database sequences (Figure 3.4).

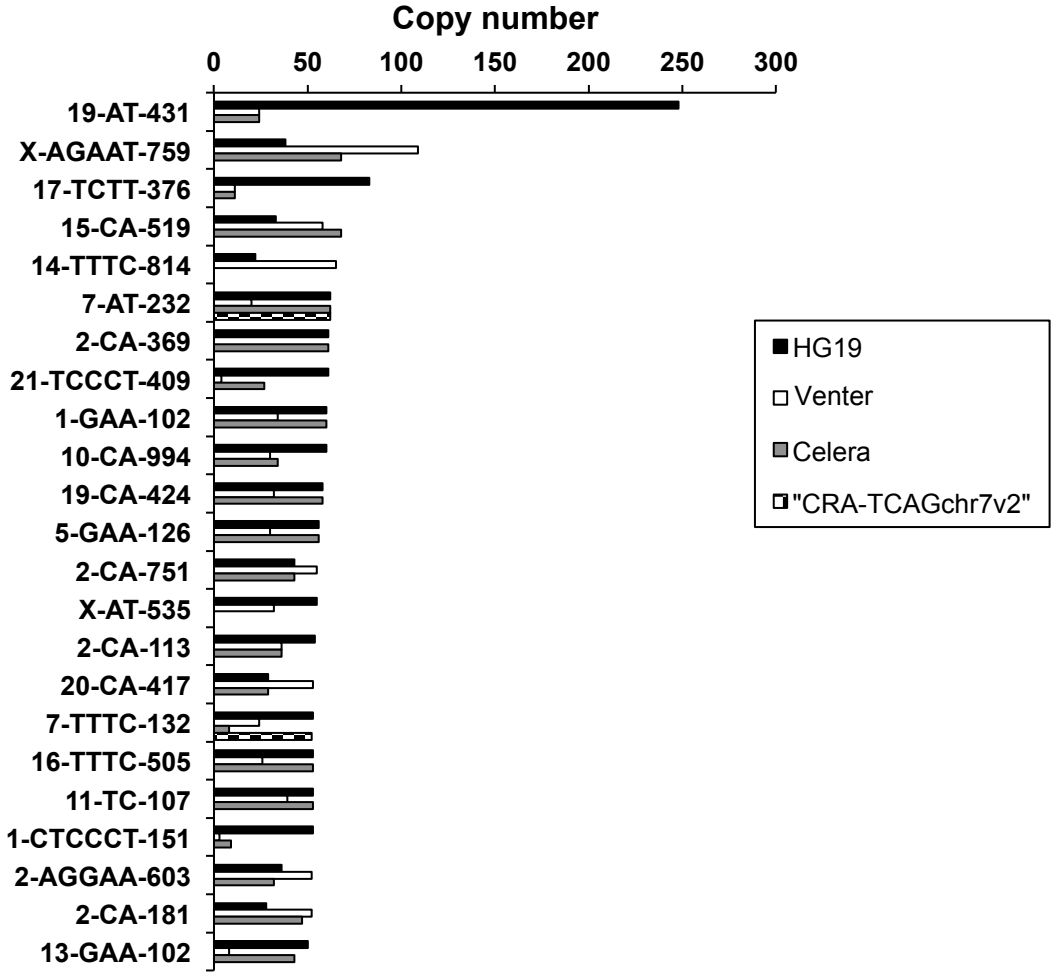


Figure 3.4: Human BLAST search for the expanded 23 microsatellites identified in TRDB

Where more than one hit were shown the search engine picked the query sequence and different copy number versions for the same pattern size which may result from general population variation within the reference sequences, or due to variation in sequencing length for the same DNA region. The majority of microsatellites have 3 hits. The human

BLAST search indicated these microsatellites are probably variable among the population and also confirmed none of the microsatellites identified in the two human reference genomes sequence overlapped. 19-AT-431 locus has large copy number compared to other expanded loci identified. This observation indicated that 19-AT-431 locus located in AT rich genomic region (AT blocks) or just due to misalignment process when mapped to the HG19. It also located in chromosome 19, which is gene dense, and Alu repeats prefer such regions.

3.1.2 Investigation of the repeat length variation in the general population

TRDB and the BLAST search identified 23 microsatellites with predicted repeat length variation for further investigation. Repeat length variation for these candidate microsatellites next needed to be confirmed in the laboratory before conducting any further analysis. Repeat length variation (copy number) was determined using bulk PCR. No variation in the population suggests that there is very low germ line mutation frequency. Nevertheless there is a strong association between germ line mutation and somatic variability (Nestor and Monckton, 2011). RepeatMasker and CLC genomic workbench were used to design primers to amplify these sequences. Then, gradient PCR was performed to identify the best annealing temperature for each microsatellite.

In order to investigate the repeat length variation in the general population, 10 human DNA samples were selected randomly from 112 DNA samples collected for this project and bulk PCR analysis was conducted for the 23 microsatellites identified to fit our stringent search criteria. Using multiple pairs of primers, only X-AT-535 and 2-CA-181 failed to generate PCR products. In addition, 13-AAG-102 181, could only be amplified in some of the samples tested and is described in detail latter.

The results obtained from 1.5% agarose gel electrophoresis for the remaining 20 microsatellites confirmed the presence of repeat length variation among the samples tested (Figure 3.5). As a result, these markers could be taken forward to the next stage of investigation, to determine the level of somatic instability in the general population.

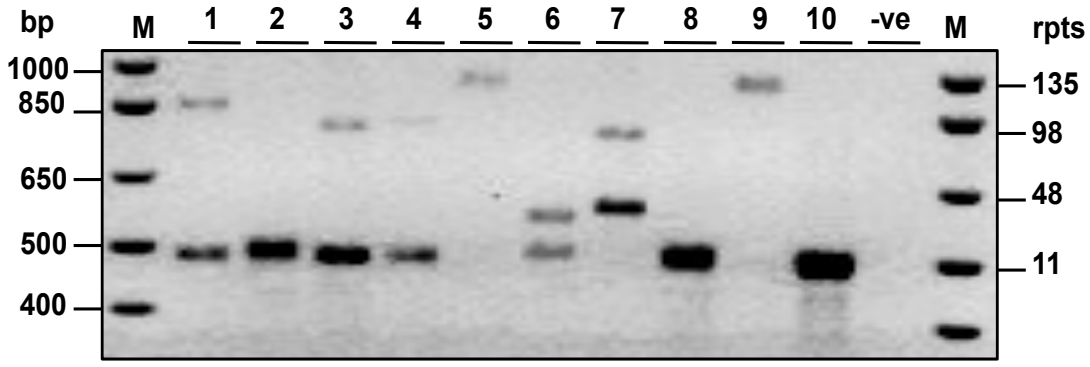


Figure 3.5: The investigation of allele length variation in 7-TTTC-132 using bulk PCR. Only 10 samples were tested at this stage. The negative (-ve) lane represents a negative control (no DNA in the reaction mix) while M lane (DNA MW marker).

In a pilot study, twenty microsatellites showed allele length variation in the general population. In order to investigate the degree of somatic instability, 10 ng of 48 human blood DNA samples with an age >40 years were used to identify individuals with an allele length of > 50 repeats. Similarly, bulk PCR was carried out and the PCR products generated were resolved by 1.5% agarose gel electrophoresis and visualised using ethidium bromide staining. The degree of allele length variation was determined (genotype) for all samples that generated PCR products. All of the loci tested showed observable allele length variation with different levels of heterozygosity.

Although each loci tested met the selection criteria in TRDB of >50 repeats, the results showed that four loci (19-AT-431, 7-AT-232, 16-TTTC-505, and 1-CTCCCT-151) identified no individuals tested with expanded allele length of > 50 repeats. An example of the amplified repeat sizes for one of these markers, 16-TTTC.505, is given in Figure 3.6. Out of the remaining 16 loci, four (19-CA-424, 20-CA-417, 15-CA-519, and 10-CA-994) have shown expansion in the repeat length at low frequency (~10%), whereas the remaining 12 frequently showed expansion in the samples examined (Figure 3.10). Both 2-AGGAA-603 and 5-AGA-126 showed high levels of allele length variations (polymorphism) suggesting their potential suitability as DNA profiling markers (Figure 3.7).

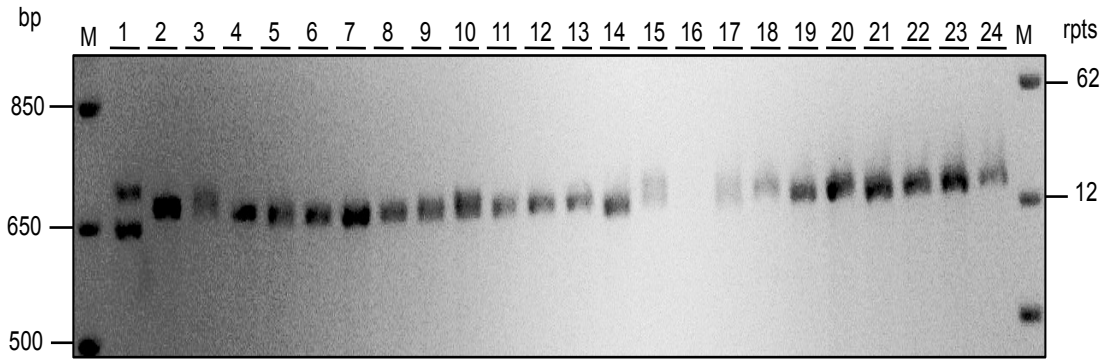


Figure 3.6: 16-TTTC.505 loci show no expanded allele (> 50 repeats) in all samples tested despite their selection criteria (only 24 samples shown).

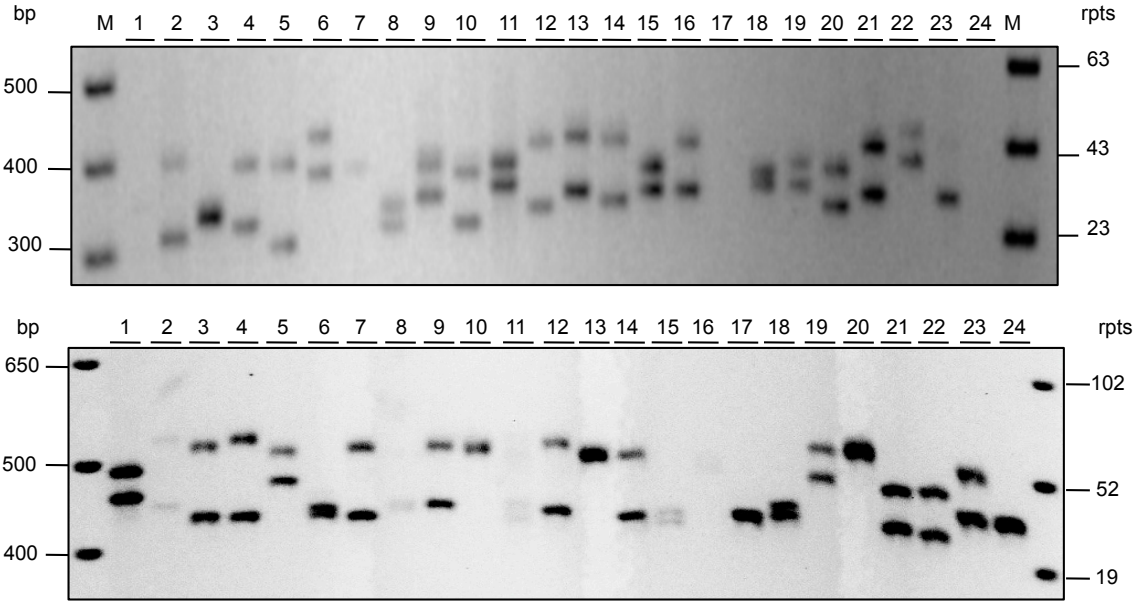


Figure 3.7: Highly polymorphic loci. A)- 2-AGGAA-603 showed high level of repeat length variations with observed 87% heterozygosity (only 24 samples shown). B) 5-AGA-126 showed high degree of repeat length variations with observed 72% heterozygosity (only 24 samples shown).

All samples generating PCR products were genotyped and the allele frequency estimated (Figure 3.10). Further analysis was conducted to estimate the number of alleles and the degree of heterozygosity (Table 3.6). 10-CA-994, 20-CA-417 and 13-AAG-102 showed alleles with length > 100 repeats with a range 35-102; 24-124 and 14-110 repeats, respectively. The highest number of estimated alleles was 47 in 20-CA-417 whereas only 9 alleles were observed in 19-AT-431. The highest heterozygosity was observed in 2-AGGAA-603 with 87% (Table 3.6).

As it is known that longer alleles undergo more mutations, and a positive correlation is expected between the mean allele length and heterozygosity. Using linear regression the relationship between mean allele length and the observed heterozygosity was estimated, which indicates no significant correlation ($r^2 = 0.17$ and P value = 0.064) between the two

compared variables. However, an ascending trend was observed (Figure 3.8). The linear regression analysis was also conducted to the estimate relationship between mean allele length and the estimated number of alleles. The result showed significant correlation (r^2 = 0.30 and P value = 0.01) (Figure 3.8). Similarly, the relationship between the estimated number of alleles and the observed heterozygosity was estimated and the linear regression result showed highly significant correlation (r^2 = 0.46 and P value = 0.0008) (Figure 3.9).

GC content of the flanking region (500 bp on either side) of the loci was calculated using Oligo calculator (http://www.basic.northwestern.edu/biotools/OligoCalc.html) (Kibbe, 2007). The highest GC content was found in 2-CA-369 (58%) whereas the lowest GC content was observed in 2-CA-751 (33%). GC content of most of the expanded microsatellite-flanking regions were ranging from 40% to 50% (Table 3.6).

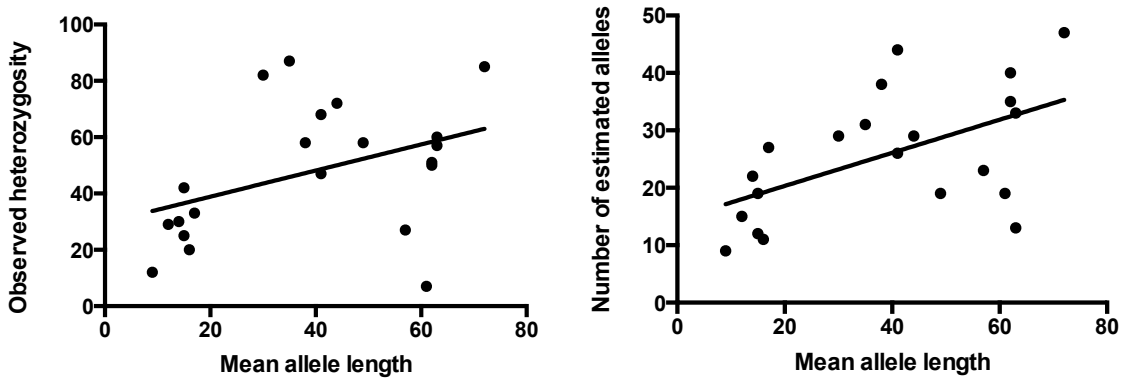


Figure 3.8: Linear regression analysis between the mean allele length and observed heterozygosity and the mean allele length and number of alleles.

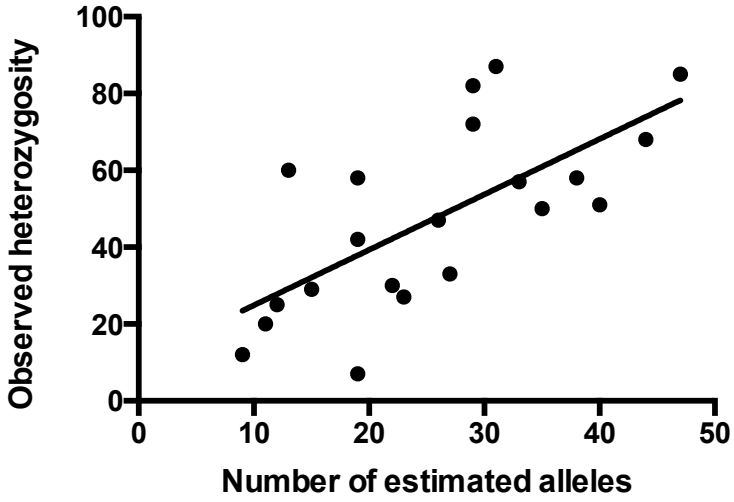


Figure 3.9: Using linear regression to estimate the relationship between estimated number of alleles and the observed heterozygosity.

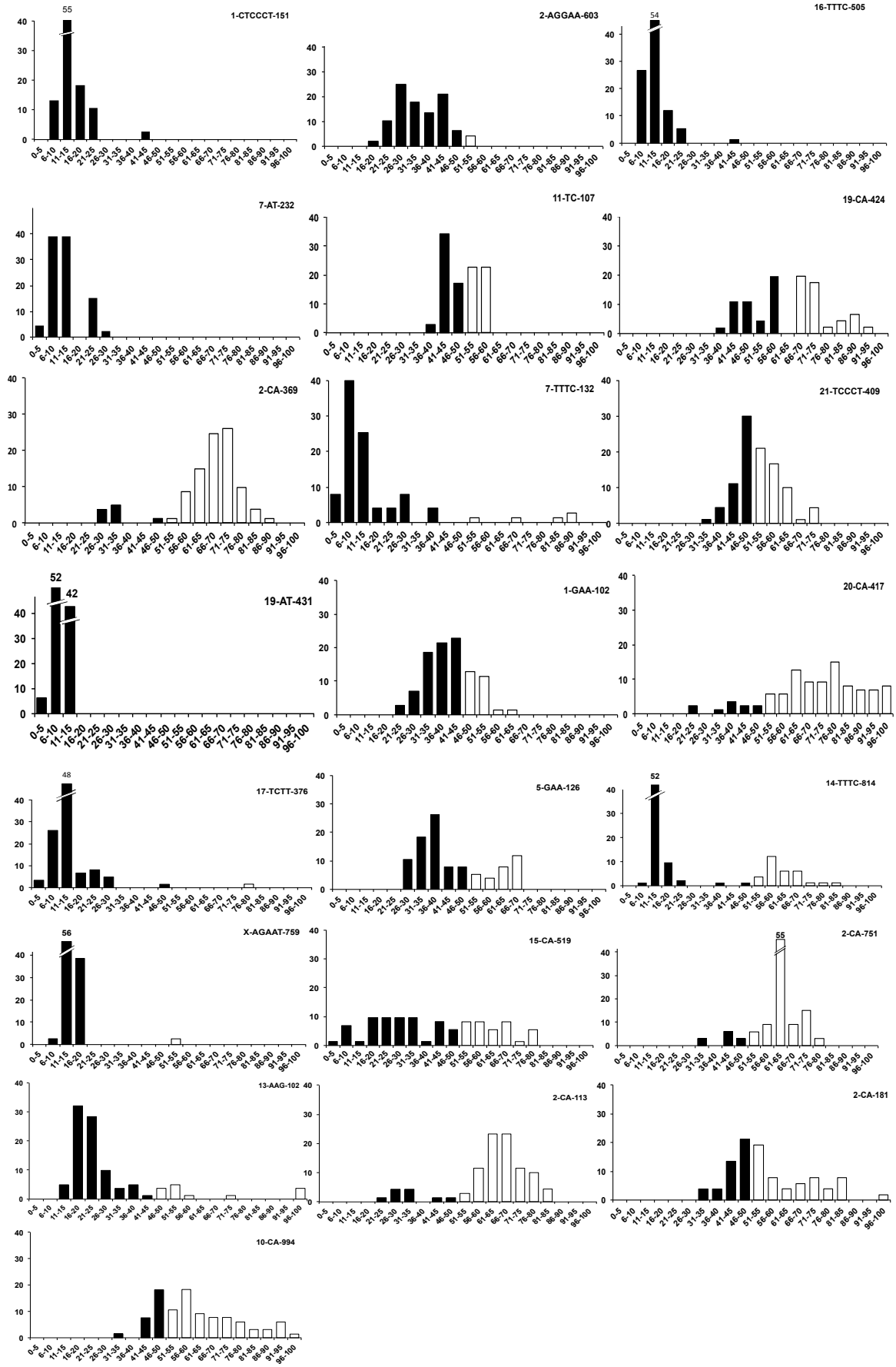


Figure 3.10: The allele frequency distribution. All the samples were genotyped and allele frequency (%) was calculated for each loci. The Y-axis showed allele frequency and the X-axis the number of repeats. Black bars represent alleles with repeat <50 and white bars alleles with repeat >50.

Table 3.6: Genotyping of microsatellites

NAME	Chromosome location	Pattern	Observed Heterozygosity	Estimated no. alleles	Repeat range	Mean allele repeat length	%GC (500 bp flanking)	TRDB
19-AT-431	19q13.1	AT	12%	9	3-12	9	44%	H
X-AGAAT-759	Xp22.3	AGAAT	20%	11	10-55	16	45%	V
17-TCTT-376	17q12	TCTT	42%	19	5-79	15	49%	H
14-TTTC-814	14q33	TTTC	82%	29	9-82	30	47%	V
7-AT-232	7p15	AT	29%	15	4-27	12	41%	H
2-CA-369	2p25	GT	51%	40	14-89	62	58%	H
21-TCCCT-409	21q22	TCCCT	N/A	27	35-73	52	36%	H
10-CA-994	10q24	GT	50%	35	35-102	62	50%	H
1-TTC-102	1p13	TTC	47%	26	22-62	41	41%	H
15-CA-519	15q22	GT	68%	44	4-80	41	41%	V
19-CA-424	19p13.3	GT	60%	13	40-92	63	54%	H
5-AGA-126	5q23	AGA	72%	29	27-69	44	33%	H
2-CA-751	2p12	CA	7%	19	35-76	61	53%	V
X-AT-535	Xp11.2	AT		No amplification				H
2-CA-113	2p15	GT	57%	33	25-85	63	52%	V
20-CA-417	20q13.1	GT	85%	47	24-124	72	49%	V
7-TTTC-132	7p21	TTTC	33%	27	5-88	17	38%	H
16-TTTC-505	16q12.1	TTTC	30%	22	8-44	14	49%	H
11-TC-107	11q23	TC	58%	19	40-59	49	49%	H
1-CTCCCT-151	1q21	CTCCCT	25%	12	10-45	15	53%	H
2-AGGAA-603	2p15	AGGAA	87%	31	19-53	35	51%	V
2-CA-181	2q32	AC	27%	23	31-98	57	41%	V
13-AAG-102	13q33	AAG	58%	38	14-110	38	36%	H

No significant correlation was observed between GC content of the flanking regions and the observed heterozygosity (Figure 3.11). The linear regression analysis showed r^2 = 0.00017 and P value = 0.96.

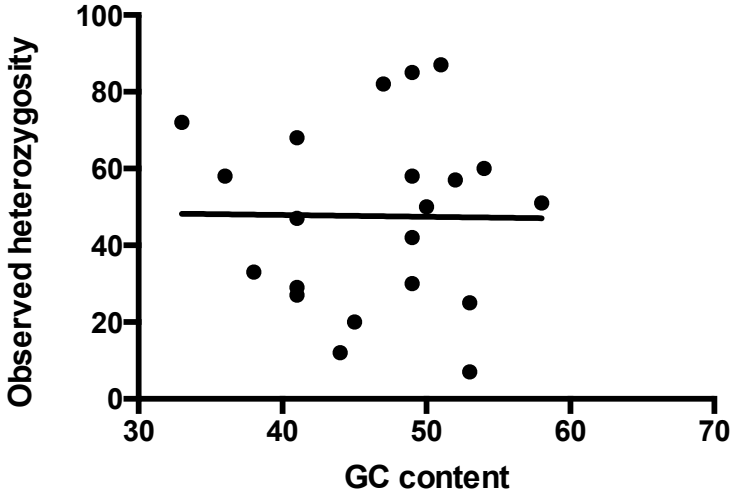


Figure 3.11: Linear regression analysis between GC content of the flanking region and observed heterozygosity.

3.1.2.1 Using small-pool PCR to investigate the level of somatic instability

DNA samples from individuals that successfully amplified a PCR product with a repeat copy number ≥ 50 at a microsatellite locus were further tested using small-pool PCR (SP-PCR). SP-PCR is a technique thorough which levels of somatic instability can be determined (Gomes-Pereira *et al.*, 2004). SP-PCR was conducted at a range of input DNA concentrations equivalent (g.e) to 100 g.e, 50 g.e, 10 g.e, and 1 g.e. The PCR products were resolved by agarose gel electrophoresis and detected by Southern blot hybridisation with a locus-specific probe.

Somatic instability was measured by sizing individual alleles. In all microsatellites tested, the estimated repeat sizes showed no or little variation within all the samples tested suggesting no or very low level of somatic instability (Figure 3.12). However, some samples showed a low level of somatic instability (Figure 3.13).

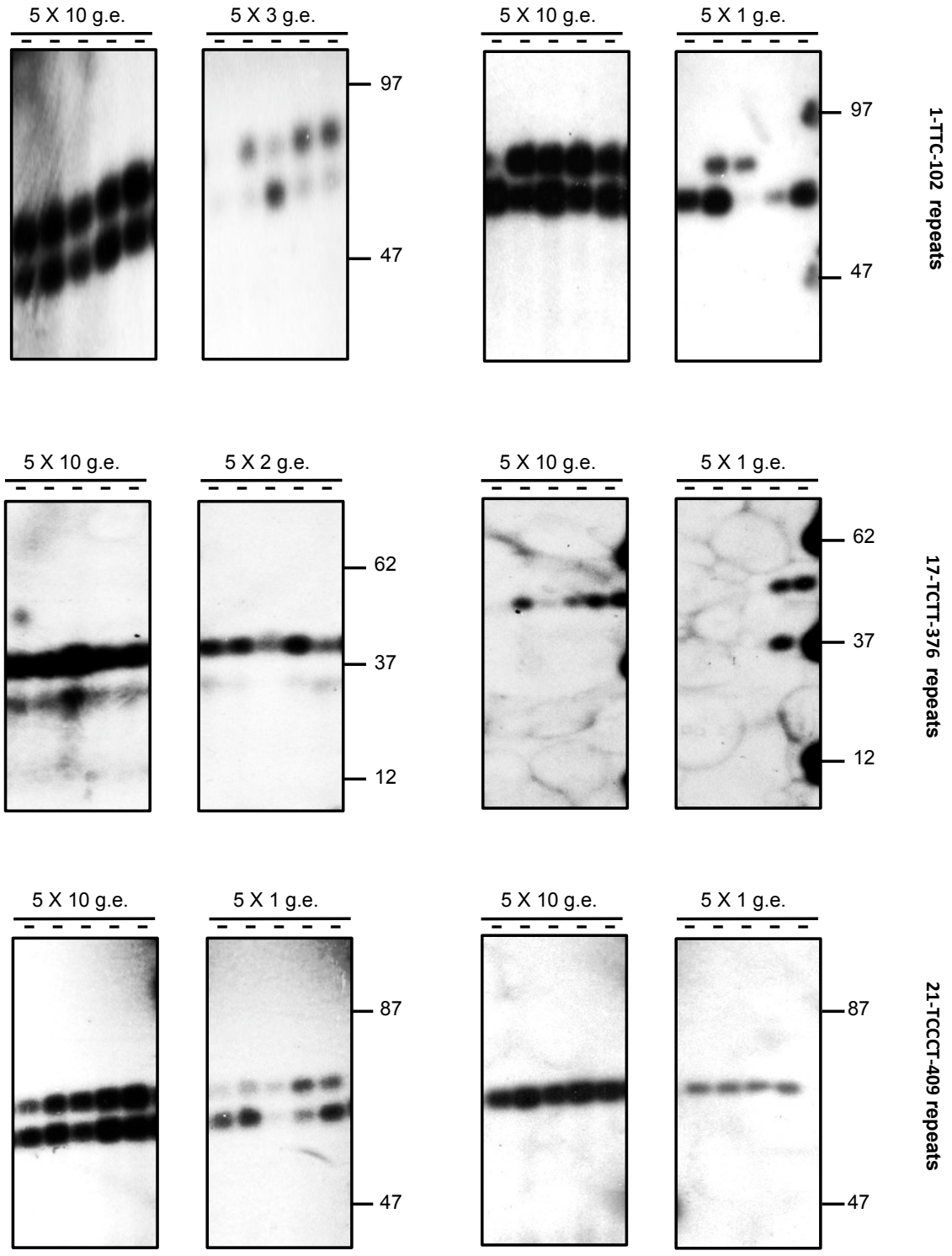


Figure 3.12: SP-PCR analysis: The result showed no or very low level somatic instability.

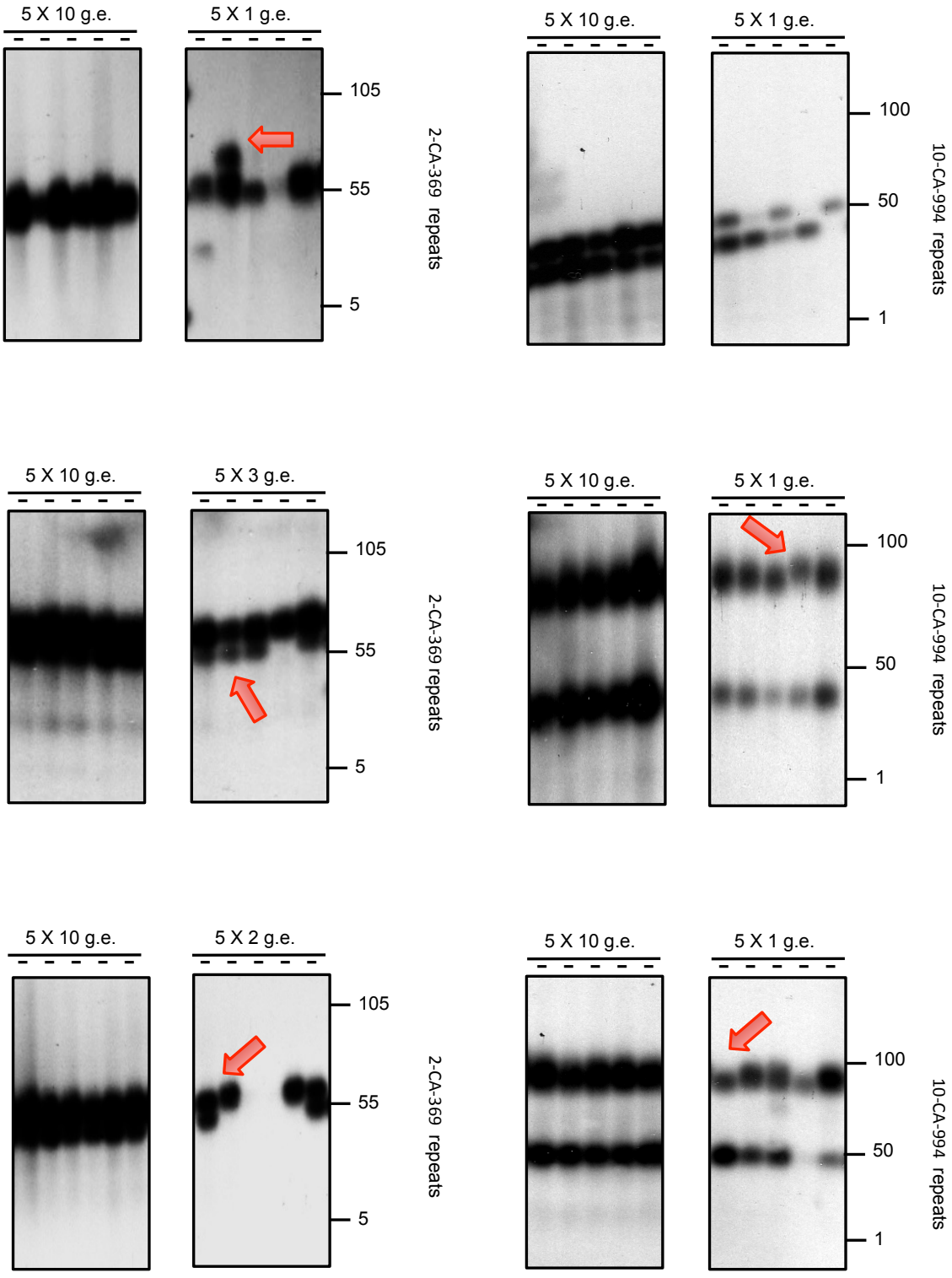


Figure 3.13:SP-PCR analysis. 10-CA-994 and 2-CA-369 loci show low level of somatic instability. The red arrows represent the mutation rendering to the difference in the relative migration of some repeat loci between different individuals

3.1.2.2 The 13-AAG-102 expansion is accompanied by a flanking deletion

The human genome reference sequence was used to select the first set of internal flanking primers (I) for 13-AAG-102 microsatellite. However, only (19/48) samples generated detectable PCR products (Figure 3.14) showing a large allele length range, from 40 to 150 repeats. Most individuals amplified were homozygous for different length alleles.

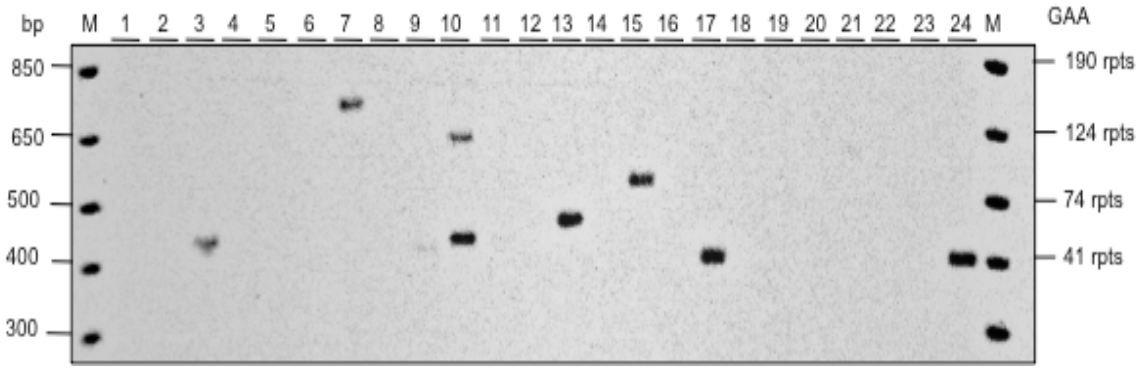
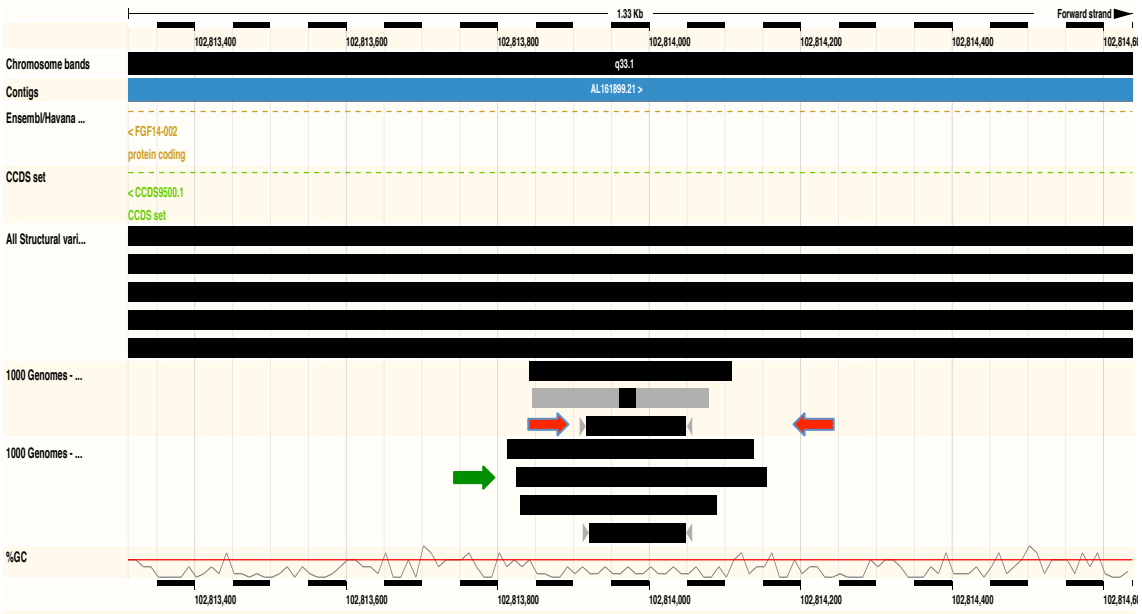


Figure 3.14: PCR products for 13-AAG-102. Only limited number of samples showed amplification using internal primers (I) located near the repeat sites showing the samples were taken from a homozygous individual for the respective allele.

The results implied the presence of a structural variant in the flanking sequences. Alignments using human genome sequence data from the 1,000-genomes Project revealed the presence of 7 different putative deletions spanning up to 113 bp into the 5′-flanking DNA and up to 79 bp into the 3′-flanking DNA (Figure 3.15).



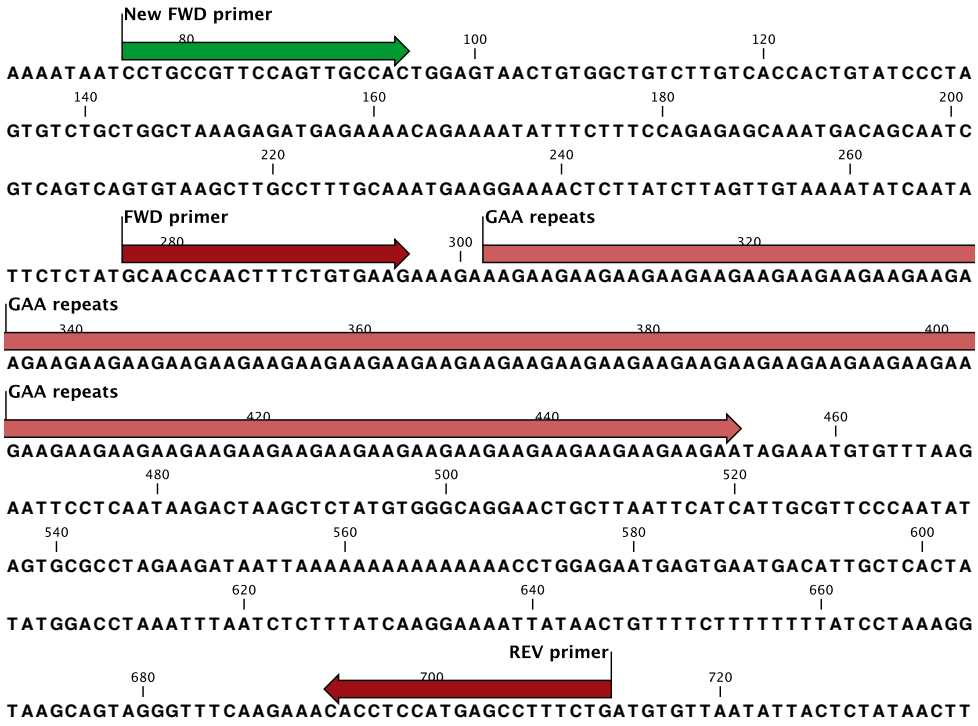


Figure 3.15: Deletion spans exist at 13-AAG-102 flanking regions as indicated using 1,000-genomes Project (Short black or grey bars). The red (I) and green (O) arrows represent the position of the primers used to amplify the respective genomic region. The latter was specifically used in case of red primer fails to show amplification due to the deletion of its complementary (binding) region.

Thus, a new forward primer (O) was designed distant from the putative deletions (Figure 3.15). From this all samples showed detectable PCR products and found to be mostly heterozygous (Figure 3.16).

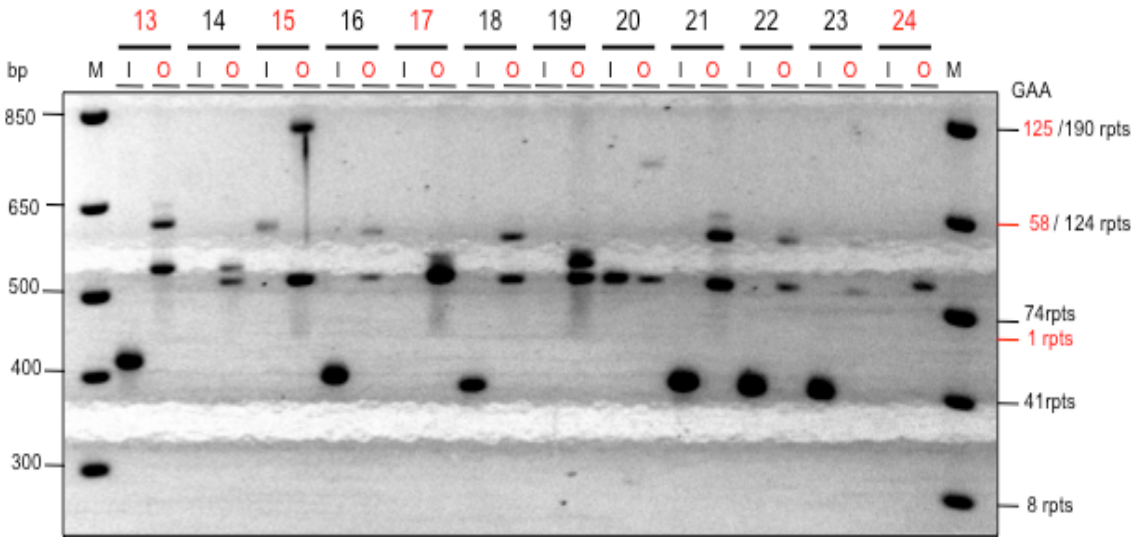


Figure 3.16: Using both I and O primers to confirm the existence of 13-AAG-102 structural variant.

However, the amplified fragment size did not match the deletion span as predicted by 1,000-genome database. Therefore, sequencing of DNA fragments was performed to

determine the deletion size. The analysis revealed the presence of two alleles. One allele is similar to the human reference genome sequence. In this sequence, the repeat sequence is lead by a 9 bp sequence GAAGAAAGA. In the other allele, the 9 bp sequence was substituted by a 15 bp sequence preceding the AAG repeats (Figure 3.17).

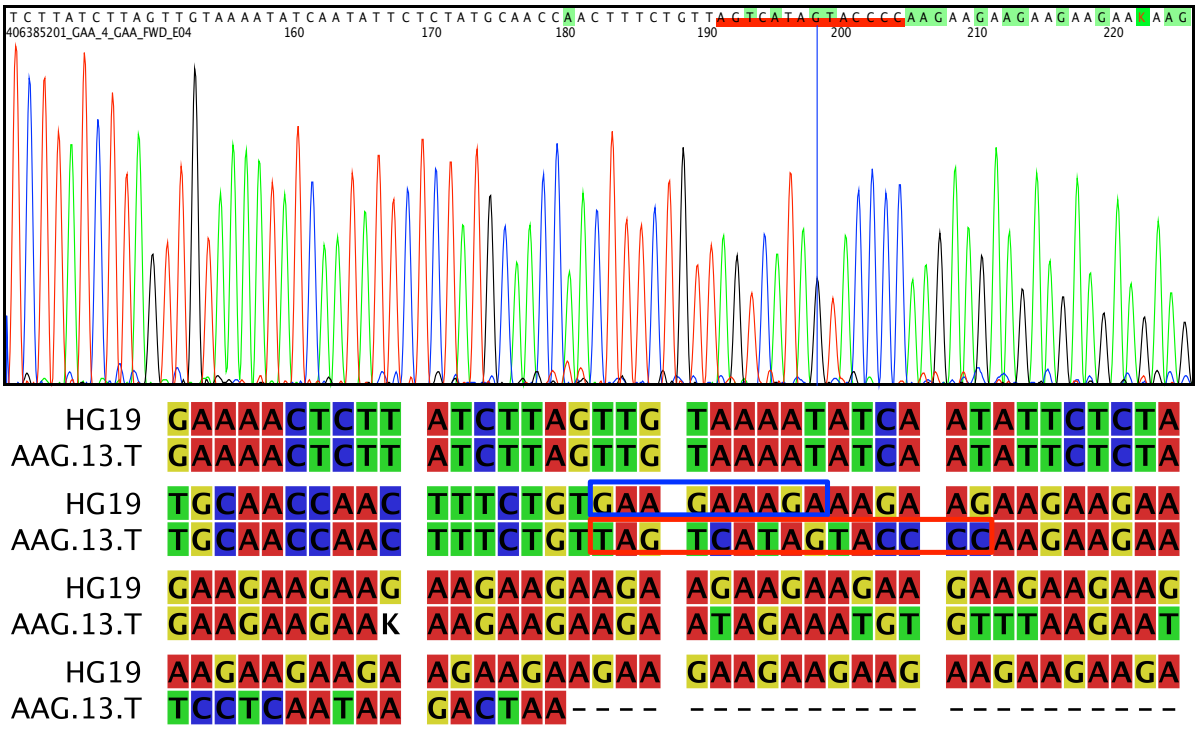


Figure 3.17: 13-AAG-102 structural variant. The 9 bp sequence (blue box) was substituted by 15 bp sequence (red box) preceding the AAG repeats. Key: HG19 (human genome reference sequence of the 13-AAG-102 repeat along with flanking region); AAG.13.T (DNA sequence of the corresponding region in the sample, where structural variation is observed.

In order to confirm this finding a new specific primer (ins) was designed to include the 15 bp sequence previously identified. The specificity meant that only those alleles containing this sequence would be amplified (Figure 3.18).

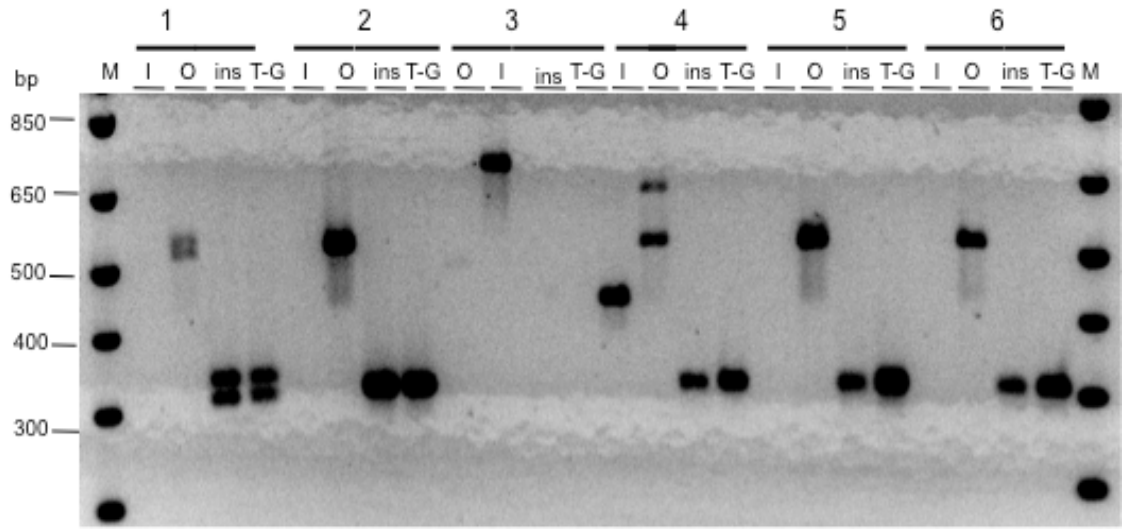


Figure 3.18: Amplification of 13-AAG-102 using ins primer.

As expected, all alleles which amplified with primer (O) but not with primer (I) were successfully amplified using new specific primers (ins). No deletion was observed in the amplified region as suggested by the 1,000-genome project. Rather it seems to be a subtle sequence change at the complementary binding site of the forward primer.

Using data obtained from primers (O) and primers (I), the allele distribution was calculated (Figure 3.19). Interestingly, each flanking allele has a different repeat length distribution. All expanded alleles are linked to the absence of the 15 bp sequence whereas a less variable allele distribution was observed when the DNA flanking sequence includes the 15 bp sequence.

Figure 3.19: Allele frequency distribution of 13-AAG-102 locus

Sequence alignment of 13-AAG-102 microsatellite region of human, chimpanzee and gorilla showed that the AAG microsatellite (at least 5 pure repeat in gorilla) is present in the common ancestor of great apes (Figure 3.20). Additionally, the data also indicate that human allele1 sequence is a derivative of a 15 bp deletion of the 5′-flanking DNA and a rearrangement of the 5′-end of the AAG microsatellite, which yielded the G/A rich 9 bp sequence, located upstream to the AAG microsatellite Figure 3.20).

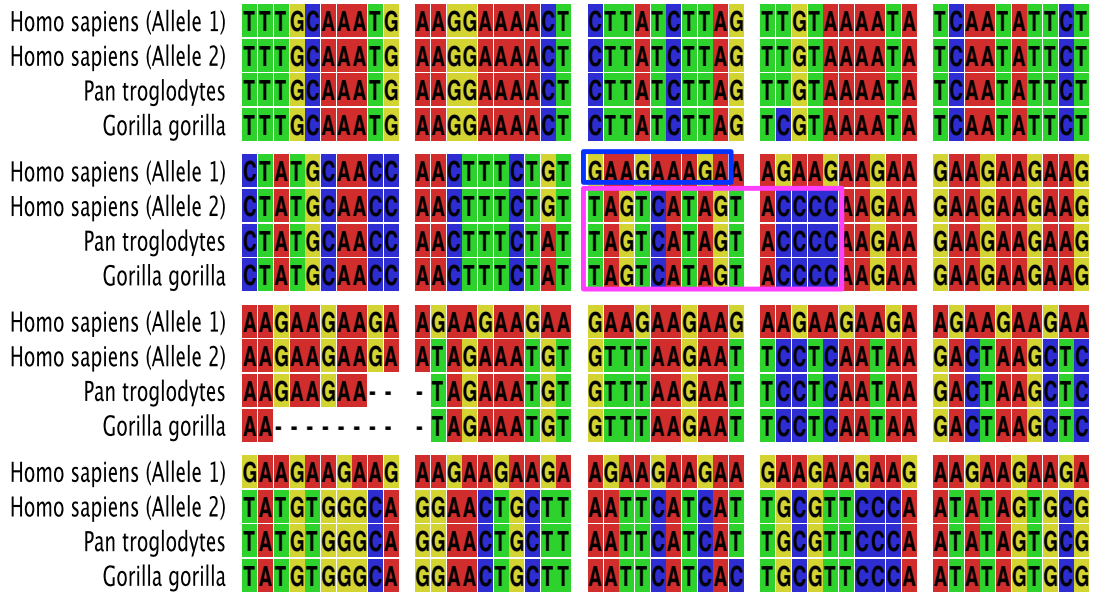


Figure 3.20: DNA sequence alignment of the 13-AAG-102 microsatellite region of human and two other greater apes as indicated. The human reference (allele 1) contains the G/A rich 9bp sequence (blue box). The human (allele 2), chimpanzee and gorilla references contain the 15 bp DNA sequence (pink box).

3.1.2.3 The 2-CA-181 is flanked by a recent polymorphic Alu insertion

The 2-CA-181 microsatellite was amplified using flanking primers based on the human genome reference. No PCR products were observed in all samples tested. Similarly, analysis of the human genome sequence data from the 1,000-genomes Project was used to identify variants that may exist in the flanking sequence of the 2-CA-181 microsatellite (Figure 3.21). The 1,000-genome data revealed the presence of 6 putative structural variants (insertion/deletions) in the 3′- flanking region of the 2-CA-181 microsatellite. These data were used to generate primers distal from these variants.

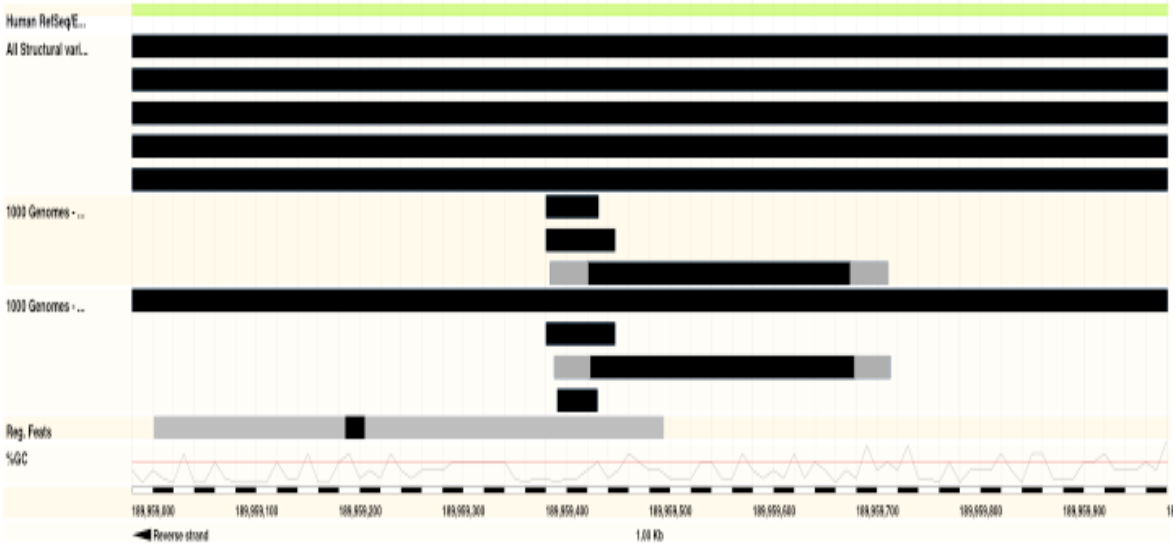


Figure 3.21: Structural variant of 2-CA-181 flanking sequence. No PCR products were observed in all samples. The 1000-genome data revealed the presence of 6 putative structural variants (insertion/deletions) in the 3'- flanking region of the 2-CA-181 microsatellite

Using a new primer successful amplification in PCR was observed (Figure 3.22). However, some of the alleles amplified were smaller than expected (677 bp) (Figure 3.22).

Eight of the smaller alleles were sequenced to identify the structural variant located in 2-CA-181 flanking sequences (Figure 3.23). The sequencing data revealed the existence of a 310 bp deletion in the very short alleles. The BLAST analysis of the 310 bp deleted region revealed it to be an Alu repeat. Correcting for the Alu insertion/deletion polymorphism, all samples were genotyped for the number of repeats and haplotyped with respect to the presence/absence of the Alu repeat. Unlike the 13-AAG-102 microsatellite, the microsatellite allele length distribution between presence/absence of the Alu repeat overlapped (Figure 3.24).

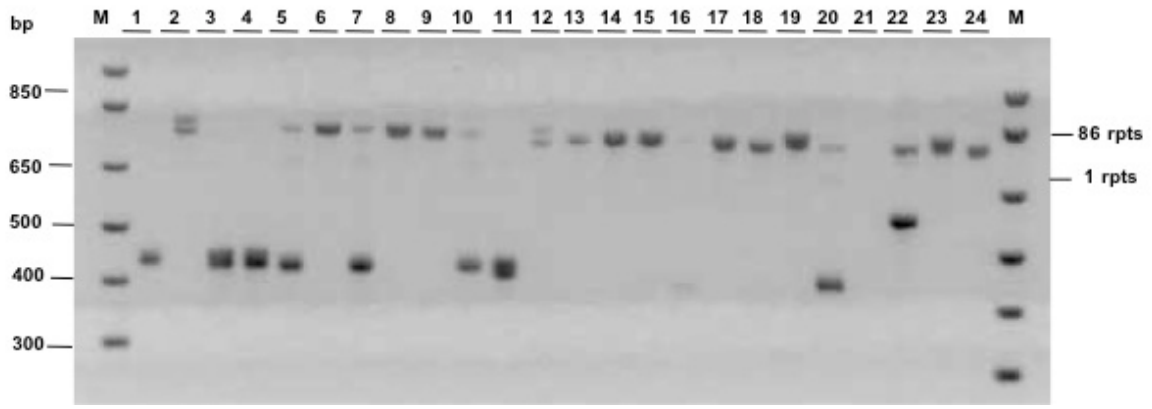


Figure 3.22: Bulk PCR using primers distal from structural variants of 2-CA-181-locus.

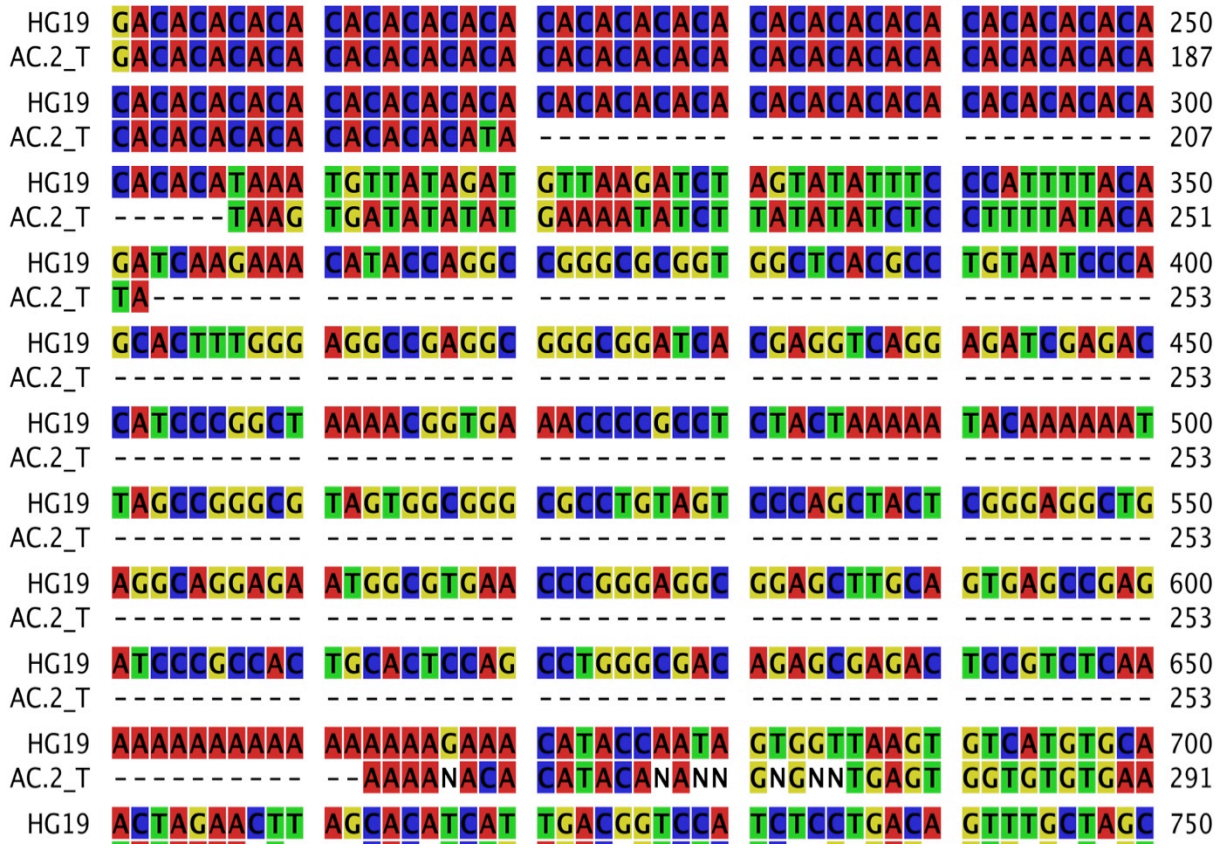


Figure 3.23: 2-CA-181 sequencing showing a 310 bp insertion/deletion. Key: HG19 (human genome reference sequence of the 2-CA-181 repeat along with flanking region); AC.2_T (DNA sequence of the corresponding region in the sample, where structural variation is observed.

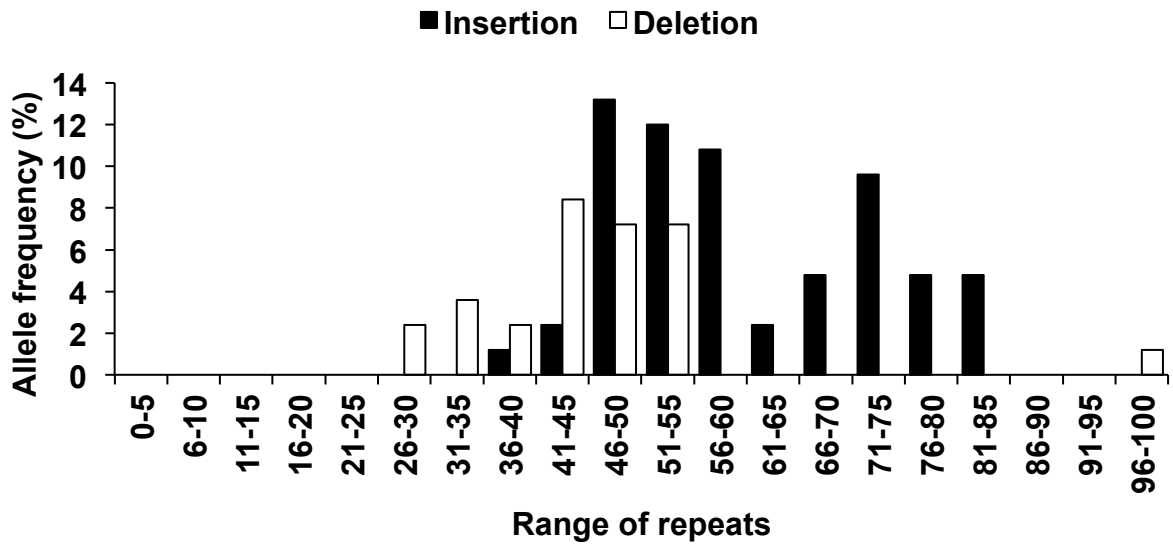


Figure 3.24: The allele frequency distribution of 2-CA-181 locus.

Sequence alignment of the 2-CA-181 microsatellite region of three species of great apes (human, chimpanzee and gorilla) revealed that the CA microsatellite exists in all species suggesting its presence in the common ancestor of great apes (Figure 3.25). The data also

show conservation of the TC-rich repeat flanking the AC microsatellite at the 5′-end and length variation of the AC microsatellite between species (Figure 3.25). Additionally, in the 3′ flanking region the Alu repeat was not found among any of the nonhuman primates examined (Figure 3.25) suggesting that the observed Alu is a recent insertion in the human lineage.

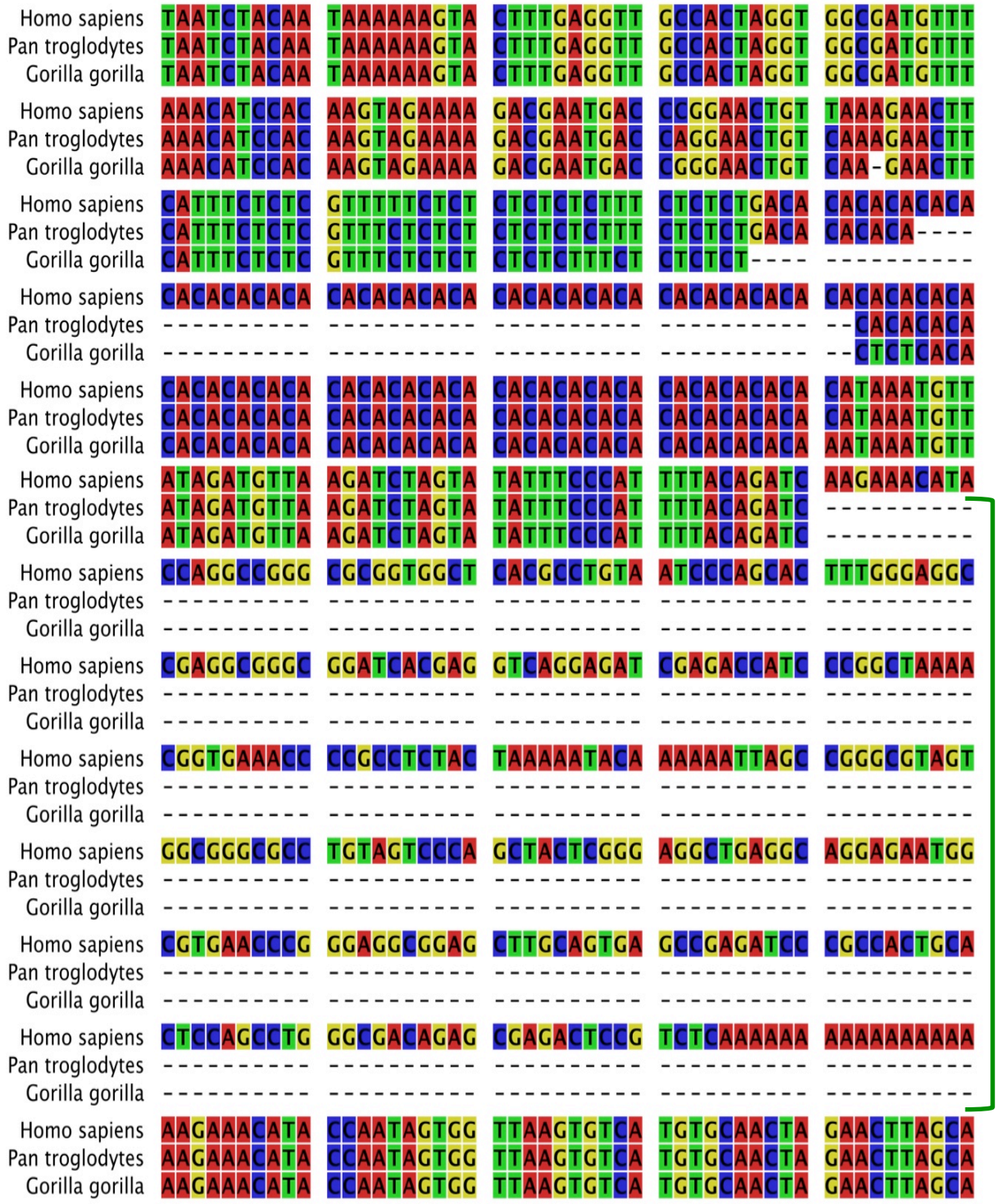


Figure 3.25: Shown here is a DNA sequence alignment of 2-CA-181 microsatellite and flanking region of three great apes (human, chimpanzee and gorilla as indicated). Note the 310 bp insertion (green bracket) in the 3' flanking region of the human sequence.

3.1.2.4 The 21-TCCCT-409 microsatellite is located within a region displaying copy number variation

The 21-TCCCT-409 microsatellite was successfully amplified revealing a high degree of variability with alleles ranging in length from ~35 to 73 repeats (Figure 3.26). Three alleles were observed in ~27% of individuals. *Ensembl* database analysis showed a number of large copy number variants spanning the 21-TCCCT-409 microsatellite ranging in size from ~47 kbp to 108 kbp (Redon *et al.*, 2006, Pinto *et al.*, 2007, Conrad *et al.*, 2010, Cooper *et al.*, 2008). Thus, individuals showing three alleles may carry at least one additional copy of the 21-TCCCT-409 microsatellite.

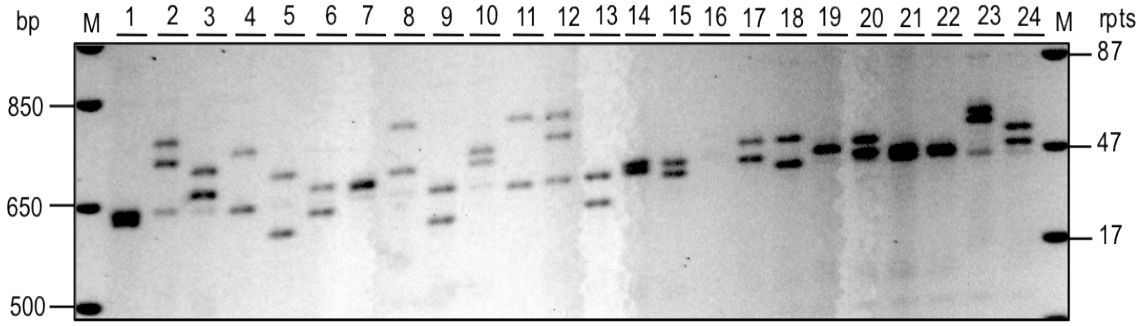


Figure 3.26: 21-TCCCT-409 microsatellite amplification shows some individuals with 3 alleles.

3.2 Discussion

The Tandem Repeat Database (TRDB) was searched to identify pure microsatellites with ≥50 repeats in the human genome (Gelfand *et al.*, 2007). Using the Tandem Repeat Finder algorithm the tandem repeats in a given genome were identified (Benson, 1999). Using the default parameters, the human genome reference sequence HG19 showed 942,700 tandem repeats with pattern sizes from 1 to 2,000 bp. To identify expanded microsatellites, our search was limited to pattern size ≥2 and ≤10 bp. Mononucleotide repeats were excluded as the vast majority are A/T repeats originated from the polyA tail of dispersed Alu repeats. Thus, flanking primer design would be complicated, and mononucleotide repeats are highly prone to PCR slippage that would consequently affect genotyping and mutation detection. HG19 analysis on TRDB showed 291,157 microsatellites with pattern sizes ≥2 to ≤10 bp. The 100% purity parameter reduced the number of tandem repeats to 101,209, whereas the ≥50 repeat copy number parameter hugely reduced the number of selected microsatellites to 15 (Table 3.3).

The HG19 reference sequence is derived from multiple anonymous donors but contains only a single allele at each locus. Thus, polymorphic loci that have some alleles that fit with our selection criteria and others that do not might be missed. The relative expanded allele frequency in the general population will determine the probability of missing a polymorphic expanded microsatellite. This suggests given the additional genomes being sequenced, other expanded microsatellites will be found. The human HG19 reference sequence was cloned in bacterial artificial chromosomes and shotgun sequencing was conducted using Sanger technology (Lander *et al.*, 2001). Given that simple repeats exhibit high instability in *E. coli* (Kang *et al.*, 1995), it is plausible to expect that the cloned alleles do not necessarily bear the true representation of the alleles present in the original donors. Although it is possible that there may be artificial expansion of the repeats in *E.coli*, microsatellites tend to be more susceptible to deletion (Kang *et al.*, 1995). Therefore, the J. Craig Venter genome was analysed to determine if the 15 expanded pure microsatellites identified in HG19 were representative (Levy *et al.*, 2007). Only 8 Venter loci were identified that fit with our search criteria (Table 3.3). Interestingly, no overlap was observed between expanded microsatellites identified in the two reference genomes. This implies that expanded alleles at these loci are variable in the population or they are artificially expanded/deleted in *E. coli*.

In both human genome references, the sum of Chi-square values showed significant deviation (P value = < 0.0001) between the total number observed/expected repeats.

However, of 23 expanded repeats only loci identified on chromosome 2 were found to be overrepresented using Poisson distribution analysis and dot plot analysis showed they were not duplicated.

To determine the degree of allele length variability, flanking primers were designed for each expanded microsatellite identified and PCR was conducted to amplify each locus from blood DNA of ~48 humans. Only the X-AT-535 microsatellite failed to amplify despite testing several pairs of PCR primers. The X-AT-535 microsatellite is an AT microsatellite derived from the polyA tail of an Alu repeat that was located within a very dense region of dispersed repeats extending ~10 kb on either side.

The 21-TCCCT-409, 13-AAG-102 and 2-CA-181 microsatellites all produced unexpected PCR banding patterns. However, all of the remaining 19 microsatellites were successfully amplified and genotyped. They showed allele length variation in the population with observed heterozygosities ranging from 7 to 87%. Surprisingly, the mean allele lengths were not highly correlated with observed heterozygosities. However, two other correlations (number of alleles and mean allele length, and observed heterozygosities and number of estimated alleles) were found to be statistically significant. These loci were selected to comprise an allele >50 repeats in one of the two reference genomes. But no expanded alleles >50 repeats were detected in four of the microsatellites (19-AT-431, 7-AT-232, 16-TTTC-505, and 1-CTCCCT-151). To rule out the possibility that ethidium-bromide stained gels might not detect large expanded alleles, all gels examined here were subjected to Southern blotting and probed with the respective repeat unit. No extra large expanded alleles were observed. Employing this methodology provides an advantage with respect to the detection of >1,000 repeats at the DM1 locus (Monckton *et al.*, 1995). Thereby, the absence of detectable expansions at these loci reflects the rarity of these alleles in the population. Alternatively their presence in the reference genomes may be attributed to cloning or sequencing errors. Four loci showed a relatively low expanded allele frequency (~10%). At the other 12 loci, expanded alleles >50 repeats were detected frequently. Using this approach does identify genuinely expanded alleles, but no locus was expanded in every individual, “the perfect locus”. However, cloning errors may results in deletions, which may explain the absence of perfect loci in HG19 and Venter human genome reference sequences.

No overlaps were observed between the expanded microsatellites identified in the HG19 and Venter genomes and none of the loci showed expanded alleles >50 repeats in all the tested samples. However, at least 5 microsatellites were highly polymorphic.

Expanded alleles >50 repeats were identified in several microsatellites and their level of somatic instability was investigated. The small-pool PCR approach was used (Monckton *et al.*, 1995, Gomes-Pereira *et al.*, 2004) to investigate the levels of somatic mosaicism in blood DNA collected from individuals with age at sampling >40 years. Nearly all of the loci examined showed a low level of somatic mosaicism. The allele length variants detected were marginally (1-3 repeats) different than the inherited progenitor alleles.

Allele length may affect the somatic variation (on at least their detection). Most alleles examined here ranged between 50-100 repeats which is much smaller when compared to the repeat range typically found in the DM1 patient population (200-500) (Morales *et al*., 2012). However, DM1 alleles in the range of 50-100 also exhibit somatic instability (Morales *et al*., 2012). In addition to allele length, *cis*-acting factors may also have reduced the somatic instability at these loci. However, given only one sequence from one reference genome (HG19) was assessed, none of the samples was examined in the context of purity of repeats. Therefore, generalisation of this observation requires assessment of more samples and/or ruling out the possibility of errors introduced during cloning and sequencing. It may be possible that HG19 and the samples sequenced are pure or may have been subjected to sequential or cloning error. Nonetheless, the data obtained from Sanger sequencing showed pure expanded microsatellites in the small number of samples tested (Figure 3.23). Previously, a strong correlation was found between the flanking GC content and relative expandability of expanded CAG•CTG repeats (Brock *et al.,* 1999). Their data showed the GC content variation of flanking region (500 bp) range from 38.5% to 79% and loci such DM, SAC2, and SAC7, which contain 66%, 79% and 71.5%, GC respectively showed high level of instability. However, our data showed GC content ranging between 33% and 58% that is much lower when compared to Brock *et al.,* data of GC content. This may explain the low level of somatic stability observed in the present study.

The 13-AAG-102 microsatellite was amplified using flanking primers constructed on the human reference sequence. However, PCR products were only observed in 19/48 individuals. The amplified alleles were all large and variable, ranging in size from ~40 to 150 repeats, and mostly homozygous. These observed data implied the presence of a high frequency non-amplifiable 'null' allele. Analysis of the 1000 Genomes Project data revealed the existence of 7 different putative deletions. Consequently, a new more distal forward primer (O) was designed to examine the putative deletions. At least one allele could be amplified using the 13-AAG-102-FO/RI in all 48 individuals. However, none of the amplified alleles fit the AAG microsatellite involving deletions predicted by the 1,000 genomes analyses. This may be due to the short reads used in the 1,000 genome project,

which may consequently result in the incorrect alignment of microsatellite alleles and their incorrect interpretation as deletions. Indeed, sequencing 8 alleles of 13-AAG-102 generated using primers of 13-AAG-102-O and 13-AAG-102-RI showed two primary structures. In the allele that matches the human reference, the 13-AAG-102 microsatellite is preceded by a G/A rich 9 bp sequence GAAGAAAGA. In the other allele the 9 bp G/A rich sequence is replaced by a 15 bp sequence (present in other greater apes) preceding the 13-AAG-102 microsatellite. The observed result was confirmed by designing another specific primer for allele 2, 13-AAG-102-ins, containing the 15 bp sequence. Successfully, alleles 2 were amplified using 13-AAG-102-Fins and 13-AAG-102-FO but not 13-AAG-102-FI. These data endorse the presence of two common alleles spanning the junction of the AAG microsatellite with the 5′-flanking DNA. Allele frequency distribution shows no overlap between the 9 bp and 15 bp flanking alleles.

Similarly, The 2-CA-181 microsatellite was amplified using flanking primers based on the human genome reference. No PCR products were observed in all samples tested. The 1,000-genome data revealed the presence of 6 putative structural variants (insertion/deletions) in the 3′- flanking region of the 2-CA-181 microsatellite. These data were used to generate new primers distal from these variants that successfully generated PCR products. However, some of the alleles amplified were smaller than expected. The sequencing data revealed the existence of a 310 bp insertion (Alu repeat). The sequence alignment of the 2-CA-181 microsatellite region of three species of great apes (human, chimpanzee and gorilla) revealed that a TC-rich repeat flanks the AC microsatellite at the 5′ end showing sequence and length variation between species. Additionally, in the 3′-flank the Alu repeat was not found among any of the non-human primates examined. There is no evidence that it is ever excised or lost from a chromosome locus**.** Unlike the 13-AAG-102 microsatellite, the allele frequency distribution showed no linkage disequilibrium with the flanking alleles.

The 21-TCCT-409 microsatellite was amplified and appeared highly polymorphic. Indeed, three alleles were observed in 27% of individuals. The *Ensembl* database showed a number of large copy number variants spanning the 21-TCCT-409 microsatellite ranging in size from ~47 kbp to 108 kbp and individuals exhibiting with three alleles likely to carry at least one additional copy of the microsatellite.

The aims of this project were to identify microsatellites with a relatively high somatic mutation frequency in the general population and age-dependent progression of repeat instability and to define a panel of microsatellites that can be used to predict age of human

tissues. We were able to identify 23 pure expanded microsatellites with relatively low level of somatic mosaicism. However, the frequency of variation was very low. Small pool PCR cannot be used to resolve the somatic instability with large amounts of template DNA, but might be feasible with low quantity of DNA and high resolving gels such as NuSieve or polyacrylamide. Alternatively, using an NGS approach to resolve somatic variants by direct sequencing might be feasible (see Chapter 6).

4 Using next generation sequencing to capture and sequence microsatellite repeat length variation

4.1 Introduction

Previously (in Chapter Three), an attempt was made to measure microsatellite instability (MSI) by separating DNA into multiple small pools containing only single or small numbers of DNA molecules. This resulted in the capturing of both progenitor and low-frequency mutant alleles into pools, where they can be identified and counted using PCR and length analysis. However, SP-PCR is laborious, time consuming and includes a radiation hazard. This technique is suitable for some research questions, but is less likely to be used in laboratories in which a high sample throughput and a rapid turnover is required, such as a forensics lab. Thus, one of the current project's aims was to investigate the potential of new technologies to achieve the same goals in a more rapid manner with higher throughput. New Next Generation Sequencing (NGS) technologies have permitted new approaches by capturing and sequencing selected DNA sequences. This technology has been developed to allow massively parallel sequencing at low cost, generating huge amounts of data (Koboldt *et al*, 2010).

Many NGS approaches use shotgun-sequencing methods in which DNA is fragmented into small segments and singly sequenced. Each sequenced DNA fragment, or read, is then aligned to a reference genome. Multiple reads are aligned to provide the depth of coverage required for the genome in question (Koboldt *et al*., 2010; Harismendy *et al*., 2009). In this experiment, we aimed to sequence a variety of human DNA sequences including microsatellites, DNA repair genes, telomeric regions, and other sequences which are of interest to our research group. A whole genome shotgun sequencing (WGS) approach is appropriate in the vast majority of genomic sequences where they can be sequenced and aligned successfully. However, WGS will sequence both desired and undesired sequences and will generate huge amounts of data in doing so. Using this approach may therefore achieve the experimental goals, but it would be expensive, and the experimental design and data analysis also would be difficult. It is impractical to apply in highly repetitive sequences of considerable length such as microsatellites and telomeric regions. For such sequences, the fragmented DNA generates short reads that may not span the whole microsatellite locus and the minimum amount of flanking sequence required for successful alignment. Moreover, the DNA shearing process is random and could produce some DNA fragments which are physically too small to contain the

required level of sequence. To minimise such obstacles, targeted sequencing was selected, in order to capture only the sequences in question, minimise the size of data to be analysed and reduce costs. Bait (probes) design is a necessary step. Using NGS target enrichment to capture unique sequences (single copy) is simple and straightforward. Custom baits which will be complementary to the desired unique DNA sequences are used to capture the sequences. The data which are generated will be more easily aligned if a single copy is present in the reference genome. Traditional NGS target enrichment is capable of capturing repetitive sequences, but there are some issues concerning the bait design needed to capture such sequences, since repetitive sequences are dispersed across the human genome. First, using bait which is constructed only from a repetitive sequence, such as $(CA)_{20}$ may also capture repetitive sequence $5 \geq CA \leq 40$ repeats. Both small and large repeat sizes may fail the intended capture and the data generated will be also difficult to align to the reference genome unless they contains a unique flanking sequence on both sides for each read. The number of repeats present in the bait will be impossible to determine because these sequences are polymorphic. The number of baits for each pattern and the ratio of bait needed for each repetitive sequence to achieve successful capture will also be difficult to calculate. Therefore, a novel approach has been developed using hybridising baits which only target the flanking regions of repetitive sequences.

4.2 Bait design and sequence capture strategy

In employing the target enrichment NGS approach, sequences of interest must be identified and then captured using probes (baits). The captured sequences are then used to generate the NGS library for each sample tested. The bait design depends on the bait size made available by the manufacturer, the properties of the sequence to be captured, and the capacity of the sequence platform. The technologies used in the experiment were Illumina GA IIx, and Agilent SureSelect^{XT} Target Enrichment System for Illumina Paired-End Sequencing Library. It was used over other target enrichment systems such as NimbleGen, provided by Roche. The NimbleGen system uses 60–90-mer DNA capture probes, whereas the Agilent system uses 120-mer RNA capture probes (Mamanova *et al*., 2010). RNA baits have a higher affinity to capture DNA fragments than DNA ones and can be selectively destroyed in the heteroduplex. The aim was to use a 150 bp sequencing read length and a longer bait size (120 bp) will ensure that the captured DNA fragments will be at least 150 bp in length. Sequencing platforms that generate longer reads, such as Ion Torrent and MiSeq, were not available at the time of

designing this experiment (Quail *et al*., 2012). Moreover, the experiments conducted were novel and the researchers preferred to use the available facilities at the University of Glasgow. The Polyomics unit uses the Illumina GA IIx platform and they are also familiar with the SureSelect^{XT} Target Enrichment System. In this experiment, Agilent's eArray free designing software was used where the maximum bait size is 120 bp and the maximum number of baits present in a single library is 57,646. The Illumina GA IIx platform generates 25-30 million reads from one lane in SE mode and twice this total in PE mode. An Illumina Paired-End sequencer platform was used, which generates a 151 bp read length. Different bait design approaches were conducted depending on the regions which were targeted. Some of these design approaches are novel.

Each finished DNA library will contain different types of captured DNA regions such as microsatellites, DNA repair genes, telomere sequences, and others. In this chapter, the target region selection, the bait design and the capture strategy are set out and explained in turn. Moreover, the results obtained from this experiment are presented and discussed. The results obtained from the captured telomere sequences are discussed in detail in the following chapter (Chapter Five).

4.2.1 Microsatellite selection

In this experiment, only pure (100% match) mono, di, and trinucleotides with copy number ≥5 were investigated. A novel approach was taken, whereby only the unique nucleotides flanking sequence in both sides was used to capture these microsatellites and to ensure that each locus was aligned to a single position on the human reference genome. Firstly, Tandem Repeat Finder (TRF) was used to identify all pure mono, di, and trinucleotides with copy number ≥5 from the *Homo sapiens* 2009 database. The number of pure microsatellites identified was 118,192 loci. Secondly, RepeatMasker® was used to identify pure microsatellites with unique flanking sequence in both sides in different flanking windows (as shown in Figure 4.1).

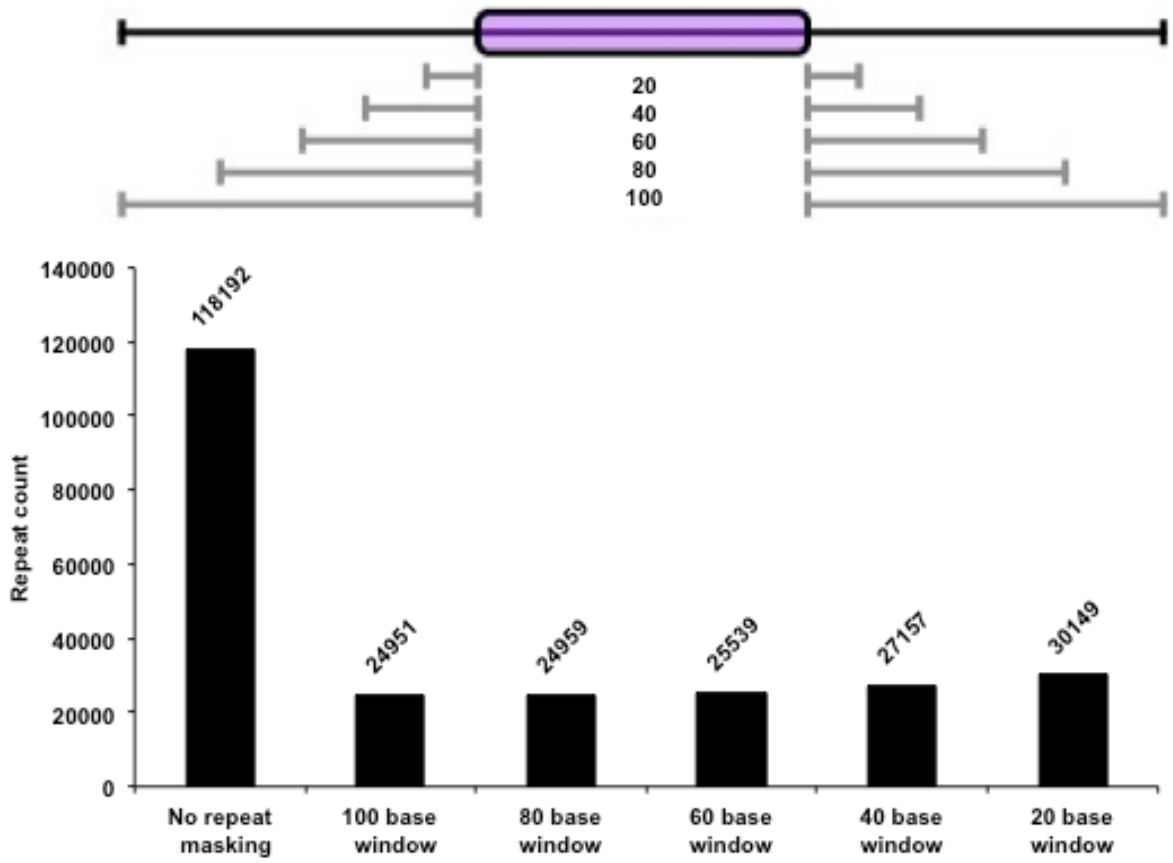


Figure 4.1: RepeatMasker's identification of microsatellites with unique flanking sequences in different flanking windows

The flanking sequence windows which were tested spanned both sides and ranged between 20 bp and 100 bp. The number of pure microsatellites with unique flanking sequences on both sides was inversely proportional to the flanking window screened. The number of microsatellites with unique sequences when a 100 bp flanking sequence on both sides was screened using RepeatMasker® was 24,951, which represents 20% of the total number of microsatellites identified. The number of unmasked microsatellites gradually increased when the flanking window size screened decreased and was at 30,149 loci when 20 bp flanking sequences were screened, representing 25.5% of the total. Thirdly, the results obtained form RepeatMasker® were also verified using BLAST®. The unmasked microsatellites were tested using BLAST® where the microsatellites that only aligned perfectly (*i.e*. a 100% match) to the top hit were selected. The pattern distribution of the microsatellites in the selected flanking windows was also counted, and the results are shown in Table 4.1.

Table 4.1:Microsatellite pattern distribution in different flanking windows

Pattern size	Repeat pattern	TRDB	100 window	80 window	60 window	40 window	20 window
Mono	**A•T**	56,849	4,426	4,426	4,505	4,748	5,236
	G•C	17	7	7	7	7	7
Dinucleotide	**AC•GT**	39,746	16,483	16,490	16,867	17,946	19,757
	AG•CT	4,387	912	912	932	980	1,112
	AT•AT	9,493	1,851	1,852	1,921	2,082	2,455
	GC•GC	8	1	1	1	1	1
Trinucleotide	**AAC•GTT**	1,939	297	297	302	326	365
	AAG•CTT	237	27	27	27	28	35
	AAT•ATT	4,232	381	381	389	417	477
	ACC•GGT	199	91	91	94	101	113
	ACG•CGT	2	1	1	1	1	1
	ACT•AGT	67	13	13	14	15	19
	AGC•GCT	213	131	131	133	138	144
	AGG•CCT	161	70	70	71	74	83
	ATC•GAT	487	175	175	182	195	230
	CCG•CGG	155	85	85	93	98	114
Total		**118,192**	**24,951**	**24,959**	**25,539**	**27,157**	**30,149**

In all of the windows which were screened, the major microsatellite pattern observed in mononucleotides was A•T, which explains the huge reduction in mononucleotide numbers when screened using RepeatMasker®. This pattern arises from Alu repeat sequences, which are dispersed across the human genome. For dinucleotides, AC•GT was the most common pattern and AAT•ATT was the most common in trinucleotides. In all the windows which were screened, the least common microsatellite patterns observed were G•C in mononucleotides, GC•GC in dinucleotides, and ACG•CGT in trinucleotides.

The flanking sequence window was chosen based on the maximum probe (Bait) length generated by the eArray® and the read size was generated by an Illumina Pair End sequencing platform. The bait length was 120 bp and the read length was 151 bp. The 60 bp windows were selected and the number of microsatellites was 25,539. The selected microsatellites are listed in Appendix I.

A detailed analysis of the selected microsatellites was carried out and they were found to consist of 4,512 mono, 19,721 di, and 1,306 trinucleotides (Figure 4.2).

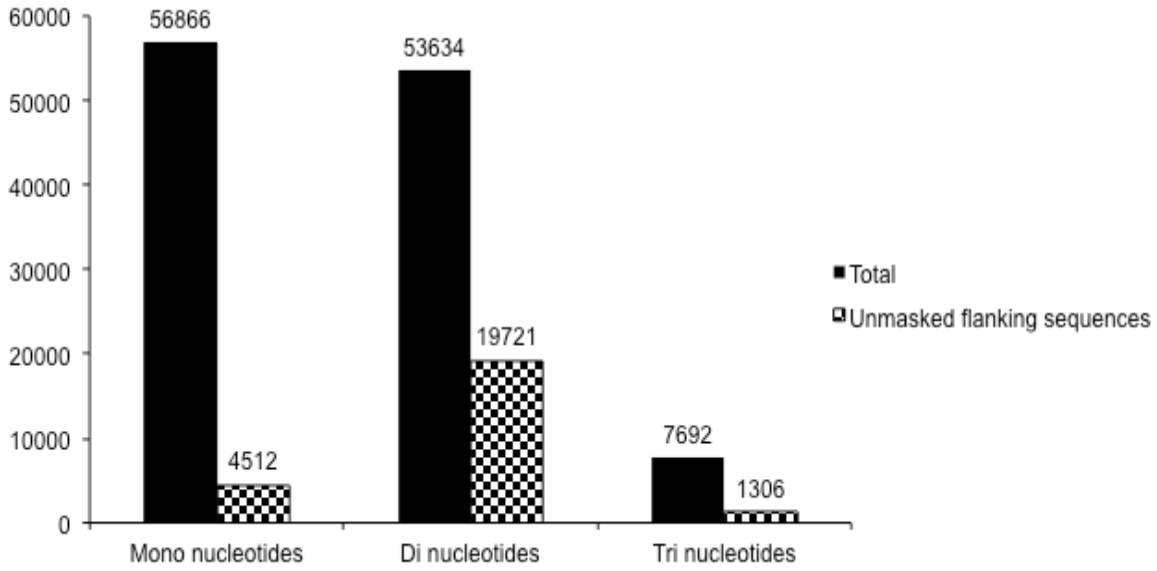


Figure 4.2: Tandem Repeat Finder and RepeatMasker results for mono, di and tri-nucleotides identified in the *Homo sapiens* 2009 database with a 60 bp unique flanking sequence.

The dinucleotides represented 77.2% of the microsatellites selected, 17.7% were mono-nucleotides, and 5.1% were tri-nucleotides. Mononucleotides were massively reduced and only 8% remained for further analysis. Similarly, only 36.7% and 17% of dinucleotides and trinucleotides remained, respectively. The microsatellites chosen were spread across all chromosomes with reasonable coverage (Figure 4.3).

More detailed analysis was conducted to identify the number and pattern of microsatellites for each chromosome (the results are shown in Table 4.2). Chromosome 2 had the highest coverage with a total of 2,371 loci, while 1,815 loci were dinucleotides, 426 loci were mononucleotides, and 130 loci were trinucleotides. Chromosome Y had the lowest coverage with a total of 217 loci, while 155 loci were dinucleotides, 54 loci were mononucleotides and 8 loci were trinucleotides.

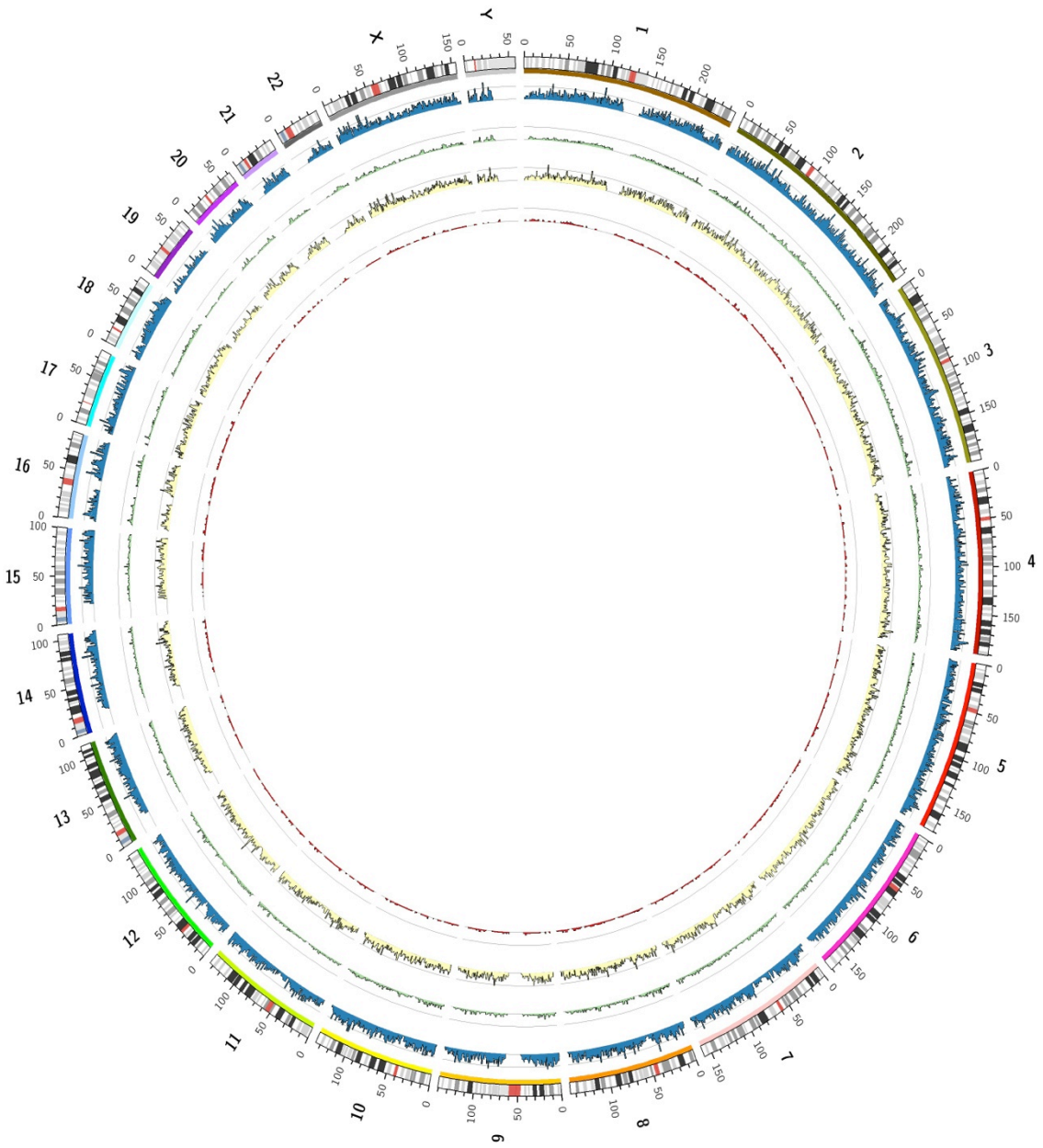


Figure 4.3: Microsatellite chromosomal distribution. Total number of microsatellites (blue) unselected and the number of mono (green), di (yellow) and trinucleotides (purple) selected produced by Circos software.

Table 4.2: Unique sequence microsatellites selected chromosomal distribution

Chromosome	All	Mononucleotides	Dinucleotides	Trinucleotides
Chr1	1,934	315	1,521	98
Chr2	2,371	426	1,815	130
Chr3	1,870	368	1,435	67
Chr4	1,605	266	1,278	61
Chr5	1,635	293	1252	90
Chr6	1,552	283	1,191	78
Chr7	1,384	252	1,059	73
Chr8	1,299	258	956	85
Chr9	1,042	195	791	56
Chr10	1,262	243	963	56
Chr11	1,120	185	874	61
Chr12	1,160	216	875	69
Chr13	959	144	777	38
Chr14	742	128	575	39
Chr15	750	132	581	37
Chr16	656	131	483	42
Chr17	656	100	509	47
Chr18	746	132	577	37
Chr19	281	37	217	27
Chr20	522	87	409	26
Chr21	305	50	245	10
Chr22	226	32	182	12
ChrX	1,245	185	1,001	59
ChrY	217	54	155	8
Total	**25,539**	**4,512**	**19,721**	**1,306**

4.2.1.1 Microsatellite bait design and capture hypothesis

Since the microsatellites which were identified had unique flanking sequences, a single probe (bait) was designed to capture each locus. The novel approach taken in the experiment assumed that baits only consist of unique flanking sequences. Repetitive microsatellite patterns were removed, except if they were present in the unique flanking sequence. Thus, for each microsatellite, the 60 bp unique flanking segments in both sides were used to build a single bait and the total number of baits was 25,539 (Figure 4.4). The microsatellite bait number was the highest, at 44% of the total baits and covering 3,064,680 bp (3 Mb). Our bait capture strategy suggested that when using fragmented DNA, both bait unique flanking sequence segments will capture both complementary microsatellite-flanking ends and a loop structure (bridge) will be formed to cover the repetitive sites that lie between both flanking segments (Figure 4.4). We assume that the bait bridge approach will be able to capture microsatellites of reasonable length up to approximately 60 bp. The full spanning reads generated should be alignable and used to study microsatellites genotypes. Larger microsatellites are also likely to be captured, but reads would be likely only to span either the left or the right arm, with very few full spanning reads expected.

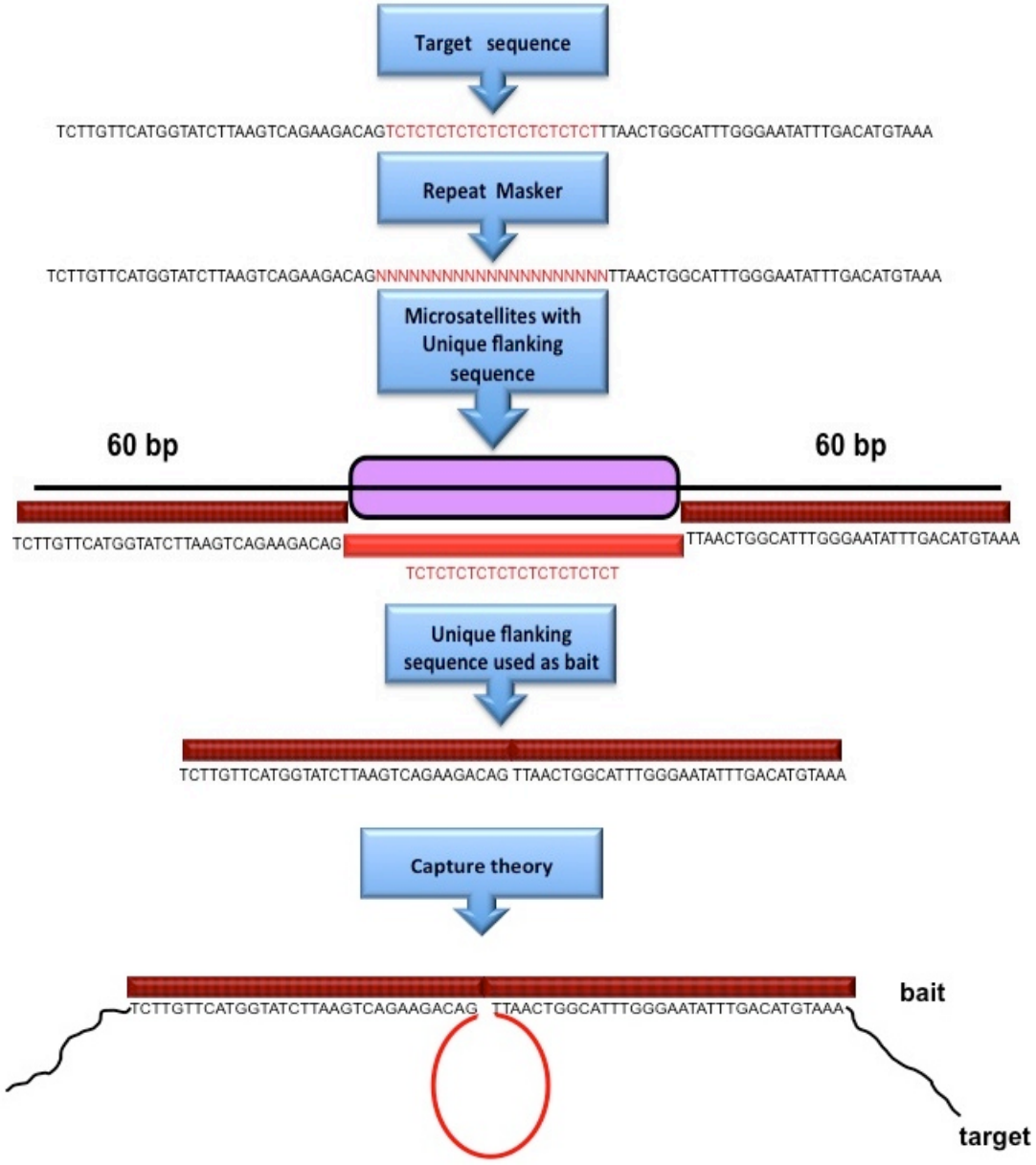


Figure 4.4: A novel probe (bait) designing strategy to capture pure microsatellites with a unique flanking sequence.

4.3 Microsatellite capturing process

The new approach used hybridising baits to target the unique flanking regions of the chosen microsatellites, in order to minimise unwanted hybridisation. The baits were manufactured by Agilent technologies (http://www.genome.duke.edu/cores/microarray/services/ngs-library/documents/SureSelectCustomerFacing_Nov212012.pdf) and based upon the SureSelect Target Enrichment system. The biotinylated baits allow targeted fragments to be captured and then separated from the unwanted fragments via the binding of biotin onto streptavidin-coated magnetic beads (a process shown in Figure 4.5). The fragments containing microsatellites can then be washed off the beads and amplified for sequencing. Each step will now be discussed in more detail.

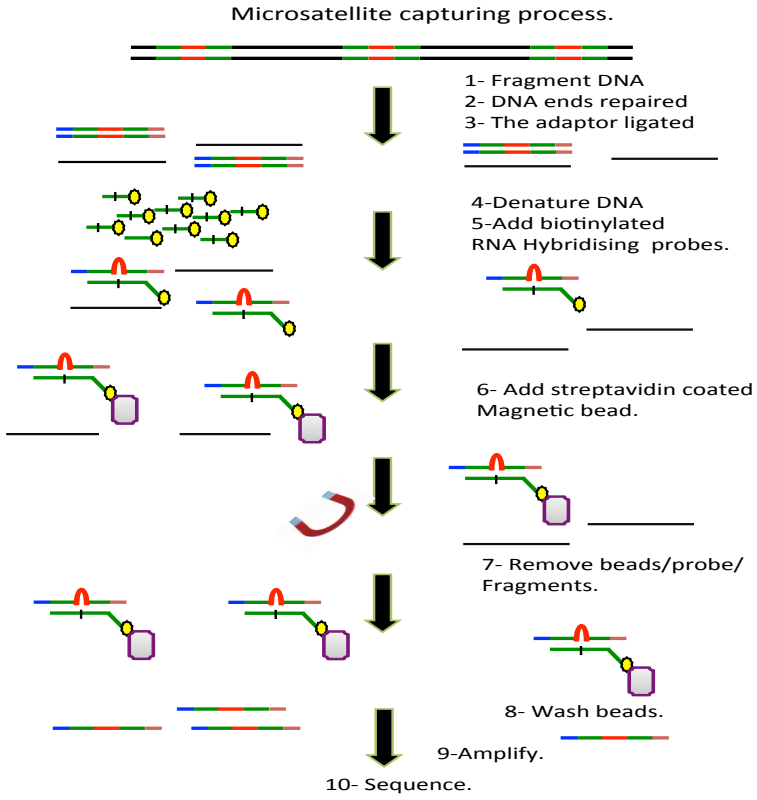


Figure 4.5: Schematic representation of the microsatellite capture process based on the SureSelect system. Step 1-3: the genomic DNA containing microsatellites (red) is sheared. Step 2: the DNA ends are repaired and the specific paired-end adaptor (blue and brown) added. Steps 4 and 5: dsDNA denaturation and biotinylated, hybridising baits are added (green/yellow) to locate and complement the microsatellite flanking sequences (green). Step 6: Magnetic beads coated in streptavidin (grey with mauve) are added and the biotin binds to the streptavidin. Step 7: These beads are then removed via magnetic force. Step 8: The fragments containing microsatellites can then be washed off the beads. Step 9: the targeted fragments are then amplified by PCR. Step 10: The generated library is then sequenced using the Illumina platform.

4.3.1 Other bait design approaches

In this experiment others sequences of interest to our research group were included, such as telomere regions, DNA repair genes, the *DMPK* region and SNPs. The baits were designed by the same tools and set at the same size (120 bp) (see Figure 4.6). In common with microsatellite baits, RepeatMasker was used to mask repetitive regions, if these proved to be present.

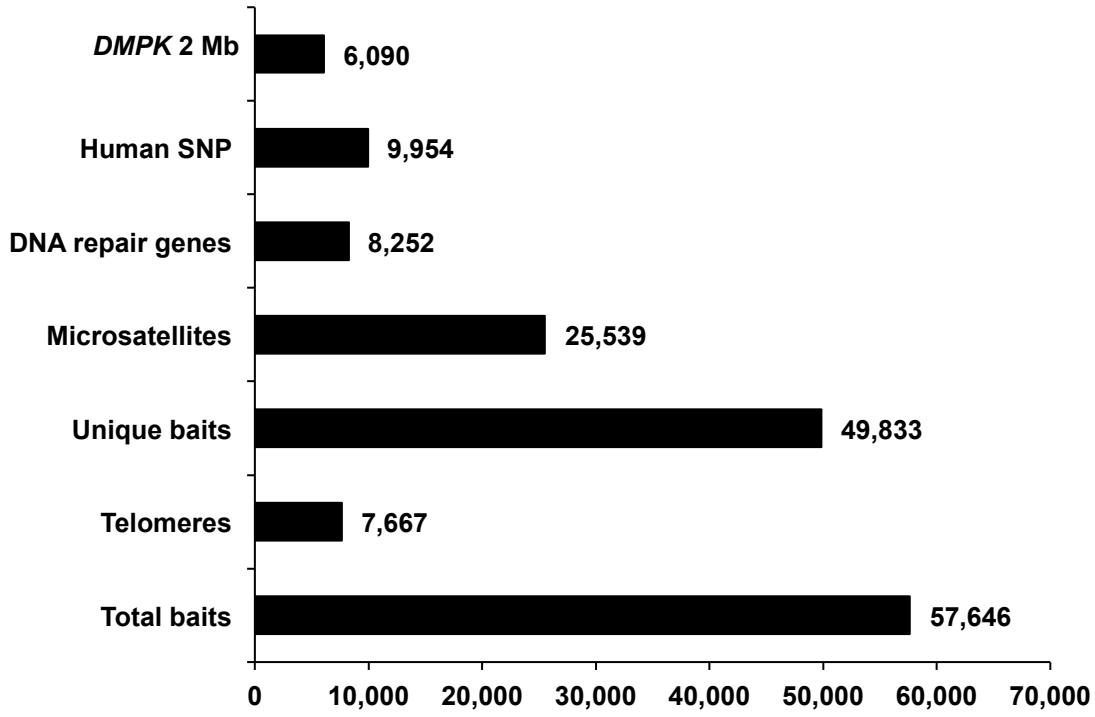


Figure 4.6: Bait distribution: The custom baits used to capture all sequences in question

The total number of DNA repair gene baits was 8,252 and these covered only the exons for the 233 genes selected. NCBI (http://www.ncbi.nlm.nih.gov/gene) was used to identify 542 DNA repair genes. Only a limited number of baits were available, and 233 genes were selected. Mismatch repair genes were prioritised to identify any variation in their sequences and the consequences of these on somatic instability. Polymerase and helicase genes were also included. The DNA repair gene baits cover 3,118 exons with 990,240 bp lengths and represent 14.3% of the total baits. The DNA repair genes are listed in Appendix II.

To capture the telomeric regions, 7,667 baits were ordered and these consisted of $(TTAGGG)_{20.}$ The telomeric region average size was estimated ~10 kb in each chromosome end (Neumann *et al*., 2002). The 46 chromosomes present in the human cell have 92 ends covering 920,000 bp. The number of baits used was calculated by dividing the size of the telomeric regions by the size of the bait (920,000 /120= 7,667) and these represented 13.3% of the total baits (as discussed in Chapter Five).

In addition, a 2 Mb region on chromosome 19 based around the *DMPK* gene was included and covered chr19: 45,000,000-47,000,000 bp. Unmasked regions on chromosome 19 including the *DMPK* gene were covered with 6,090 baits and represented 10.5% of the total baits. The data generated will be used to study *cis*-acting modifiers and disease haplotypes, since the samples were collected from myotonic dystrophy families.

A single nucleotide polymorphism (SNP) is a single site variation in genomic DNA and is the most frequent type of variation (Sherry *et al*., 2001). A single bait was designed to capture each human SNP, which was identified by dbSNP (http://www.ncbi.nlm.nih.gov/snp) with a minor allele frequency of >40% in the general population. NCBI dbSNP Build ID132 was used to identify 15,026 SNPs, of which only 9,954 SNPs were within unique DNA sequences, and these occupied 17.2% of the total baits. A list of SNPs is presented in Appendix III.

A set of 23 loci that are commonly used in forensic labs to conduct DNA profiling analysis (including all the 13 loci from U.S Combined DNA Indexing System (CODIS) were also included in this experiment (Budowle *et al*, 1999). 23 baits were generated to capture these loci. The sum of unique baits included in this experiments was 49,833, covering ~6 Mb and representing 86.4% of the total baits. The total number of baits was 57,646.

4.3.2 Samples

The samples selected to be tested were obtained from myotonic dystrophy type 1 patients and include nine samples separated by 5-10-year age intervals, and 12 samples derived from two different sampling dates for six individuals, one family trio and one mother and son. The total number of samples to be sequenced was 26, and these covered an age range of between 18-75 years. Thousands of microsatellites will be sequenced to an approximate read depth of 173x, as a typical paired-end run can generate up to 10×10^7 reads run which enable the detection of somatic mutations with a frequency of ~1%. Theoretically, the results then can be used to

make reliable estimates of somatic instability. Variation within individuals can be measured and single repeat differences can be detected. This experiment aims to measure individual repeat lengths (genotype) that can be used as a means of relationship testing. Using related family groups (such as the trios and mother and son) could also help to aid understanding of the intergenerational mutation rate.

4.3.3 Sheared DNA Quality

In order to capture the selected sequences using baits, the DNA size must be relative to the size of the baits. High-power ultrasound (~20 kHz frequency) generated by a Bioruptor UCD 200 was used to fragment 3 µg of genomic DNA which had previously been extracted from peripheral blood lymphocytes to generate a DNA library for each sample to be sequenced. The DNA samples used were highly pure and, have an OD260/280 ratio of between 1.8 and 2.0, and should not be degraded. The ratio of absorbance at260nm and280nm is used to assess the purity of DNA and RNA. A ratio of ~1.8 is generally accepted as pure for DNA; a ratio of ~2.0 is generally accepted as pure for RNA. If the ratio is appreciably lower in either case, it may indicate the presence of protein, phenol or other contaminants that absorb strongly at or near 280 nm. The samples were fragmented in 100 µl volume at a low power setting for 7 x 10 minutes, with 30 seconds on and 30 seconds off, or until the samples were fragmented to the correct size (Figure 4.7). This process was carried out in the Polyomics unit at the University of Glasgow.

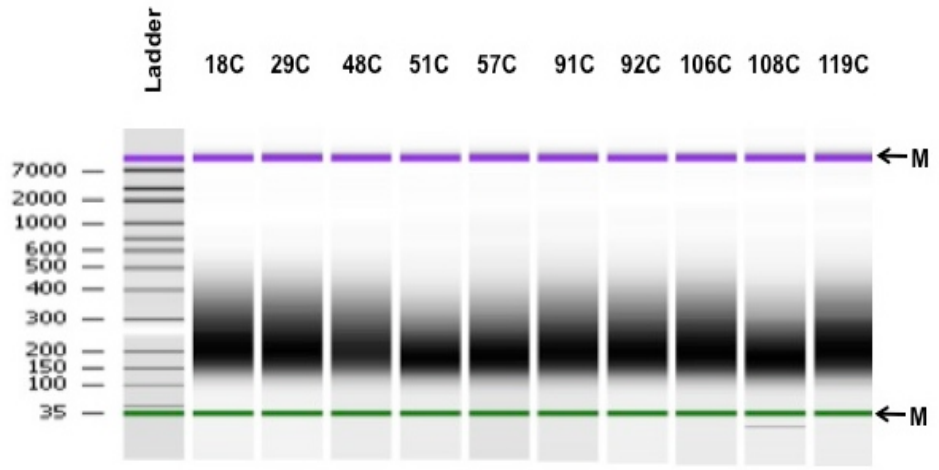


Figure 4.7: Sheared DNA Quality. Pseudo gel was generated by Agilent 2100 Bioanalyser**.** Bioruptor DNA shearing was used to sonicate 3 µg of sample DNA to achieve the desired fragment size between 150 bp and 300 bp. Only 10 samples are shown, where M is a DNA marker.

The samples were run on an Agilent 2100 Bioanalyser on High sensitivity DNA chips to assess the quality of the sheared DNA and to ensure that the finished sheared DNA was between 150 and 300 bp (Figure 4.8). Different peaks were identified and their average size was calculated.

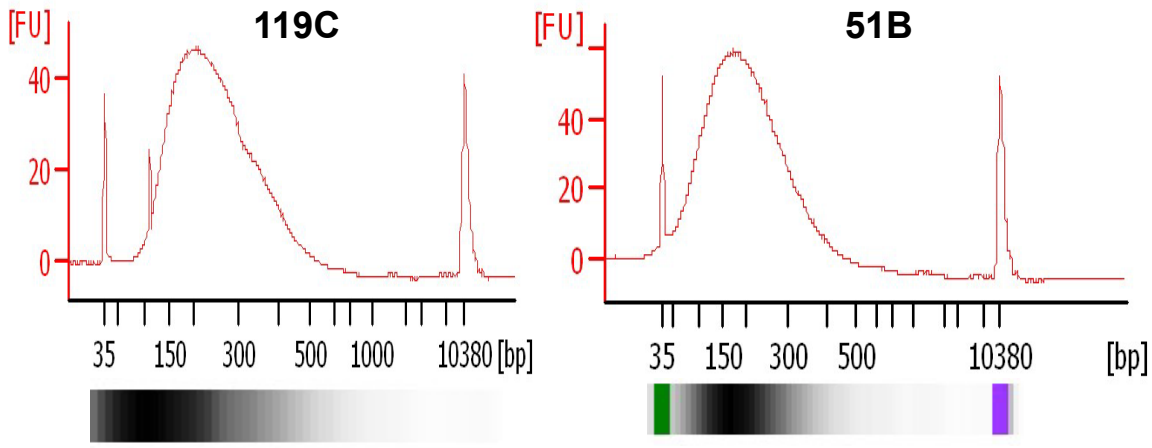


Figure 4.8: Assessing the quality of sheared DNA using an Agilent 2100 Bioanalyser. The electropherogram generated by Agilent Bioanalyser shows a distribution with a peak size of between 150-300 bp**.** Only sample 119C for the individual with the oldest age (75 years) and average size 291 bp and sample 51B for the individual with the youngest age (18 years) and average size 254 bp are shown. The pseudo gel of each sample is shown below the corresponding graph.

4.3.4 The effect of individual age at sampling on the sheared DNA fragment size

The total number of samples to be sequenced was 26 and these covered an age range of between 18-75 years. The sheared DNA fragment size peak ranged between 183 - 302 bp.

The DNA chromatin structure is lost with age and becomes less compact (Feser and Tyler, 2011). The quality of DNA also depends on the extraction method used, since the oldest DNA sample was extracted ~18 years ago (Simbolo *et al*., 2013). This may affect the size of fragmented DNA. The effect of age at sampling on the sheared DNA fragment size was calculated and is shown in Figure 4.9. A linear regression analysis was calculated to measure the effect of age at sampling and showed that $r^2 = 0.0004$ and the P value = 0.92. This result indicated no significant effect for age at sampling on the DNA fragment size.

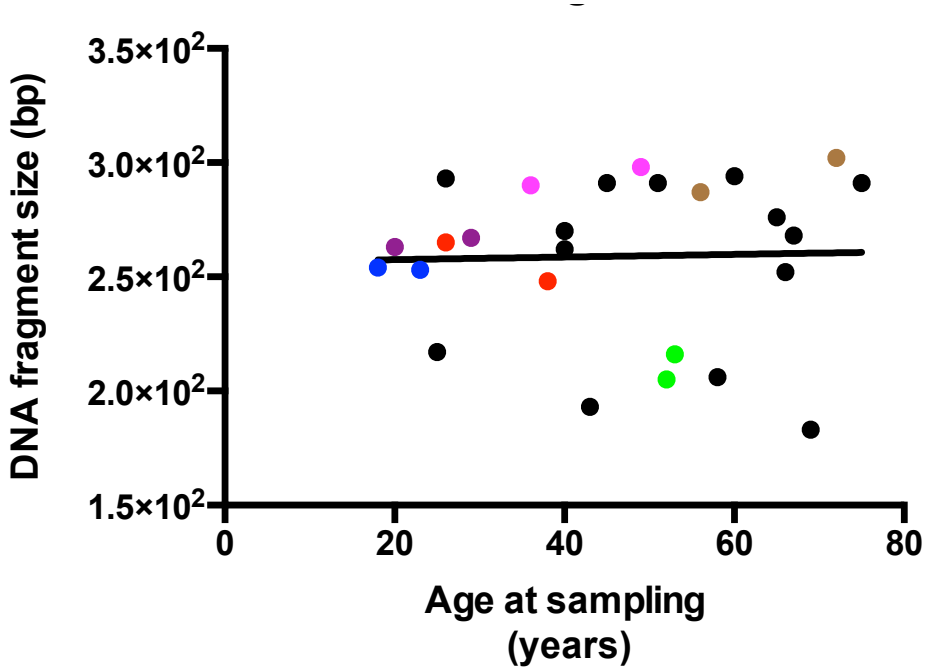


Figure 4.9: The effect of age at sampling on sheared DNA size. Paired samples from the same individual are shown in colour.

4.3.5 The effect of sample storage time on sheared DNA fragment size

Similarly, the sample storage times (-80°C) ranged between 4 months for sample 134C male, which had an age at sampling of 40 years, and 222 months and for sample 18B male, which had an age at sampling of 56 years. Long-term storage may cause degradation of extracted DNA and (Madisen *et al.,* 1987). The effect of sample storage time on sheared DNA fragment size is shown in Figure 4.10. A linear regression analysis was calculated to measure the effect of sample storage time on the sheared DNA's size and showed that $r^2 = 0.021$ and the P value = 0.54. This result revealed no significant effect of sample storage time on sheared DNA fragment size.

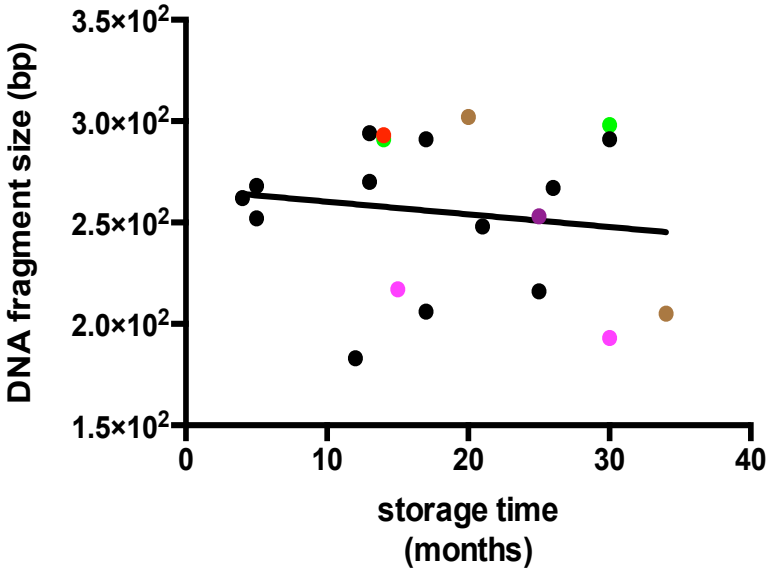


Figure 4.10: The effect of sample storage time on sheared DNA fragment size. A linear regression analysis conducted on samples with storage time ≤ 50 months. Paired samples in colour.

4.4 NGS library

Shearing genomic DNA using sonication is a random process and can cause damage to the DNA fragment ends, and it also generates mixtures of DNA fragments containing 3′ or 5′ overhangs and blunt ends. The overhang ends will be of varying lengths and ends. Damaged ends need to be repaired prior to hybridisation. End repair converts the overhang ends generated from physical fragmentation into blunt ends using T4 DNA polymerase and a Klenow DNA polymerase. Klenow DNA polymerase has 3′→5′ exonuclease activity, but lacks 5′→3′ exonuclease activity and removes 3′ overhangs, and the polymerase activity fills in the 5′ overhangs. The 5′ ends of the DNA fragments are phosphorylated using T4 Polynucleotide Kinase. The addition of 5′ phosphates to oligonucleotides allows for subsequent ligation. Adding single `A` bases to the 3′ ends of the blunt DNA fragments prevents them from ligating to one another during the adapter ligation step. A single T-base on the 3′ end of the adapter offers a complementary overhang for the ligation step. Ligation of the indexing-specific paired-end adapter using a 10:1 ratio of adapter to genomic DNA insert was also carried out prior to the hybridisation step. The optimal amount of indexing of the adaptor-ligated library in the PCR is 250 ng, as quantified on a Bioanalyser DNA1,000 chip. Five PCR cycles were used to amplify the adapter-ligated library in order to generate linear DNA sequences, as Table 4.3 shows.

Table 4.3: PCR program used to amplify the library

Step	*Temperature*	*Time*
Step 1	98°C	2 minutes
Step 2	98°C	30 seconds
Step 3	65°C	30 seconds
Step 4	72°C	1 minute
Step 5	98°C	Repeat step 2 through step 4 a total of 5 times
Step 6	72°C	10 minutes
Step 7	4°C	Hold

Based on DNA quality, different library preparations can generate slightly different results. In most cases, an adequate yield can be produce from five cycles for subsequent capture without introducing bias or non-specific products.

Purification steps were conducted after each step using AMPure XP beads. The sheared DNA was now ready for the hybridisation step, which involved mixing the sheared genomic DNA with the specific biotinylated library baits in a hybridisation buffer. The presence of the specific baits with unique sequences minimises unwanted hybridisation. Bound targeted fragments were removed from the unwanted unbound fragments by the binding of biotin to streptavidin-coated magnetic beads. The captured fragments were then washed off the beads and amplified for sequencing. The library has insert sizes ranging from 250–350 bp. After the capture, 14 PCR cycles were conducted in order to add the index tags (Table 4.4). The concentration of the baits as well as the primer and adaptor sequences is proprietary. The finished library DNA size was within the range of 350–450 bp. An accurate quantification of the finished DNA library is crucial to the data which are generated on Illumina sequencing platforms. Any overestimation of finished DNA library concentration will result in a lower cluster density after amplification. Inversely, any underestimation of finished DNA library concentration will result in too high a cluster density on the flow cell, which can produce low cluster resolution. The finished library DNA fragments were first quantified on the Qubit and then run on the Bioanalyser, pre- and post- capture, to establish the library profile and measure the average fragment size (Figure 4.11).

Table 4.4: PCR program used to amplify the capture library and to add index tags

Step	*Temperature*	*Time*
Step 1	98°C	2 minutes
Step 2	98°C	30 seconds
Step 3	57°C	30 seconds
Step 4	72°C	1 minute
Step 5	98°C	Repeat step 2 through step 4 a total of 14 times
Step 6	72°C	10 minutes
Step 7	4°C	Hold

Finally, the library was quantified by real-time PCR using the Kapa library quantification kit. This contains DNA Standards (six 10-fold dilutions) and a 10X Primer Premix, paired with KAPA SYBR® FAST qPCR Kits to accurately quantify the number of amplifiable molecules in an Illumina library. The KAPA Illumina DNA Standard consists of a linear DNA fragment flanked by qPCR primer binding sites with 452 bp lengths. The Illumina adaptors were constructed containing the KAPA qPCR primer sequences. The qPCR which was conducted consisted of 35 cycles starting with an initial denaturation step at 95°C for 5 minutes and each cycle included a denaturation step at 95°C for 30 seconds, and annealing and extension steps which were both at 60°C for 45 seconds. Quantification was achieved by inference from a standard curve generated using the six DNA Standards. Library quantification is highly reliant on the accurate dilution of the library DNA. The finished library DNA was between 350 and 450 bp.

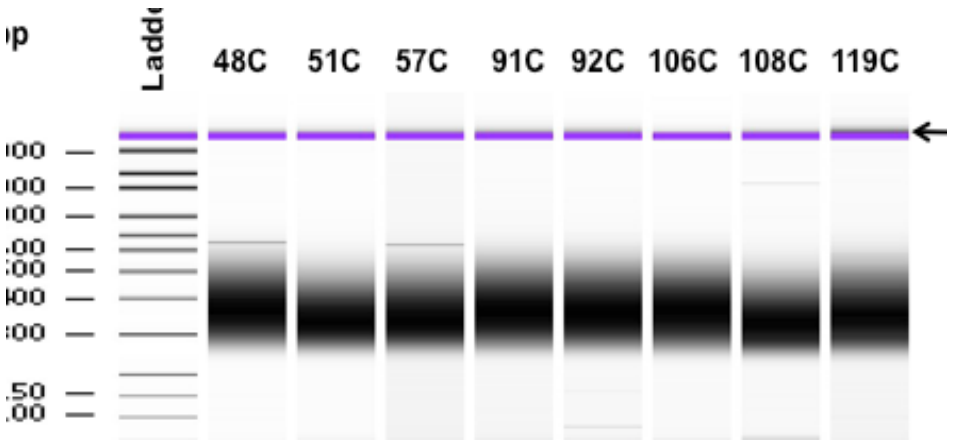


Figure 4.11: Finished library DNA fragment size quality: Pseudo gel generated by Agilent 2100 Bioanalyser, used to assess the quality of the finished library DNA. The finished library DNA was between 200 and 400 bp. 8 samples are shown, where M is a DNA marker.

The finished DNA library quality was assessed by identifying the number of peaks for each sample, and the peak average was then determined (Figure 4.12). The average sizes of the finished DNA library were 430 bp for the oldest individual (119C) with an age at sampling of 75 years, and 343 bp for the youngest individual (51B) with age at sampling of 18 years. Library preparation was carried out by the Polyomics unit at the University of Glasgow.

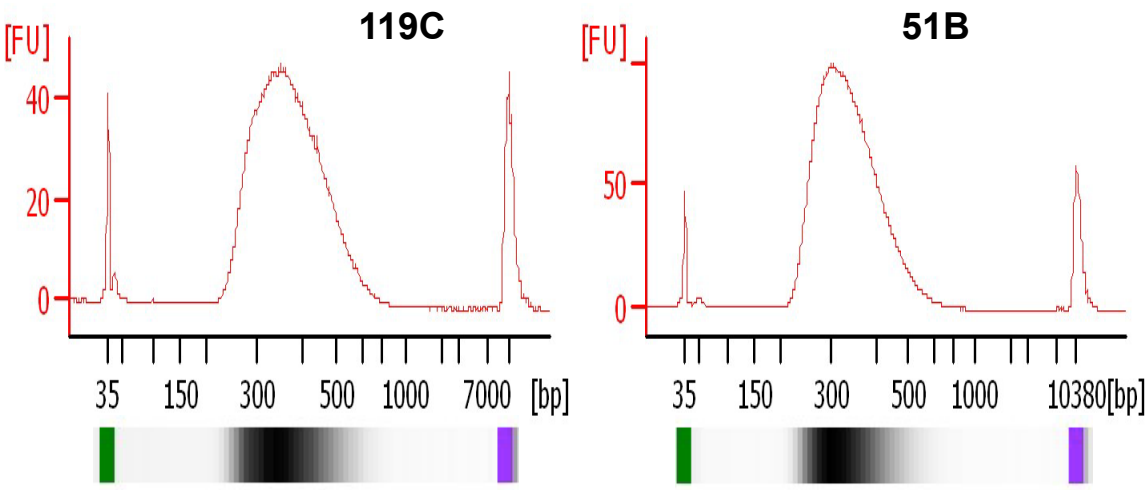


Figure 4.12: Assessment of the quality of finished library DNA fragments using Agilent 2100 Bioanalyser. The electropherogram shows a distribution with a peak size of between 200-400 bp**.** Pseudo gel, generated by Agilent 2100 Bioanalyser, are shown below the peak curve graph. Only samples 119C and 51B are shown.

4.4.1 The effect of age at sampling on the finished library DNA fragment size

Sonication was used to fragment the DNA, a process which could raise the temperature of the DNA solution above ~20°C. Small DNA fragments that may be generated from DNA collected from older individuals (especially those with a high AT content), may be denatured by heating. Single-stranded fragments will not ligate to the double-stranded adapters, which would result in their under-representation in the final library (Illumina Catalog # PE-930-1001, 2011). The age effect on the finished library was calculated and is shown in Figure 4.13. The linear regression analysis produced results of $r^2 = 0.06$ and the P value = 0.23. This indicated no significant effect of age at sampling on the finished library DNA fragment size. However, we can observe a positive trend where the finished DNA library increases with age.

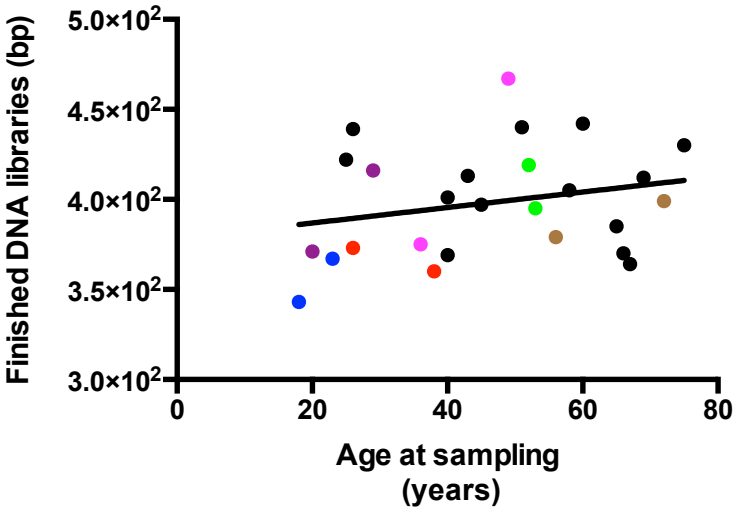


Figure 4.13: The effect of age at sampling on the finished library DNA fragment size (paired samples are shown in colour).

4.4.2 The effect of sample storage time on the finished library DNA fragment size

The quality of the DNA is also reliant on the efficiency of the storage method employed, and the age of the oldest sample stored ~ 18 years ago (Simbolo *et al*., 2013). The effect of sample storage time on the finished library DNA fragment size was analysed and the results are presented in Figure 4.14. The linear regression analysis showed that $r^2 = 0.087$ and the P value = 0.21. This indicated that there was no significant effect of age at sampling on the finished library DNA fragment size.

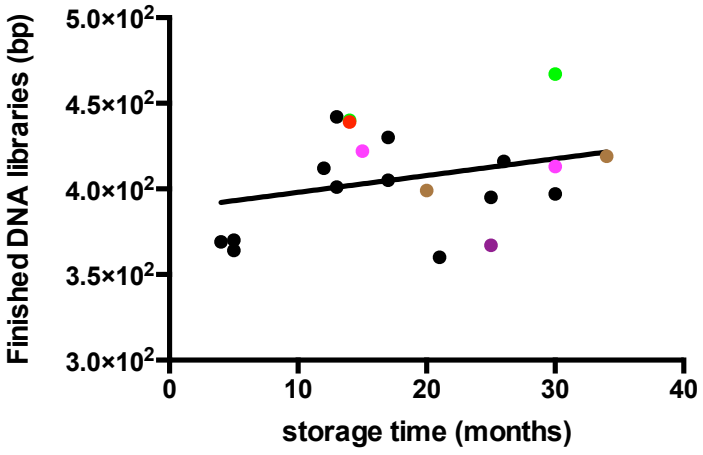


Figure 4.14: The effect of sample storage on the finished library DNA size (paired samples are shown in colour).

4.4.3 The effect of sheared DNA fragment size on the finished library quality

The captured DNA was assessed by analysing the effect of sheared DNA size on the size of the finished library DNA, and the results are shown in Figure 4.15. A linear regression analysis showed that $r^2 = 0.01$ and the P value = 0.63. This indicated that there was no significant effect of sheared DNA size on the finished library DNA size which was generated.

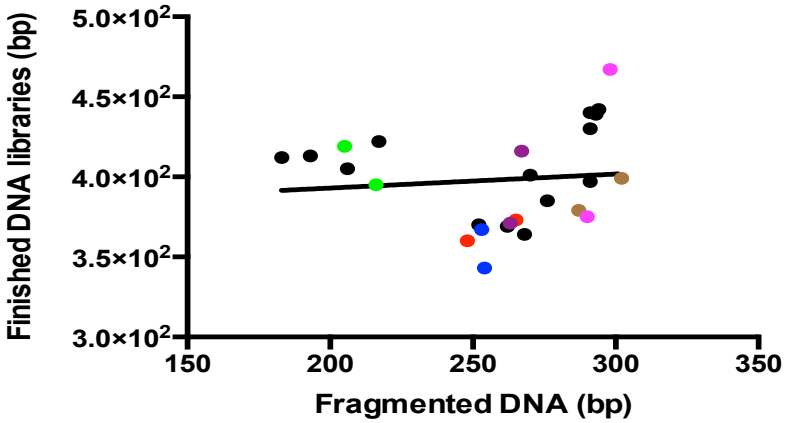


Figure 4.15: The effect of sheared DNA size on the finished library DNA size. (Paired samples are shown in colour).

4.5 Results

4.5.1 Total captured NGS reads

Next generation sequence (NGS) reads were generated using the Illumina Platform. Its Pair Ended (PE) mode generates 151 bp reads. Beside the novelty in some of the sequence capture methods and novelty of approaches in others, this experiment has successfully generated millions of reads for each sample (Figure 4.16). The total number of reads generated span between 5.5 and 9.5 million reads for the samples tested. The average read count was 8.0 x10^6. The highest read count was 9,484,008, observed in male subject 134C who was 40 years of age at time of sampling. The lowest read count was 5,745,730 reads, and this was observed in male subject 119C who was 75 years of age at sampling. These sequencing processes using Illumina Platform and bioinformatics work (reads count and mapping) were carried out by the Polyomics unit at the University of Glasgow.

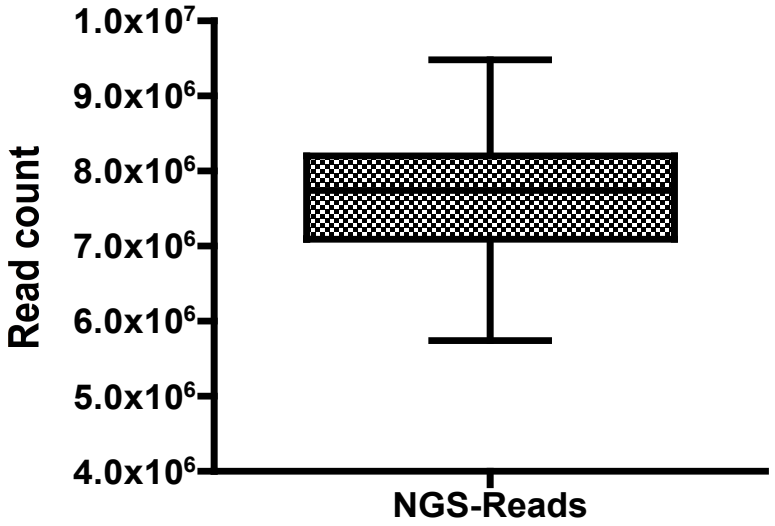


Figure 4.16: The total number of reads generated using the Illumina sequencer platform. The average read count was 8.0 x10^6. The highest read count was 9,484,008, where the lowest read count was 5,745,730 reads.

4.5.1.1 The effect of age at sampling on total reads count

The reads obtained from these experiments include unique sequence reads and age-dependent telomere reads. The effect of age at sampling on the total read count is shown in Figure 4.17. A significant effect was observed when the linear regression analysis showed that $r^2 = 0.19$ and the P value = 0.028. This indicated that age at sampling influences the total read count. Specifically, the total number of reads captured decreases as the age at sampling increases. This can be explained by the presence of telomere reads, which are known to shorten with age and which therefore cause a decrease in the total read counts.

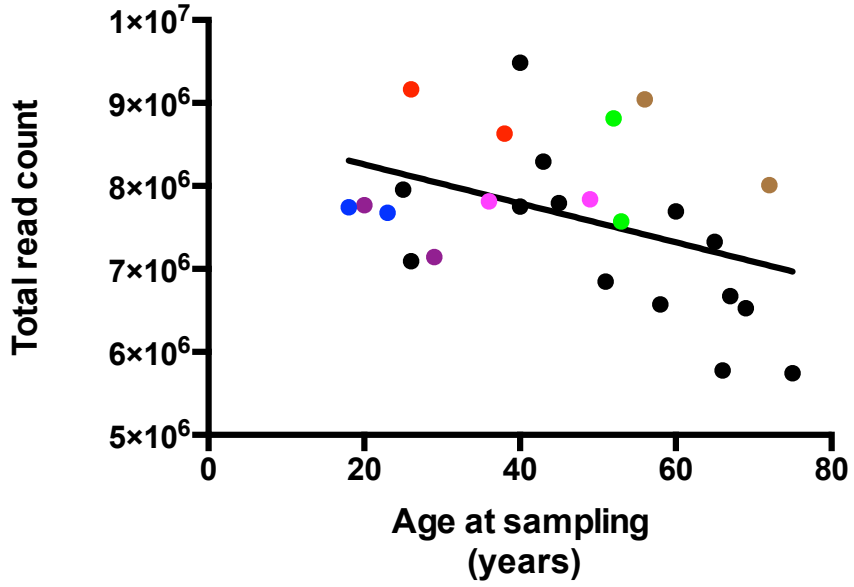


Figure 4.17: The effect of age at sampling on the total reads count (paired samples are shown in colour).

4.5.2 Unique sequences read count

In this experiment, unique DNA sequences were involved, such as microsatellites with unique flanking sequences and DNA repair genes. These unique DNA sequences were separately counted. Firstly, the telomere reads were removed and a trimming process was carried out to trim the reads of sequence adapters. Following that, a quality trim was performed to remove very poor quality reads. The remaining trimmed reads were mapped to the bait sequence padded with 100 bp on either side. The remaining trimmed reads that were not mapped to the unique sequence baits were mapped to the HG19. The mapping process was conducted using Bowtie 2, and the results are shown in Table 4.5.

Of the total reads captured, the unique DNA sequences captured represent 60% against 86% of the bait ratio. The read ratio of SNP, DNA repair genes and CHR19 (2Mb region) is very similar to the ratio of the bait used (as shown in Table 4.5). By contrast, against 44% of bait, only 24% of the microsatellites reads were obtained, suggesting a 50% lower result than the expected rate of capture of reads. Moreover, around 26% of the reads captured that did not map to the baits, were mapped to the human genome reference.

Table 4.5: Detailed analysis of NGS captured reads

Sample ID	SNP	DNA repair genes	*DMPK*	Disease loci	*DHFR-MSH3*	CHR19 2 Mb	Microsatellites	Reads mapped HG19	Total unique sequence reads	(Unique reads/Total reads)%	Total reads
DMGV119C	899,596	750,937	35	1,434	377	384,593	1,395,726	1,453,716	3,432,698	60%	5,745,730
DMGV18C	1,221,164	1,040,758	32	2,142	486	557,177	1,983,579	2,222,341	4,805,338	60%	8,010,376
DMGV99C	989,061	848,069	23	1,694	418	470,809	1,492,667	1,768,032	3,802,741	58%	6,527,096
DMGV129C	1,080,185	888,269	22	1,727	367	467,583	1,680,458	1,709,826	4,118,611	62%	6,672,786
DMGV133C	941,215	795,857	12	1,346	368	389,700	1,453,850	1,501,368	3,582,348	62%	5,776,950
DMGV22B	1,152,549	974,835	19	1,732	449	501,236	1,824,668	1,885,767	4,455,488	61%	7,328,262
DMGV106C	1,234,961	1,040,636	29	1,981	461	539,585	1,874,588	2,001,987	4,692,241	61%	7,693,758
DMGV118C	1,022,891	861,330	26	1,696	386	469,733	1,552,375	1,799,301	3,908,437	59%	6,574,880
DMGV18B	1,445,228	1,188,197	18	2,207	519	606,358	2,260,023	2,269,120	5,502,550	61%	9,043,646
DMGV13C	1,153,101	989,731	38	2,019	474	561,124	1,794,679	2,004,122	4,501,166	59%	7,572,124
DMGV13B	1,275,360	1,104,033	26	2,275	545	633,496	1,943,490	2,599,971	4,959,225	56%	8,814,984
DMGV92C	1,036,506	884,226	31	1,717	345	453,202	1,643,340	1,697,333	4,019,367	59%	6,849,546
DMGV29C	1,191,003	1,004,666	39	2,106	481	563,600	1,904,067	2,153,864	4,665,962	60%	7,838,174
DMGV48C	1,177,344	980,764	43	1,900	413	514,430	1,859,796	2,006,495	4,534,690	58%	7,794,014
DMGV47C	1,226,031	1,055,507	38	2,126	474	595,295	1,890,750	2,317,729	4,770,221	58%	8,293,476
DMGV108C	1,202,867	989,725	31	2,120	451	549,642	1,938,405	1,752,886	4,683,241	60%	7,752,890
DMGV134C	1,527,578	1,272,066	34	2,366	589	656,862	2,344,811	2,448,933	5,804,306	61%	9,484,008
DMGV70C	1,427,451	1,178,429	29	2,028	500	604,699	2,202,607	2,090,416	5,415,743	63%	8,632,068
DMGV29B	1,247,391	1,032,444	32	1,916	463	546,650	1,970,763	1,978,745	4,799,659	61%	7,815,304
DMGV57C	1,158,083	956,151	47	1,741	390	488,728	1,806,298	1,677,197	4,411,438	62%	7,142,628
DMGV70B	1,430,335	1,222,908	30	2,272	519	618,605	2,237,907	2,327,076	5,512,576	60%	9,164,438
DMGV91C	1,047,497	887,754	18	1,857	417	482,420	1,709,950	1,704,904	4,129,913	58%	7,095,050
DMGV86C	1,170,761	1,020,519	33	2,065	492	558,316	1,784,445	2,237,466	4,536,631	57%	7,955,916
DMGV51C	1,245,700	1,004,493	30	1,945	412	517,312	1,909,609	1,759,666	4,679,501	61%	7,675,374
DMGV57B	1,235,093	1,026,295	30	1,906	476	527,673	1,924,012	1,874,728	4,715,485	61%	7,768,158
DMGV51B	1,217,633	966,456	46	1,857	323	522,215	2,053,284	1,589,776	4,761,814	61%	7,742,960
Average	1,190,638	998,656	30	1,930	446	530,040	1,862,929	1,955,106	4,584,669	60%	7,644,792
Average % of total reads	**16%**	**13%**		**-**	**-**	**7%**	**24%**	**26%**	**60%**		
Bait ratio	**17%**	**14%**				**11%**	**44%**		**86%**		

4.5.3 The effect of age at sampling on the unique sequence read counts

We expected that the unique sequence read count would be the same for all individuals if a typical sequence capture was achieved and that the age effect observed previously was due to the age-dependent telomere shortening. The telomere reads were excluded from the total read count and the results obtained are shown in Figure 4.18. The correlation between age at sampling and the unique sequences reads was calculated. A linear regression analysis of the unique sequences reads' correlation with age at sampling showed that $r^2 = 0.19$ and the P value = 0.03, which represents a significant correlation.

The data used to calculate the linear regression contained paired (dependent) samples, whereby an individual DNA sample was collected twice at different ages: B (collected at an early time point) and C (collected at a relatively older time point). The significant observation may be due to the presence of the dependent samples. In order to rule out this possibility, a similar analysis was conducted with either B or C samples. A linear regression analysis of the unique sequences reads which only involved B-samples showed that $r^2 = 0.26$ and the P value = 0.02 whereas a linear regression analysis of the unique sequences reads which only included C-samples showed that $r^2 = 0.21$ and the P value = 0.04. Both results showed that a significant effect was observed (Figure 4.19).

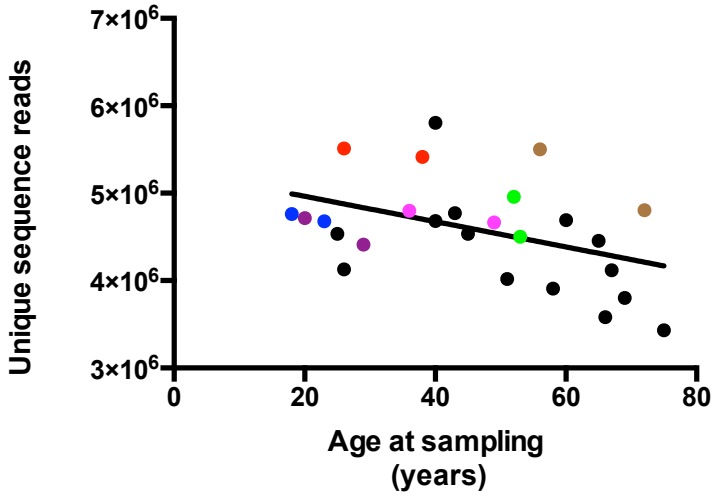


Figure 4.18: The effect of age at sampling on unique sequence read counts (paired samples are shown in colour).

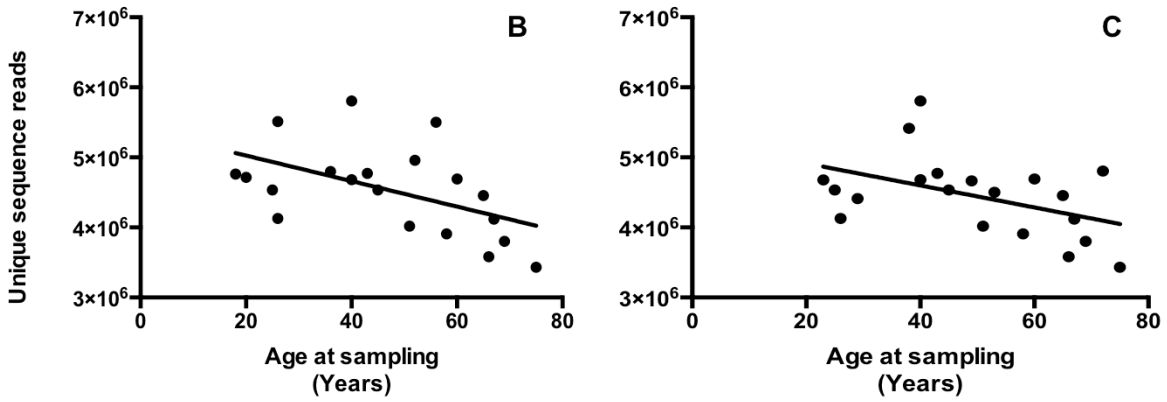


Figure 4.19: The effect of age at sampling on unique sequence read counts including either B- or C-samples of the paired samples.

4.5.3.1 The effect of sheared DNA size on the total and unique DNA sequence read count

The effect of sheared DNA size on the total reads captured was investigated to ensure that no sheared DNA size differences could contribute to the declining trend observed in the generated results, with respect to both total and unique DNA sequences. The linear regression analysis showed that $r^2 = 0.00$ and the P value = 0.98 for total reads and $r^2 = 0.007$ and the P value = 0.7 for unique DNA sequences (Figure 4.20). This indicated no significant effect of the sheared DNA size on either the total number of reads generated or the unique DNA sequences calculated.

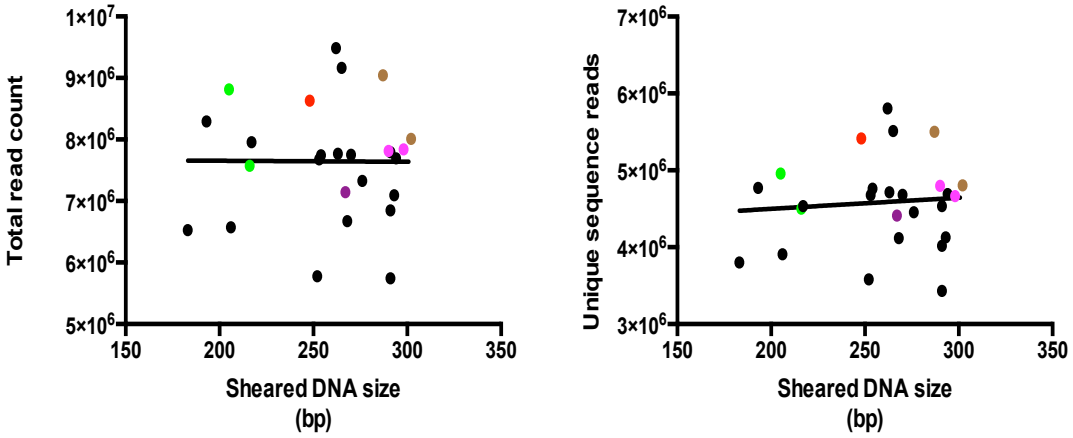


Figure 4.20: The effect of sheared DNA size on total reads count and unique DNA sequences reads (paired samples are shown in colour).

4.5.3.2 *The effect of finished library DNA size on the total and unique DNA sequence read counts*

To generate the finished library DNA, sheared DNA ends were repaired and adapters, primers, and index sequences were also added. Moreover, two PCR amplifications were conducted. This may have resulted in differences in the finished library DNA size, resulting in turn in a decrease in the total and unique sequence reads. So, the effect of finished library DNA size on the total and unique reads which were captured was also investigated. A linear regression analysis showed that $r^2 = 0.055$ and the P value = 0.25 for total read count and $r^2 = 0.12$ and the P value = 0.07 for unique reads were obtained (Figure 4.21). This indicated that no significant effect existed of finished library DNA size on the total number of unique DNA sequences reads generated. However, a declining trend was observed in both reads counts. Earlier, an ascending trend ($r^2 = 0.06$ and the P value = 0.23) was observed between the finished library DNA size and age at sampling. The quality of data generated on the Illumina sequencing platforms depends on constructing optimal cluster densities across every lane of every flow cell. Longer finished library DNA may fail to produce good quality clusters and may reduce the number of reads generated; this could be the reason for the observed declining trend in unique sequences reads with age at sampling.

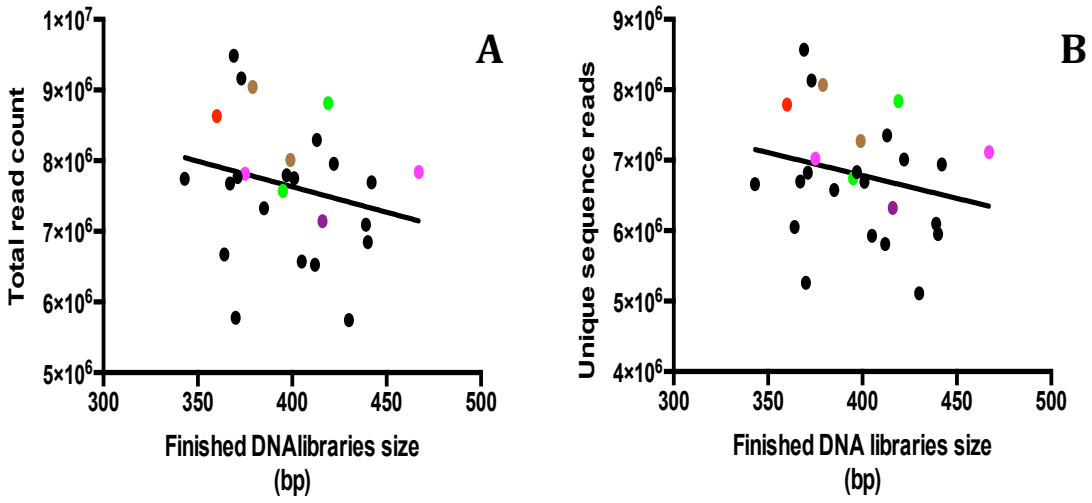


Figure 4.21: The effect of finished library DNA size. A): the effect of finished library DNA size on total reads count. **B):** the effect of finished library DNA size on unique DNA sequences reads.

4.5.4 Microsatellites

Previously, 25,539 pure repetitive sequences with unique 60 bp in both flanking sequences were selected. The chosen pure repetitive sequences were only mono-, di-, and trinucleotides and were identified by TRF in TRDB (hg19). Each microsatellite was covered by a single bait (1X coverage). Using Bowtie, the total reads generated using NGS were aligned to the human genome reference (HG19) in order to identify captured loci. The data showed that the proposed bait bridge capture strategy worked successfully, as Table 4.6 shows. The counts for the microsatellite were performed using Bowtie. The reads were mapped to the bait lengthened to 100 bp in both flanking sites and the count for each bait was recorded. Each read was forced to map only once. Bowtie 2, at the highest sensitivity option, is argued to be the best mapping method (Highnam *et al*., 2013).

Table 4.6: Aligned microsatellites reads to the human genome reference (HG19)

Sample ID	*Age at sampling (years)*	*Mono nucleotides*	*Di-nucleotides*	*Tri nucleotides*	*Total nucleotides read*
DMGV119C	75	276,214	1,042,385	77,127	1,395,726
DMGV18C	72	388,139	1,485,192	110,248	1,983,579
DMGV99C	69	263,593	1,148,067	81,007	1,492,667
DMGV129C	67	332,679	1,252,131	95,648	1,680,458
DMGV133C	66	291,532	1,077,705	84,613	1,453,850
DMGV22B	65	350,327	1,370,602	103,739	1,824,668
DMGV106C	60	355,634	1,413,364	105,590	1,874,588
DMGV118C	58	282,376	1,184,863	85,136	1,552,375
DMGV18B	56	430,081	1,701,103	128,839	2,260,023
DMGV13C	53	308,077	1,390,563	96,039	1,794,679
DMGV13B	52	308,562	1,529,415	105,513	1,943,490
DMGV92C	51	319,168	1,232,292	91,880	1,643,340
DMGV29C	49	351,108	1,448,143	104,816	1,904,067
DMGV48C	45	350,182	1,407,249	102,365	1,859,796
DMGV47C	43	308,234	1,482,237	100,279	1,890,750
DMGV108C	40	378,411	1,452,812	107,182	1,938,405
DMGV134C	40	468,269	1,741,968	134,574	2,344,811
DMGV70C	38	440,195	1,634,997	127,415	2,202,607
DMGV29B	36	372,593	1,487,940	110,230	1,970,763
DMGV57C	29	366,725	1,330,792	108,781	1,806,298
DMGV70B	26	444,040	1,663,407	130,460	2,237,907
DMGV91C	26	330,359	1,286,506	93,085	1,709,950
DMGV86C	25	311,377	1,375,680	97,388	1,784,445
DMGV51C	23	379,228	1,422,594	107,787	1,909,609
DMGV57B	20	382,309	1,428,884	112,819	1,924,012
DMGV51B	18	423,882	1,520,228	109,174	2,053,284

This was repeated for all the baits and for all the samples. Microsatellites were successfully captured using this strategy using only flanking sequences. The average microsatellite read count was 1,862,929; this represented 24% of the total read count and ~ 41% of unique DNA

sequences. The presence of a unique flanking sequence on both sides of the microsatellite helped to reduce mapping errors. Finally, uncovered bait sequences were identified for each sample, as shown in Table 4.7. The data obtained revealed how appropriate this approach was. The captured microsatellites were 4,491 mono-, 18,924 di-, and 1,234 trinucleotides. The alignment process provided an average coverage of ~80X. Different levels of coverage were observed for each loci in the NGS data obtained, which for some loci was as high as 1,819 reads, while other loci had no coverage at all. The 'bait' strategy worked as expected. The alignment process showed matches in ~97% of the loci. The overall quality of the alignment was very high. However, ~18% of the reads showed a total quality drop after the microsatellites or at the read end (Figure 4.22).

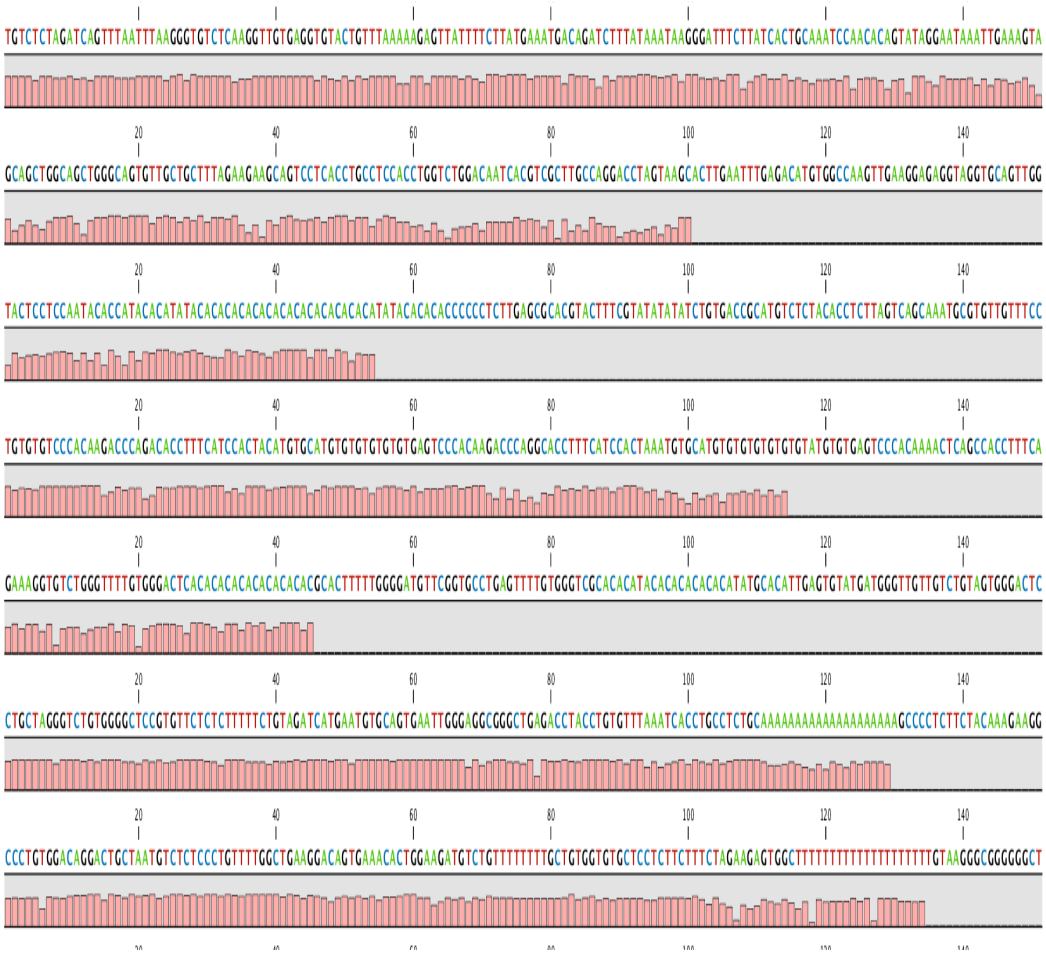


Figure 4.22: The microsatellite reads quality: The total quality drop after the microsatellites or at the end.

Table 4.7: Detailed analysis to identify the number of non-covered microsatellites

Sample ID	*Age at sampling (years)*	*Mono-aligned*	*Mono-non-covered*	*Di-aligned*	*Di non-covered*	*Tri-aligned*	*Tri-non-covered*	*Total aligned*	*Total covered (%)*
DMGV119C	75	4,489	23	18,845	876	1,226	80	24,560	96
DMGV18C	72	4,503	9	19,025	696	1,238	68	24,766	97
DMGV99C	69	4,483	29	19,115	606	1,224	82	24,822	97
DMGV129C	67	4,501	11	18,930	791	1,234	72	24,665	97
DMGV133C	66	4,453	59	18,940	781	1,230	76	24,623	96
DMGV22B	65	4,493	19	19,127	594	1,231	75	24,851	97
DMGV106C	60	4,490	22	18,959	762	1,231	75	24,680	97
DMGV118C	58	4,482	30	18,634	1,087	1,228	78	24,344	95
DMGV18B	56	4,503	9	18,936	785	1,234	72	24,673	97
DMGV13C	53	4,483	29	19,125	596	1,224	82	24,832	97
DMGV13B	52	4,481	31	18,645	1,076	1,223	83	24,349	95
DMGV92C	51	4,451	61	18,926	795	1,230	76	24,607	96
DMGV29C	49	4,495	17	18,962	759	1,247	59	24,704	97
DMGV48C	45	4,494	18	19,118	603	1,235	71	24,847	97
DMGV47C	43	4,488	24	19,137	584	1,226	80	24,851	97
DMGV108C	40	4,496	16	19,212	509	1,238	68	24,946	98
DMGV134C	40	4,506	6	18,581	1,140	1,243	63	24,330	95
DMGV70C	38	4,497	15	18,632	1,089	1,241	65	24,370	95
DMGV29B	36	4,501	11	19,131	590	1,241	65	24,873	97
DMGV57C	29	4,501	11	19,093	628	1,227	79	24,821	97
DMGV70B	26	4,499	13	19,009	712	1,244	62	24,752	97
DMGV91C	26	4,500	12	18,651	1,070	1,240	66	24,391	96
DMGV86C	25	4,486	26	18,877	844	1,237	69	24,600	96
DMGV51C	23	4,494	18	18,991	730	1,230	76	24,715	97
DMGV57B	20	4,499	13	18,733	988	1,233	73	24,465	96
DMGV51B	18	4,499	13	18,966	755	1,236	70	24,701	97

The selected non-covered microsatellites were further analysed using IGV. The majority of these microsatellites were found to have high levels of AT rich patterns and / or A/T rich flanking sequences. Sonication was used to fragment the DNA, a process which could raise the temperature of the DNA solution to above ~20°C. It is possible that the small DNA generated during processing, may become denatured due to this heating, especially those fragments with a high AT content. Single-stranded fragments would not ligate to the double-stranded adapters, resulting in their under-representation in the final library (Illumina, Catalog # PE-930-1001, 2011).

4.5.4.1 The effect of age at sampling on microsatellite read count

Microsatellites are polymorphic, as they vary in length among individuals. Shorter microsatellites can be captured and sequenced more efficiently compared to longer ones. As the length of microsatellites is not expected to change substantially over time, no significant correlation between age at sampling and microsatellites read count is expected. However, the effect of age at sampling on microsatellite reads was investigated, and the results are shown in Figure 4.23. Surprisingly, the data showed a significant correlation between the age at sampling and the microsatellite read count. A linear regression analysis showed that $r^2 = 0.22$ and the P value = 0. 016. This indicated that the number of microsatellite reads decreased as age at sampling increased.

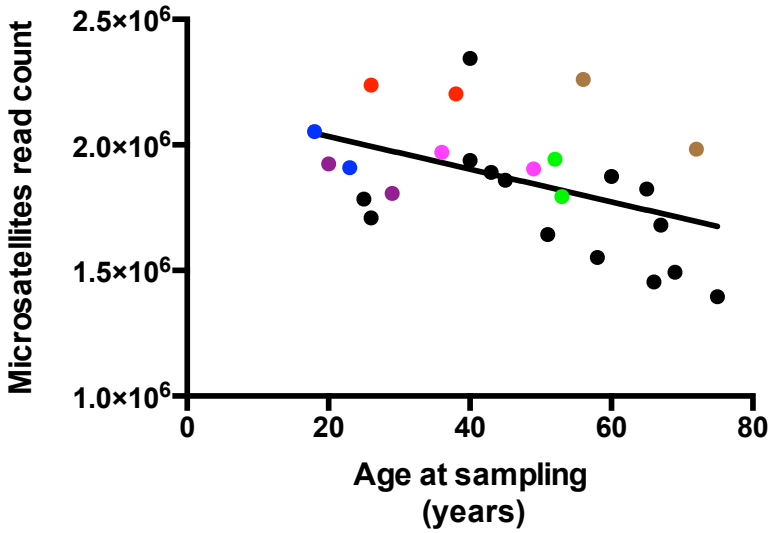


Figure 4.23: the effect of age at sampling on microsatellite reads

4.5.4.2 The effect of sheared DNA and finished DNA library size on microsatellite reads

The total microsatellite read count showed an age-associated decrease. Two PCR steps were used to generate the finished DNA library. PCR amplification can generate errors when DNA sequences include microsatellites, especially for mononucleotides and dinucleotides (Clarke *et al*., 2001). These errors can create products with different alleles. To exclude the possibility of any technical errors or unwanted artifacts being generated that may contribute to the result observed, the effect of both sheared DNA and finished DNA library size on microsatellite reads was studied, as Figure 4.24 shows.

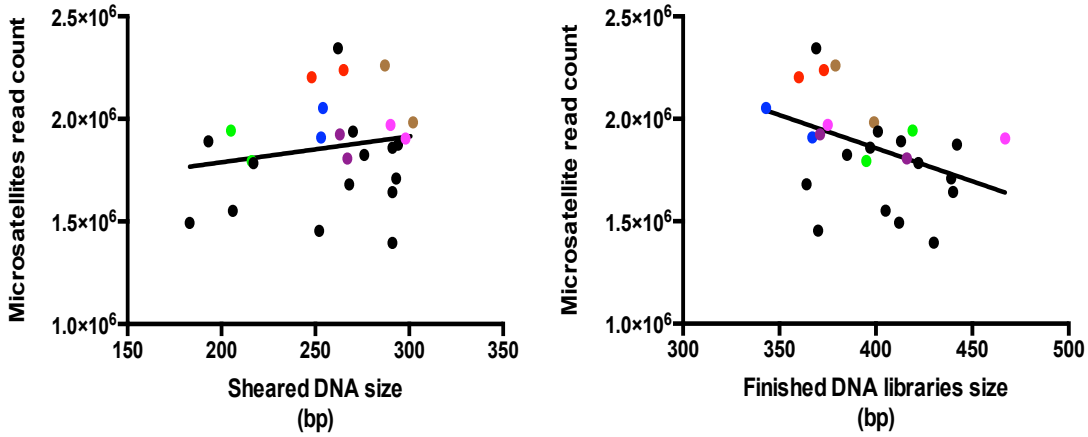


Figure 4.24: The effect of both sheared DNA and finished DNA library size on microsatellite reads.

The data showed no significant correlation with microsatellite read counts when the sheared DNA size linear regression was analysed, and the result obtained was $r^2 = 0.03$ and the P value = 0.4. However, a linear regression analysis of the finished DNA library size showed a marginally significant correlation with microsatellite reads. The result was $r^2 = 0.17$ with a P value = 0.04 which indicated that the number of microsatellites decreased when the finished DNA library size increased. Previously, we showed that the finished DNA library size increased with age. This may explain the significant decrease in microsatellite read counts with increased age, due to the increased size of the finished DNA library.

4.5.5 Relative microsatellite read length and age at sampling

In order to confirm that the number of microsatellites decreased when the finished DNA library size increases (which may explain the significant decrease in microsatellites read count with age), the autosome gene reads were used to normalise and calculate the relative microsatellite read count. Then, the relative microsatellite read count was correlated with age at sampling, as shown in Figure 4.25. The result of the linear regression ($r^2 = 0.12$ and the P value = 0.08) showed no significant correlation between relative microsatellite read length and age at sampling.

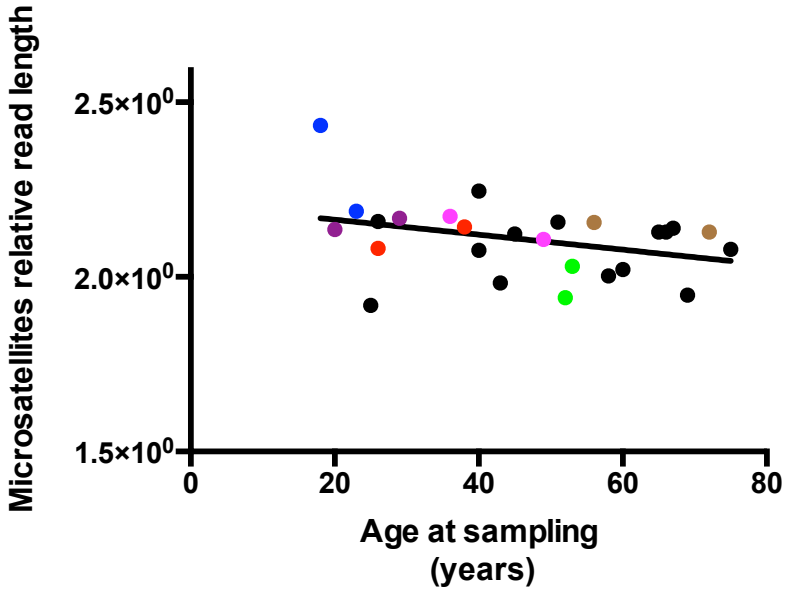


Figure 4.25: Microsatellites relative read length and age at sampling

4.5.5.1 Microsatellite genotyping

Microsatellite genotypes must be allocated in terms of allele length or the number of sequenced bases within a read which separate the unique flanking regions aligned to the reference. Furthermore, reads must span an entire repeat length in order to be accurately genotyped. In this experiment, 11 microsatellites were randomly selected to investigate the suitability of the captured reads for genotyping. Among those included were: 3-A-112, 4-T-184, 6-A-250, 7-C-105, 10-A-563, and 11-A-676 as mononucleotides; 1-AC-578, 9-AT-415, 13-AC-241, 14-CA-687, 15-GT-384, and 16-TC-781 as dinucleotides; 2-CTG-133, 5-ACA-996, 12-ATC-272, 18-TGT-417, 19-TAA-360, 20-TTG-477, and 21-TCA-390 as trinucleotides and finally seven genetic markers, D3S1358, D5S818, CSF1PO, D7S820, D8S1179, D13S317, and D16S539, as tetranucleotides. A CLC Genomic Workbench was

used to identify the genotype for the selected microsatellites; the map reads to reference tools were employed in doing so.

4.5.5.1.1 Tetranucleotide genotyping

Seven microsatellites were selected as tetranucleotides, from the Combined DNA Index System (CODIS). All the selected microsatellites were analysed (Table 4.9), but only D16S539 will be discussed in detail here. A custom reference was generated using the 100 bp flanking sequence in both sides, with a single repeat addition from 0 to 25 repeats. Custom references were generated for each locus to overcome misalignment that may be generated when using HG19 due to gap formation arise from allele length variation among individuals tested as HG19 encompass only one reference for each locus location. By adjusting the length fraction (70%) and similarity fraction (90%) parameters, we were able to genotype all the 7 microsatellites by counting the reads that map successfully to the custom reference, as shown in Figure 4.26.

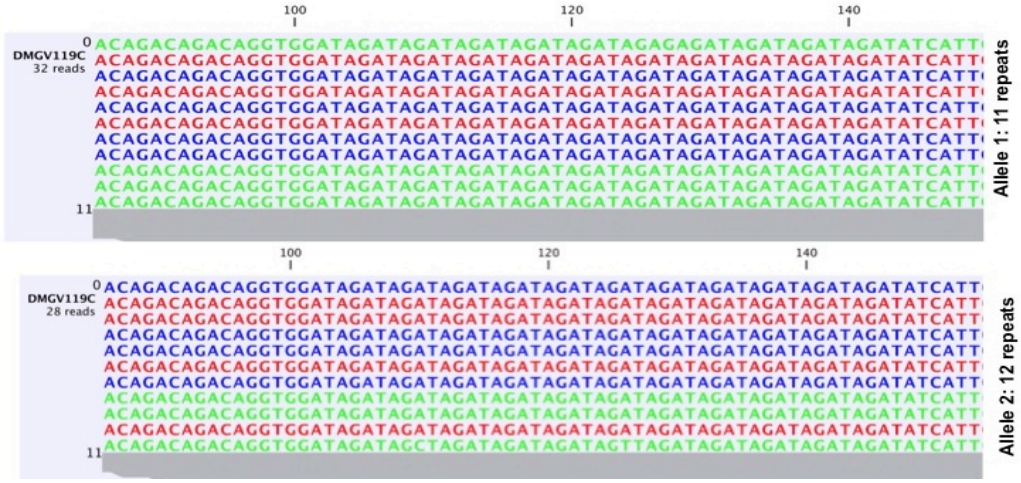


Figure 4.26: D16S539 mapped reads for sample DMGV119C: Image obtained from CLC Genomic workbench analysis.

In sample 119C, the mapped repeat count starts at a position of 101 bp where the D16S539 pattern starts (GATA). Thirty-two reads were mapped to the custom reference with 11 repeats, and 28 reads were mapped to the custom reference with 12 repeats. Only one read was mapped to the custom reference with 10 repeats (see Table 4.8). This table shows the trios, paired samples (collected at two different time points) and samples collected at single time points. Only the data of these samples will be shown in the analysis of the other microsatellites.

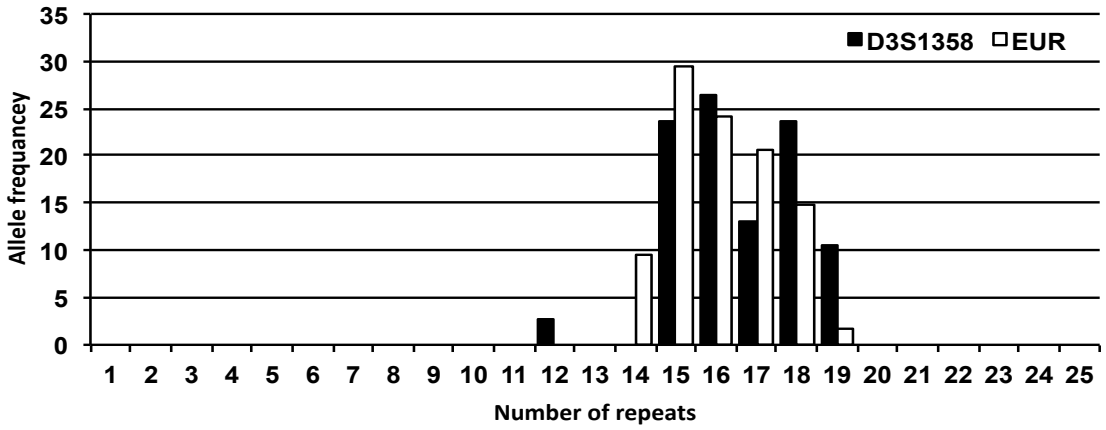
Table 4.8:Tetranucleotide reads matching the custom reference count for D16S539

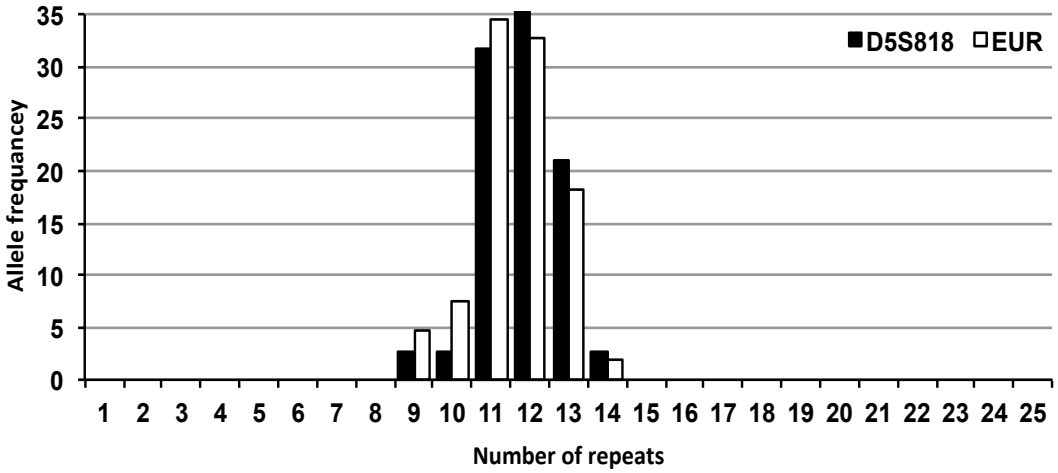
	Sample ID	DMGV 119C	DMGV 129C	DMGV 133C	DMGV 134C	DMGV 51C	DMGV 51B
Custom reference repeat count	Number of repeats	-	Father	Mother	Son	C	B
	6	0	0	0	0	0	0
	7	0	0	0	0	0	0
	8	0	0	2	0	0	0
	9	0	22	31	0	0	0
	10	1	0	0	0	0	0
	11	32	0	3	1	0	0
	12	28	0	28	49	3	6
	13	0	0	0	1	72	125
	14	0	28	0	57	1	0
	15	0	0	0	0	0	0
	16	0	0	0	0	0	0
	Genotype	11/12	9/14	9/12	12/14	13/13	13/13

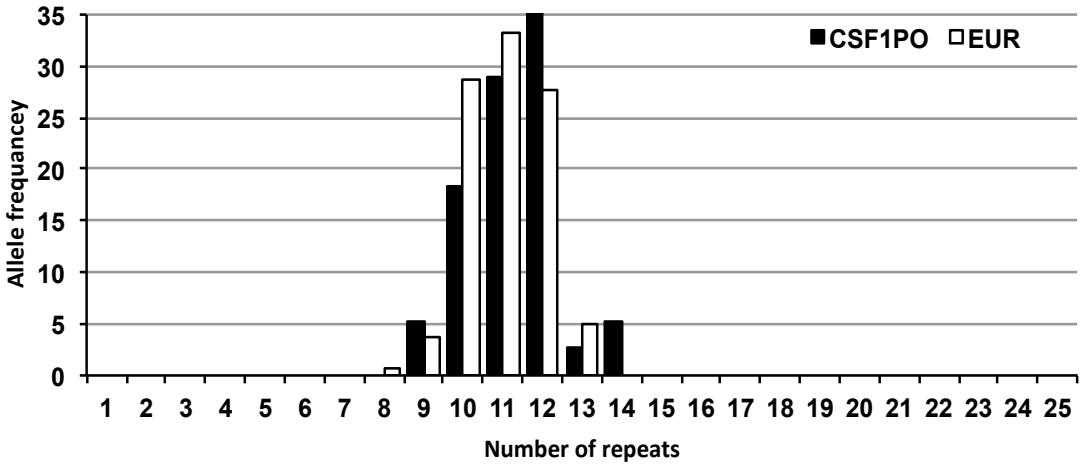
The custom references showing the highest read counts were selected to be the individual allele and the genotype for 119C was predicted to be (11/12) (Table 4.8). In some samples, a few reads were generated due to PCR slippage that shortened the allele length mapped to the custom reference (such as in 133C), but this did not compromise our ability to predict the correct genotype. The analysis was conducted for all 26 samples, and it proved easy to genotype all the seven tetranucleotides.

The allele frequency was also calculated for the 19 individuals in the seven tetranucleotides selected, after excluding one sample from each pair and the son from the trios, and comparing them with European data generated from the analysis of global variability in 15 traditional and five new European Standard Set (ESS) STRs using the CEPH human genome diversity panel (Phillips *et al*., 2011). The European population analysis was conducted on 158 individuals (Phillips *et al*., 2011). In D16S539, eight alleles were observed, which ranged between 8-15 repeats length with 80% heterozygosity, which is identical to the data generated by Phillips *et al*. (see Figure 4.27). Both fit within the general population range of 5-16 in 19 alleles seen (Butler, 2005). In D3S1358, six alleles were observed ranging between 12-20 repeats length whereas in the Phillips *et al*. study, data were identified for a similar number of alleles but ranging between 14-20 repeats. Both allele ranges lay within the population range (8-21) where 20 alleles were seen (Butler, 2005). In our data, a single allele with a length of 12 repeats was observed in sample 48C (12/19). The same allele was observed once in an African population for the same locus with an allele frequency (0.005) (Phillips *et al*., 2011). For D8S1179, eight alleles were observed in our data, ranging between 9-16, with 74% heterozygosity, while the European data had showed 10 alleles observed for the same loci ranging between 8-17

and both fit within the population range 7-20, with 84% heterozygosity, where 15 alleles were seen (Butler, 2005). This difference in allele ranges may have been due to the difference in sample size between the studies. Similar analysis was conducted on the other four tetranucleotides (D5S818, CSF1PO, D7S820, and D13S317,) and the results are shown in Figure 4.27. The genotypes of these microsatellites were identified and the results are shown in (Table 4.9).







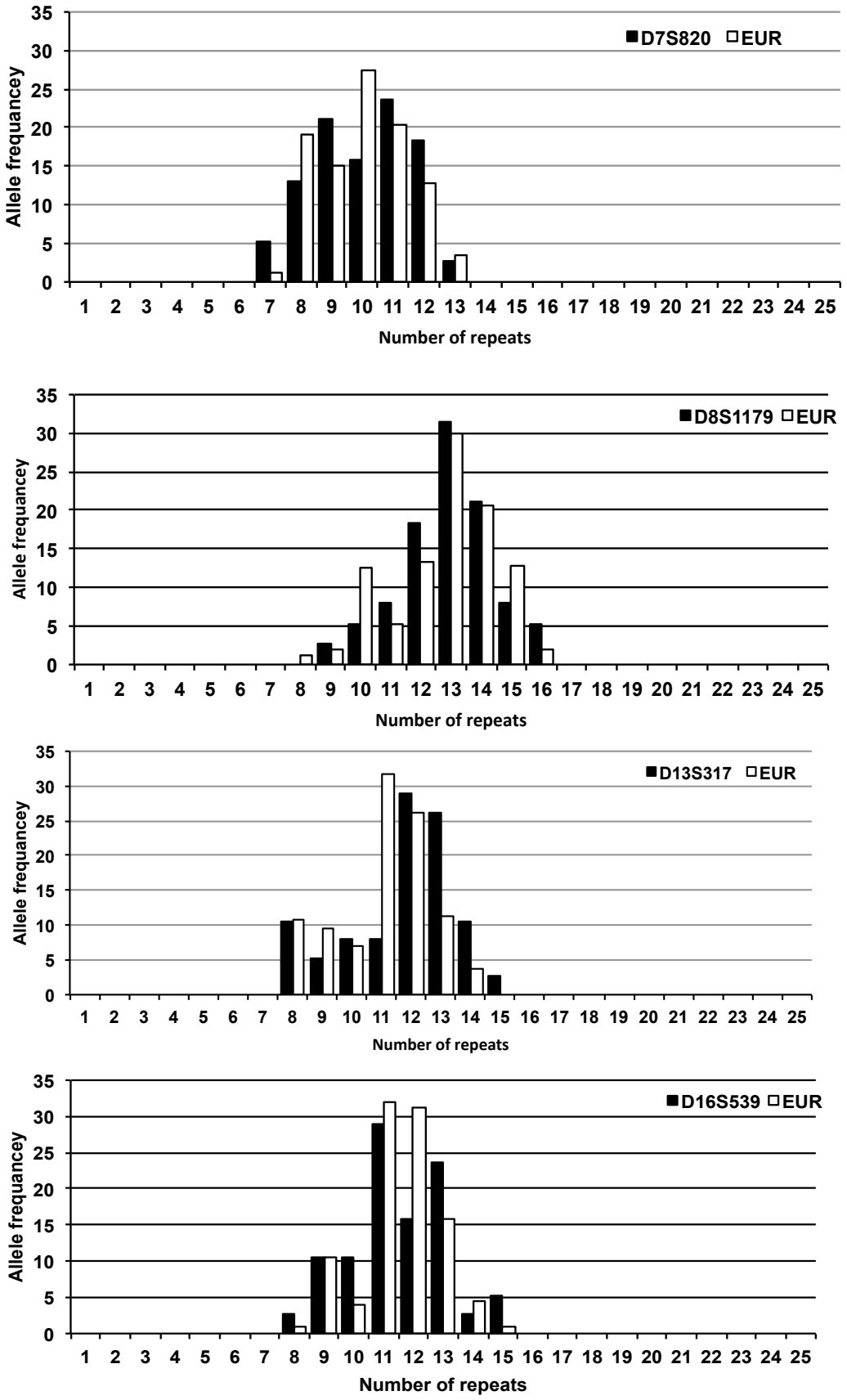


Figure 4.27: CODIS loci allele distributions: The black bars show the observed allele distribution for the three selected CODIS markers and the white bars show the European population's allele frequency for these loci.

Table 4.9: Tetranucleotide genotypes

	D3S1358	D5S818	CSF1PO	D7S820	D8S1179	D13S317	D16S539
Motif	**CTAT**	**TCTA**	**CTAT**	**TCTA**	**CTAT**	**ATCT**	**GATA**
DMGV119C	17/17	12/11	12/11	11/8	15/16	15/12	11/12
DMGV18C	15/19	11/10	12/10	10/9	13/15	14/13	11/13
DMGV18B	15/19	11/10	12/10	10/9	13/15	14/13	11/13
DMGV99C	15/16	12/11	12/9	10/10	13/14	14/12	10/11
DMGV129C	17/18	12/9	13/11	11/9	12/13	13/12	9/14
DMGV133C	16/16	11/11	12/12	11/8	11/13	14/13	9/12
DMGV134C	16/17	11/9	12/11	11/8	12/13	13/12	12/14
DMGV22B	16/19	13/12	11/11	11/8	13/14	12/10	9/11
DMGV106C	18/18	12/11	10/10	12/11	10/13	12/10	9/12
DMGV118C	15/16	14/13	12/12	12/9	13/14	8/8	11/12
DMGV13C	15/18	13/11	12/11	9/7	12/12	13/11	10/10
DMGV13B	15/18	13/11	12/11	9/7	12/12	13/11	10/10
DMGV92C	15/15	11/11	12/12	13/12	12/14	13/12	11/15
DMGV 91C	15/16	12/11	14/14	12/10	14/14	13/12	13/15
DMGV48C	12/19	13/13	11/11	11/8	13/14	13/12	13/13
DMGV29C	16/19	12/12	12/11	10/10	10/13	14/11	8/11
DMGV 29B	16/19	12/12	12/11	10/10	10/13	14/11	8/11
DMGV47C	15/18	13/12	12/11	9/7	13/15	13/8	11/12
DMGV70C	17/18	13/12	11/10	9/8	11/13	12/10	11/12
DMGV 70B	17/18	13/12	11/10	9/8	11/13	12/10	11/12
DMGV 108C	16/18	12/12	12/10	12/11	11/12	13/8	13/13
DMGV 57C	16/17	12/11	10/9	12/11	14/16	12/11	10/11
DMGV 57B	16/17	12/11	10/9	12/11	14/16	12/11	10/11
DMGV 86C	18/18	12/11	12/11	12/11	9/13	9/9	11/13
DMGV 51C	15/16	13/12	12/10	9/9	12/12	13/12	13/13
DMGV 51B	15/16	13/12	12/10	9/9	12/12	13/12	13/13
Allele range	**12-19**	**9-14**	**9-14**	**7-13**	**9-16**	**8-15**	**8-15**
Heterozygosity (%)	**74**	**75**	**65**	**85**	**84**	**80**	**80**

4.5.5.1.2 Trinucleotide genotyping

Similarly, seven trinucleotides were selected (2-CTG-133, 5-ACA-996, 12-ATC-272, 18-TGT-417, 19-TAA-360, 20-TTG-477, and 21-TCA-390) and 31 custom references were generated with a 60 bp flanking sequence on both sides for each loci selected. The reads that were aligned to the custom references were counted (Figure 4.28).

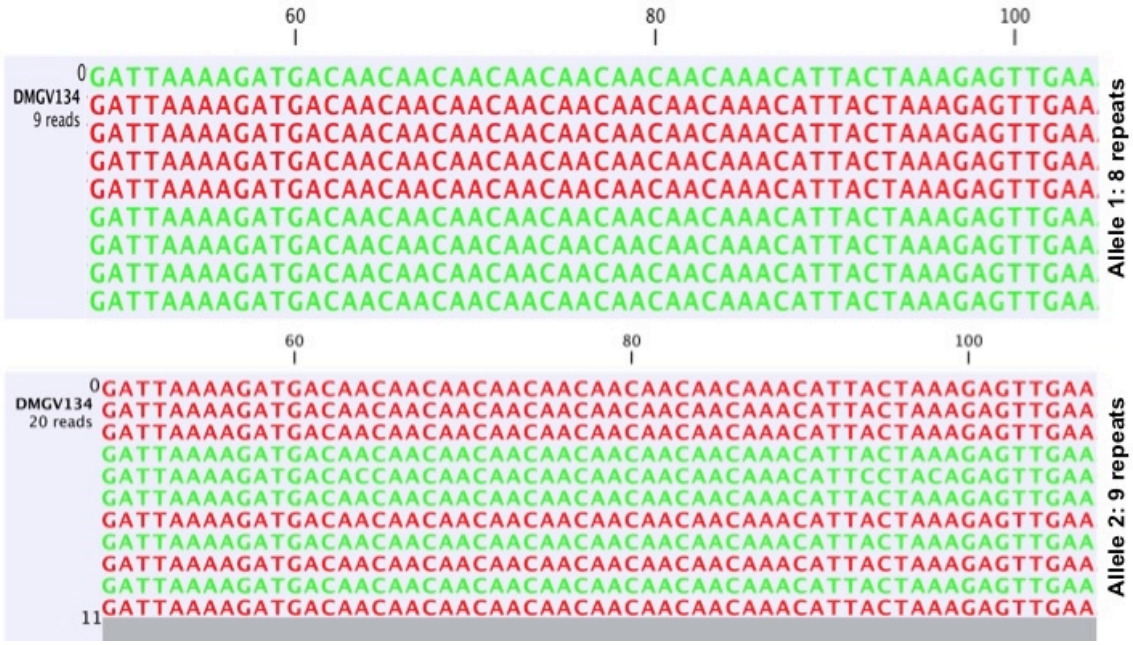


Figure 4.28: 5-ACA-996 aligned reads to the custom reference for sample 134

In sample 134C, the repeat count starts at the position of 61 bp where the 5-ACA-996 pattern (ACA) starts. Nine reads were mapped to the custom reference with eight repeats, and 20 reads were mapped to the custom reference with nine repeats. The predicted observed genotype was (8/9).

The same analysis was conducted on all 26 samples and the data obtained from 5-ACA-996 for five individuals are shown (an individual sampled once, and trios and an individual sampled at two time points) in Table 4.10.

Table 4.10: Trinucleotides read count mapped to the custom reference for 5-ACA-996

	Number of repeats*	DMGV 119C	DMGV 129C	DMG V133C	DMGV 134C	DMGV 51C	DMGV 51B
Custom reference repeat count			Father	Mother	Son	C	B
	6	0	0	0	0	0	0
	7	0	0	0	0	0	0
	8	0	6	12	9	10	16
	9	16	8	0	20	0	0
	10	0	0	0	0	0	0
	11	0	0	0	0	0	0
	Genotype	**9/9**	**8/9**	**9/9**	**8/9**	**8/8**	**8/8**

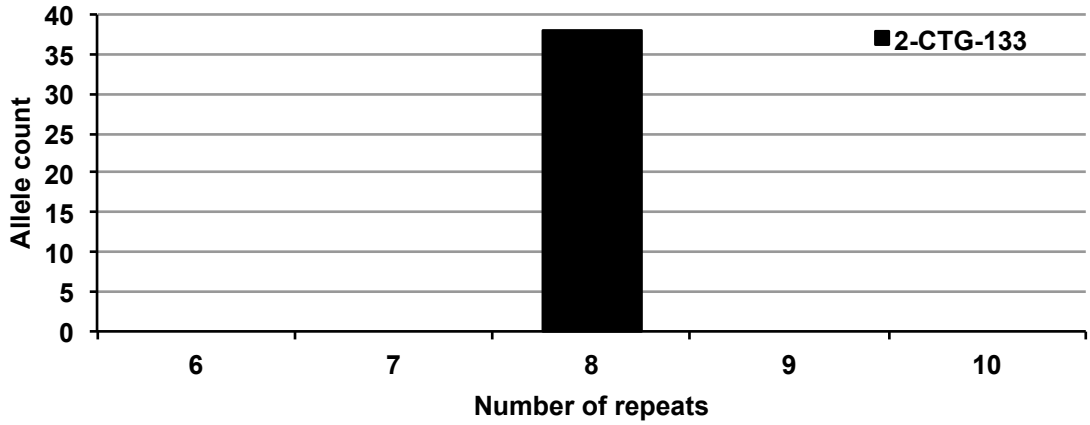
* Number of repeats present in custom reference.

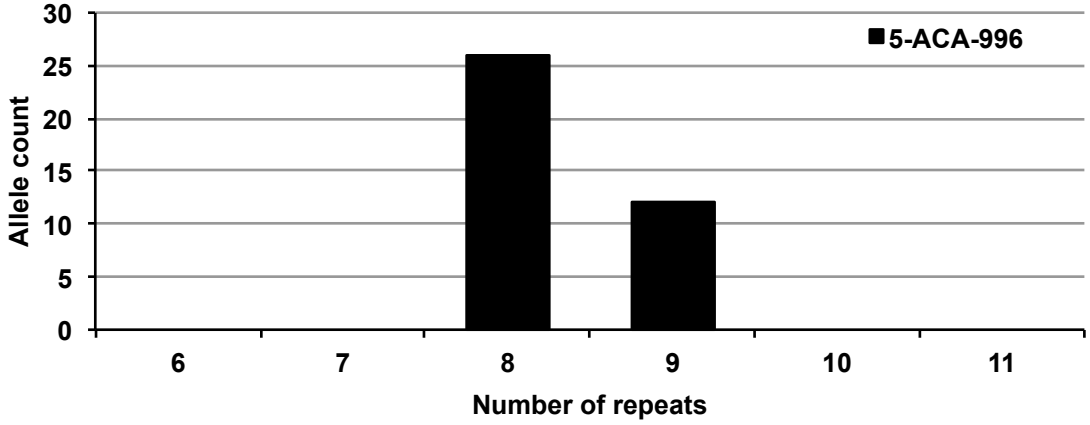
Genotyping trinucleotides were also straightforward and the genotype for 20 individuals was predicted in 26 samples. The observed allele frequency in 5-ACA-996 was calculated and the results are shown in Table 4.11.

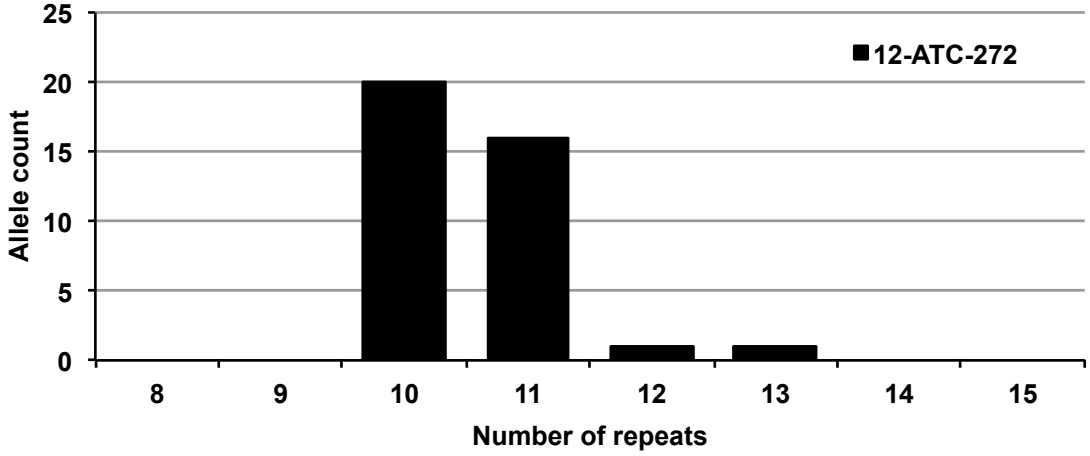
Table 4.11: allele frequency observed in 5-ACA-996

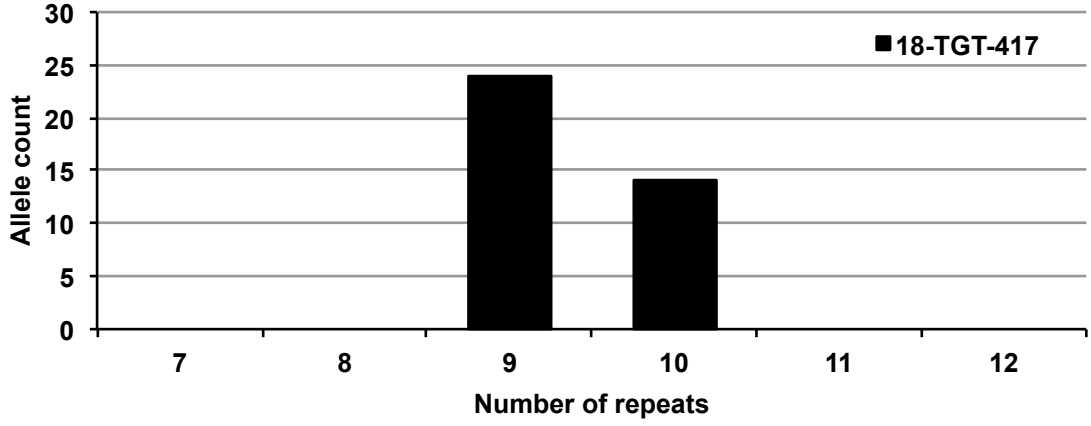
Alleles	Allele count	Frequency observed
8	26	0.68
9	11	0.32
Total	**38**	**1**

Only two alleles were observed (8 and 9) in the 26 samples tested. The minor allele frequency was 0.32 in allele with a length of eight repeats. Four alleles were observed in 12-ATC-272 and they ranged between 10 to 13 repeats. In 2-CTG-133, all samples were homozygous, with an identical single allele observed with eight repeats. A similar analysis was conducted on another four loci 18-TGT-417, 19-TAA-360, 20-TTG-477, and 21-TCA-390. The allele distribution for the seven trinucleotides selected is shown in Figure 4.29. The results showed that two alleles were observed in 18-TGT-417 and three alleles were observed in 19-TAA-360. The maximum number of alleles was seen in 20-TTG-477 with five alleles ranging 7-14 repeat, whilst for 21-TCA-390, three alleles were observed. The trinucleotide genotyping was conducted in all samples tested and the results are shown in Table 4.12.


2-CTG-133
Allele count
40
35
30
25
20
15
10
5
0
6
7
8
9
10
Number of repeats


5-ACA-996
Allele count
30
25
20
15
10
5
0
6
7
8
9
10
11
Number of repeats


12-ATC-272
Allele count
25
20
15
10
5
0
8
9
10
11
12
13
14
15
Number of repeats


18-TGT-417
Allele count
30
25
20
15
10
5
0
7
8
9
10
11
12
Number of repeats

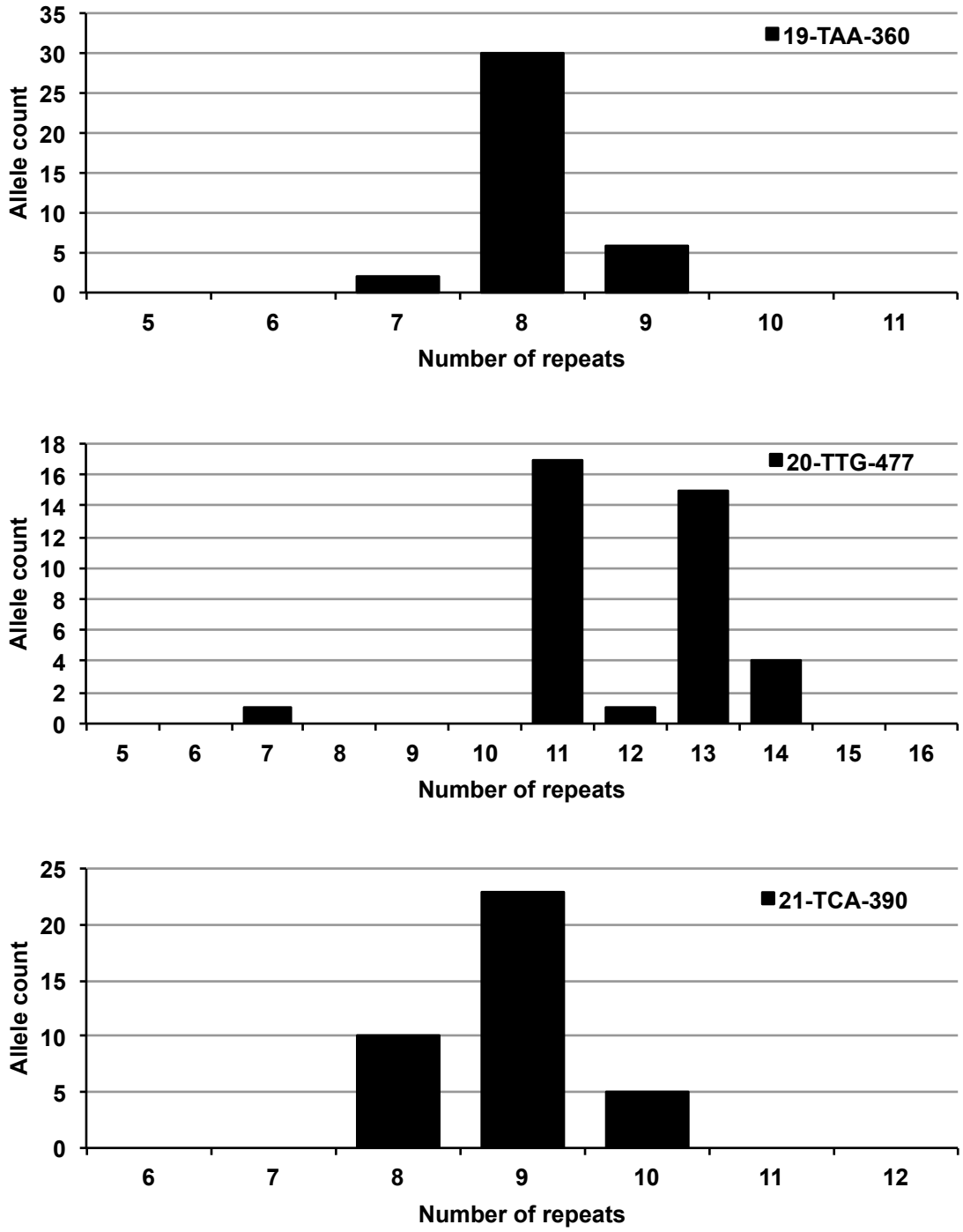


Figure 4.29: Trinucleotide allele distributions: the alleles observed in 2-CTG-133, 5-ACA-996, 12-ATC-272, 18-TGT-417, 19-TAA-360, 20-TTG-477, and 21-TCA-390.

Table 4.12: Trinucleotide genotypes

Microsatellites ID	5-ACA-996	12-ATC-272	2-CTG-133	18-TGT-417	19-TAA-360	20-TTG-477	21-TCA-390
Motif	**ACA**	**ATC**	**CTG**	**TGT**	**TAA**	**TTG**	**TCA**
DMGV119C	9/9	10/12	8/8	10/9	9/8	11/11	10/8
DMGV18C	8/8	10/11	8/8	9/9	8/8	11/11	9/9
DMGV18B	8/8	10/11	8/8	9/9	8/8	11/11	9/9
DMGV99C	8/8	10/10	8/8	9/9	8/8	11/11	9/9
DMGV129C	8/9	10/11	8/8	9/9	8/8	13/13	9/8
DMGV133C	8/8	10/13	8/8	10/9	9/8	11/11	9/8
DMGV134C	8/9	10/10	8/8	10/9	9/8	13/11	8/8
DMGV22B	8/8	10/10	8/8	10/10	8/8	13/13	9/9
DMGV106C	9/9	11/11	8/8	10/9	9/8	13/11	9/9
DMGV118C	8/8	10/11	8/8	9/9	8/8	11/7	10/8
DMGV13C	8/8	11/11	8/8	9/9	8/8	14/13	9/8
DMGV13B	8/8	11/11	8/8	9/9	8/8	14/13	9/8
DMGV92C	8/9	10/10	8/8	10/10	8/8	13/11	10/10
DMGV 91C	8/9	10/11	8/8	10/10	8/8	13/13	10/9
DMGV48C	8/8	10/10	8/8	9/9	8/8	14/11	9/8
DMGV29C	8/9	10/11	8/8	10/10	8/8	11/11	9/8
DMGV 29B	8/9	10/11	8/8	10/10	8/8	11/11	9/8
DMGV47C	8/8	10/11	8/8	10/9	8/8	13/13	9/9
DMGV70C	8/9	11/11	8/8	9/9	9/9	12/11	9/9
DMGV 70B	8/9	11/11	8/8	9/9	9/9	12/11	9/9
DMGV 108C	8/8	10/11	8/8	10/9	8/7	13/13	9/8
DMGV 57C	9/9	10/11	8/8	10/9	8/8	14/11	9/9
DMGV 57B	9/9	10/11	8/8	10/9	8/8	14/11	9/9
DMGV 86C	8/9	10/11	8/8	9/9	8/7	13/13	8/8
DMGV 51C	8/8	10/11	8/8	9/9	9/8	14/11	9/9
DMGV 51B	8/8	10/11	8/8	9/9	9/8	14/11	9/9
Allele range	**8-9**	**10-12**	**8-8**	**9-10**	**7-9**	**7-14**	**8-10**
Heterozygosity (%)	**32**	**63**	**-**	**35**	**35**	**45**	**50**

Paired samples shaded blue; Trios samples shaded yellow; Mother and son samples shaded orange

4.5.5.1.3 Dinucleotide genotyping

Seven dinucleotides were also selected (1-AC-578, 9-AT-415, 13-AC-241, 14-CA-687, 15-GT-384, 16-TC-781 and 17-TC-319). A custom reference was generated using the 60 bp flanking sequence on both sides. Thirty-one custom references were generated for each locus. CLC Genomic workbench was used to align the reads to the custom reference. The data obtained from 1-AC-578 analyses are shown in Figure 4.30.

Figure 4.30: 1-AC-578 mapped reads to the custom reference in 134C.

The reads aligned to the custom references were counted for all 26 samples tested. The data obtained from the analysis of 1-AC-578 for five individuals are shown in Table 4.13.

Table 4.13: 1-AC-578 read mapped to the custom reference

	Sample ID	DMGV119C	DMGV129C	DMGV133C	DMGV134C	DMGV51C	DMGV51B
Custom reference repeat	No. repeats		Father	Mother	Son	C	B
	7	0	0	0	0	0	1
	8	0	2	0	0	5	5
	9	0	100	1	122	202	170
	10	0	0	35	0	0	0
	11	0	0	0	0	0	0
	19	0	0	0	0	0	0
	20	0	0	0	0	0	0
	21	0	0	2	6	0	0
	22	39	0	3	20	0	0
	23	0	0	0	0	0	0
	24	0	0	3	0	0	0
	25	0	0	0	0	0	0
	Genotype	22/22	9/9	10/22	9/22	9/9	9/9

More PCR slippage was observed in the seven dinucleotides which were tested. Nevertheless, a good genotype can be predicted. All the samples appeared to have been genotyped successfully, but when the trio was analysed, sample 133C was problematic. The sample showed that 35 reads had mapped to custom reference with ten repeats but only a few reads were mapped to the custom references with 21, 22, and 24 repeats. The predicted genotype based on this observation was a homozygous (10/10). However, the son showed reads which mapped to custom references with 9 and 22 repeats, for which the predicted genotype was (9/22). The father was homozygous for allele 9 and the reads mapped only to the custom reference with 9 repeats; no other reads were mapped to other references. This showed that the son should inherit the 9 allele from his father and the 22 from his mother. Turning to the number of reads in the son, 122 reads were mapped to custom reference custom references with nine and 20 reads only mapped to custom references with 22 repeats. A huge reduction in the read count was therefore observed (84%). This pattern of a low read count mapped to the bigger allele was observed in other samples, for the same loci analysed. This showed that the actual genotype for the mother is (10/22). We could not reliably call the genotype for the mother because of the low read count mapped to the bigger allele. This showed that shorter alleles were captured or sequeced more efficiently and that large dinucleotide alleles are more prone to PCR slippage, and a broader mapped read spread with large alleles was observed, and that slippage has occurred between molecules and was generated during PCR steps before generating the DNA cluster. However, we also observed errors as the cluster generated in large alleles will not pass the quality threshold, and the read quality will drop. Moreover, full spanning reads are required to achieve a good genotyping and large alleles show a lower spanning read count than short alleles (Table 4.13).

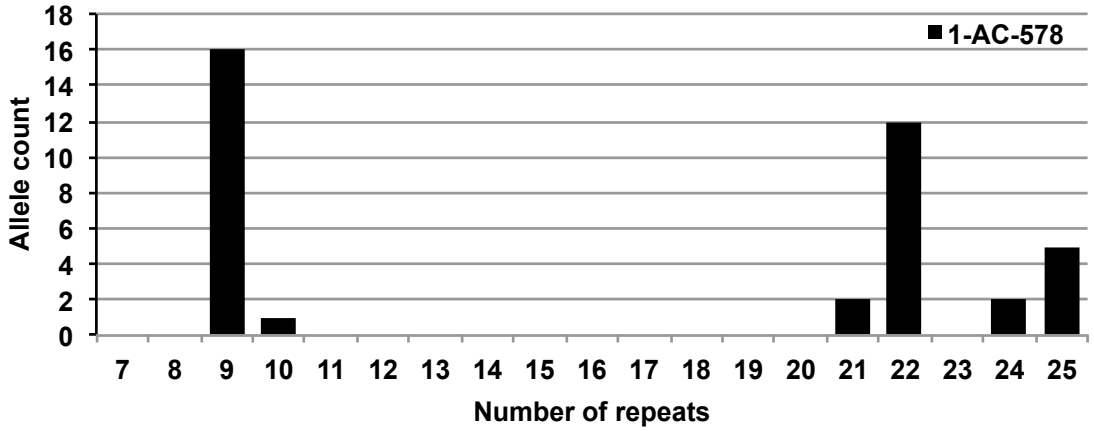
The allele frequency was calculated for 1-AC-578 and 6 alleles were observed, ranging between nine and ten repeats (Table 4.14).

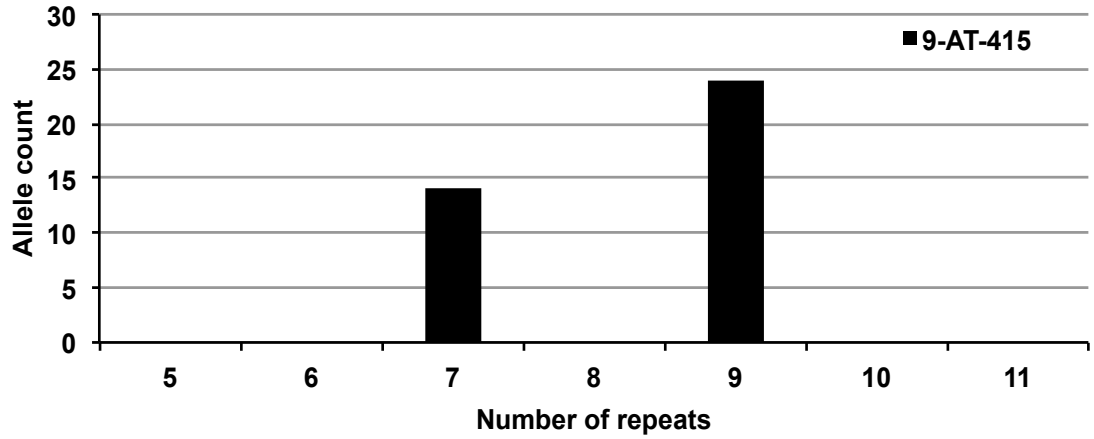
Table 4.14 the observed allele frequency for the 1-AC-578 locus

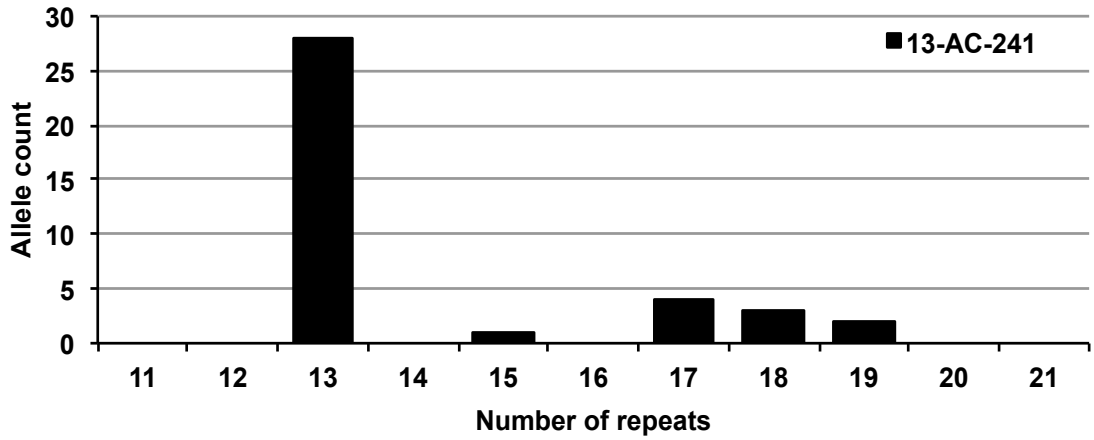
Alleles	Allele count	Frequency observed
9	16	0.42
10	1	0.03
21	2	0.05
22	12	0.32
24	2	0.05
25	5	0.13
Total	**38**	**1**

The allele frequency distribution of the seven dinucleotides selected is shown in Figure 4.31. For the 16-TC-781, five different alleles were observed ranging between 11 and 19

repeats. In 9-AT-415, only two alleles were observed which had seven and nine repeats. The dinucleotides genotyping was conducting in all samples tested and the result was shown in Table 4.15.







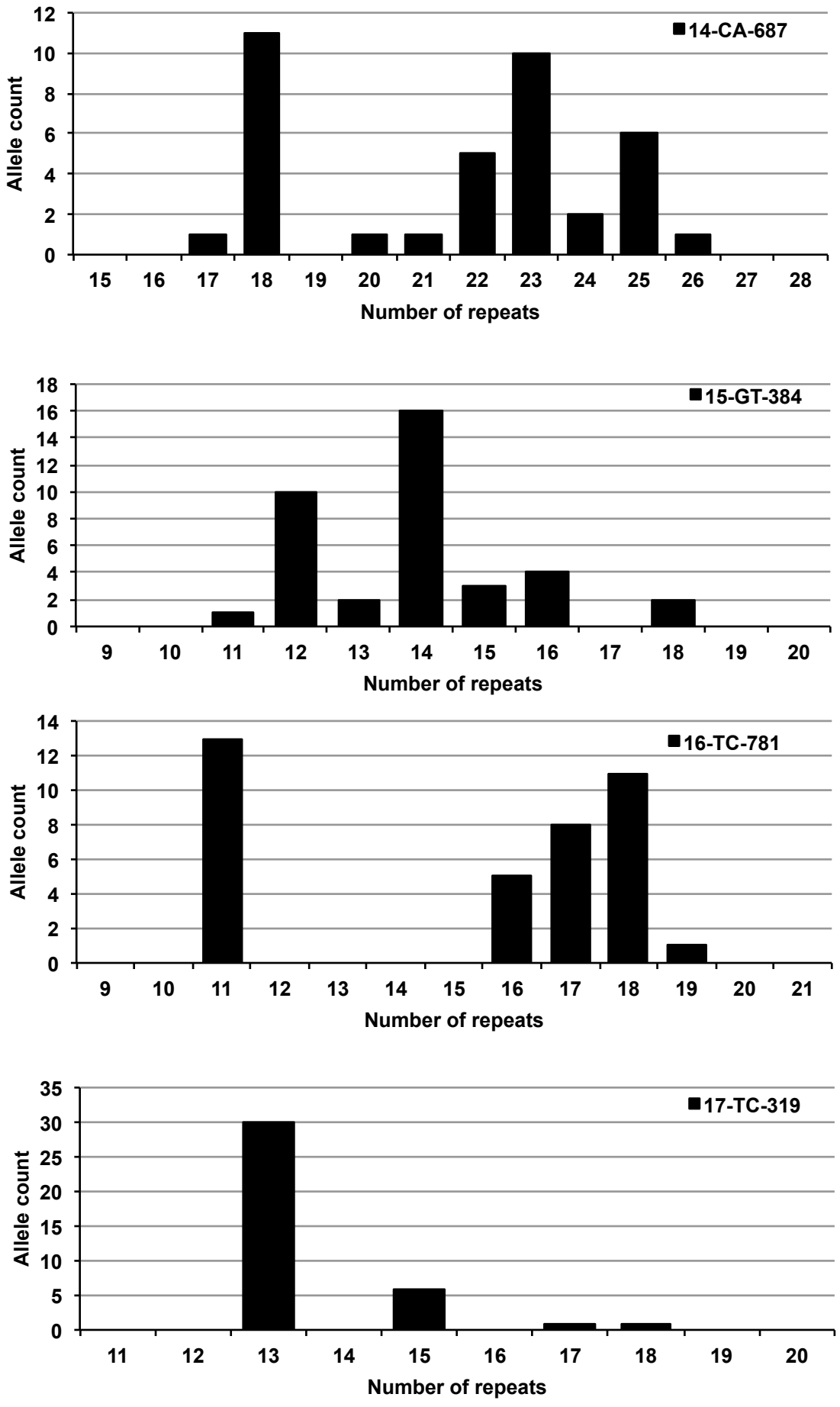


Figure 4.31: The allele distribution for the dinucleotides selected.

Table 4.15: Dinucleotide genotypes

Microsatellites ID	1-AC-578	16-TC-781	9-AT-415	13-AC-241	14-CA-687	15-GT-384	17-TC-319
Motif	AC	TC	AT	AC	CA	GT	TC
DMGV119C	22/22	11/11	9/9	13/13	23/22	14/14	13/13
DMGV18C	9/9	18/18	7/9	18/17	23/18	12/12	18/13
DMGV18B	9/9	18/18	7/9	18/17	23/18	12/12	18/13
DMGV99C	21/21	18/18	7/9	13/13	23/22	14/12	15/13
DMGV129C	9/9	16/17	7/9	13/13	25/23	15/12	15/13
DMGV133C	10/22	18/18	9/9	18/17	24/23	14/12	13/13
DMGV134C	9/22	17/18	7/9	18/13	25/24	12/12	15/13
DMGV22B	24/25	11/11	7/9	13/13	26/18	15/11	13/13
DMGV106C	22/25	11/11	7/9	13/13	23/22	16/14	17/13
DMGV118C	9/22	16/16	9/9	17/13	23/18	18/14	13/13
DMGV13C	9/25	17/17	7/9	13/13	20/18	16/14	13/13
DMGV13B	9/25	17/17	7/9	13/13	20/18	16/14	13/13
DMGV92C	9/9	11/16	7/9	13/13	25/22	14/12	15/13
DMGV 91C	9/22	11/11	7/9	13/13	25/23	14/13	13/13
DMGV48C	9/25	18/18	7/9	13/13	25/18	14/12	15/13
DMGV29C	9/25	17/19	7/9	19/18	24/18	16/14	13/13
DMGV 29B	9/25	17/19	7/9	19/18	26/18	16/14	13/13
DMGV47C	9/22	17/17	9/9	19/13	25/21	18/12	13/13
DMGV70C	9/22	11/11	7/9	13/13	18/18	14/12	13/13
DMGV 70B	9/22	11/11	7/9	13/13	18/18	14/12	13/13
DMGV 108C	9/22	17/18	9/9	13/13	23/17	16/14	13/13
DMGV 57C	22/24	16/17	7/9	15/13	18/18	14/13	15/13
DMGV 57B	22/24	16/17	7/9	15/13	18/18	14/13	15/13
DMGV 86C	22/22	11/18	7/9	13/13	25/18	15/14	15/13
DMGV 51C	9/9	11/18	7/9	17/13	23/22	14/12	13/13
DMGV 51B	9/9	11/18	7/9	17/13	23/22	14/12	13/13
Allele range	**9-25**	**11-19**	**7-9**	**13-19**	**17-26**	**11-18**	**13-18**
Heterozygosity (%)	**63**	**37**	**74**	**40**	**90**	**80**	**45**

Paired samples shaded blue; Trios samples shaded yellow; Mother and son samples shaded orange

4.5.5.1.4 Mononucleotide genotyping

Six mononucleotides (3-A-112, 4-T-184, 6-A-250, 7-C-105, 10-A-563, and 11-A-676) were selected. Thirty one custom references were generated with 60 bp flanking sequence in both sides for each of the loci selected. The reads were aligned to the custom reference (Figure 4.32).

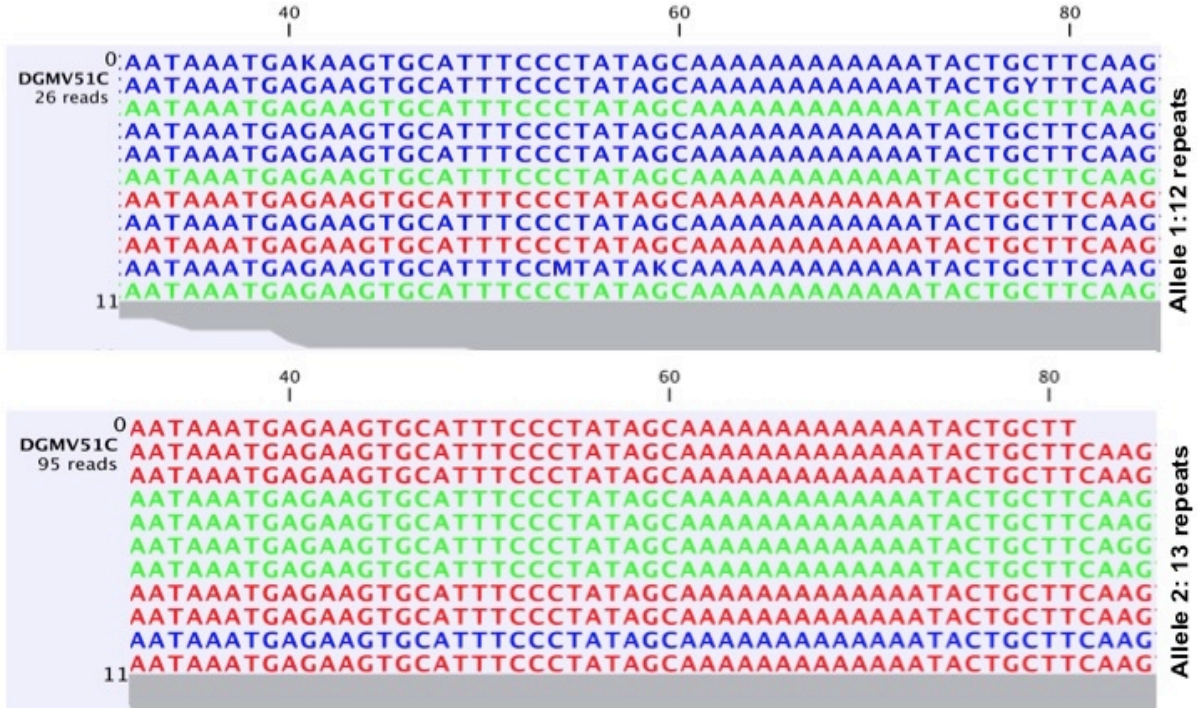


Figure 4.32: 3-A-112 reads mapped to the custom references in sample 51C . The repeats count starts at position 61.

The mononucleotides showed more PCR slippage, and, in only a few cases were reliable genotypes observed. In samples 51B and 51C we were able to predict the genotype (12/13). But in other samples, reasonable read numbers with border spread, as in sample 134C, was observed Table 4.16. With the current approach, the genotyping of mononucleotides failed to produce any reliable inference in the majority of cases Table 4.17.

In connection with this, a stringent parameter significantly reduces the mapped reads and by contrast a relaxed parameter tends to detect non-specific reads.

Table 4.16: Mononucleotide read count mapped to the custom reference for 3-A-112 in 5 individuals

Number of repeats	DMGV 119C	DMGV 129C	DMGV 133C	DMGV 134C	DMGV 51C	DMGV 51B
		Father	Mother	Son	C	B
10	0	0	0	0	0	0
11	0	2	0	0	4	0
12	0	7	0	0	26	17
13	0	34	0	0	95	89
14	0	3	0	1	1	4
15	0	1	0	0	3	3
16	0	0	0	0	2	0
17	1	0	0	0	0	1
18	1	0	0	2	0	0
19	0	0	3	2	0	0
20	2	0	3	3	0	0
21	3	0	2	5	0	0
22	3	1	6	4	0	0
23	4	2	3	9	0	0
24	2	3	5	16	0	0
25	3	4	4	13	0	0
26	2	2	7	11	0	0
27	1	1	2	15	0	0
28	1	2	3	5	0	0
29	0	3	1	4	0	0
30	0	1	0	0	0	0

In addition to the success of the bait bridge strategy to capture microsatellites, the analysis of the NGS data generated from this data were also used to develop a new strategy of genotyping microsatellites by our research group (Carlos del Ojo Elias and John Cole). Software called 'SatNav' has been developed to genotype the microsatellites and to help us to understand the evolution of these tandem repeats. The software has showed a good level of performance for trinucleotide repeats. Similarly, low performance was observed with regard to mononucleotide repeats. The data investigation also revealed coverage problems resulting from sequencing depth; different coverage for each locus was identified in some loci to be high as 800x and other loci had no coverage at all. Their microsatellite data showed that the flanking and nearby DNA sequences could be a crucial factor in these extreme examples of coverage differences. The data also showed that variable repeats and SNPs surrounding microsatellites could affect both the sequencing and the alignment process.

Table 4.17: Mononucleotide genotypes

Microsatellites	3-A-112	4-T-184	6-A-250	7-C-105	10-A-563	11-A-676
Motif	**A**	**T**	**A**	**C**	**A**	**A**
DMGV119C	N	N	24/25	N	16/15	13/10
DMGV18C	N	N	24/23	N	16/15	13/10
DMGV18B	N	N	24/22	N	16/16	13/10
DMGV99C	13/13	N	23/22	N	16/16	14/10
DMGV129C	13/13	N	23/24	N	16/15	13/10
DMGV133C	N	N	N	N	N	13/10
DMGV134C	N	16/16	24/22	N	16/16	13/10
DMGV22B	14/14	N	23/23	N	16/16	13/10
DMGV106C	13/13	16/16	21/21	N	16/16	13/10
DMGV118C	13/14	N	N	N	16/15	13/10
DMGV13C	13/13	N	23/23	N	16/15	13/10
DMGV13B	13/13	N	23/23	N	16/15	13/10
DMGV92C	13/13	N	24/24	N	16/15	13/10
DMGV 91C	13/13	16/16	N	N	16/15	13/10
DMGV48C	14/14	N	23/22	N	16/15	13/10
DMGV29C	13/13	16/16	29/29	N	16/16	14/11
DMGV 29B	13/13	16/15	29/29	N	16/16	13/10
DMGV47C	N	16/16	N	N	16/15	13/10
DMGV70C	N	16/16	N	N	16/15	13/10
DMGV 70B	N	16/16	29/29	N	16/16	13/10
DMGV 108C	13/13	N	25/25	N	16/16	13/10
DMGV 57C	13/13	16/16	23/21	N	16/15	13/10
DMGV 57B	13/13	N	23/23	N	16/16	13/10
DMGV 86C	N	16/16	N	N	17/16	13/10
DMGV 51C	12/13	16/16	24/23	N	16/15	13/10
DMGV 51B	12/13	16/16	24/23	N	16/15	13/10

*Paired samples shaded blue; Trios samples shaded yellow; Mother and son samples shaded orange.

4.5.6 DNA repair gene captured reads

DNA repair genes were included in this experiment as they have unique sequences and will be used when estimating age-dependent telomere shortening (see Chapter Five). Specifically, the exons of 233 genes were selected for capture using 8,252 baits to cover 3,118 exons. The target enrichment approach successfully captured the targeted genes regions and Bowtie was used to align reads to match bait positions. These were counted and are shown in Table 4.18.

Table 4.18: DNA Repair Genes NGS reads count*

Sample ID	*Age at sampling (years)*	*Gene read counts*	*Autosome gene reads*	*X-chromosome reads*	*Autosome/Gene read*
DMGV119C	75	750,937	726,045	24,892	97%
DMGV18C	72	1,040,758	1,006,365	34,393	97%
DMGV99C	69	848,069	819,907	28,162	97%
DMGV129C	67	888,269	859,047	29,222	97%
DMGV133C	66	795,857	747,431	48,426	94%
DMGV22B	65	974,835	943,397	31,438	97%
DMGV106C	60	1,040,636	1,006,294	34,342	97%
DMGV118C	58	861,330	832,304	29,026	97%
DMGV18B	56	1,188,197	1,149,794	38,403	97%
DMGV13C	53	989,731	956,709	33,022	97%
DMGV13B	52	1,104,033	1,066,055	37,978	97%
DMGV92C	51	884,226	829,697	54,529	94%
DMGV29C	49	1,004,666	971,075	33,591	97%
DMGV48C	45	980,764	948,691	32,073	97%
DMGV47C	43	1,055,507	1,019,652	35,855	97%
DMGV108C	40	989,725	957,069	32,656	97%
DMGV134C	40	1,272,066	1,230,412	41,654	97%
DMGV70C	38	1,178,429	1,140,108	38,321	97%
DMGV29B	36	1,032,444	999,015	33,429	97%
DMGV57C	29	956,151	924,856	31,295	97%
DMGV70B	26	887,754	858,211	29,543	97%
DMGV91C	26	1,222,908	1,183,857	39,051	97%
DMGV86C	25	1,020,519	986,108	34,411	97%
DMGV51C	23	1,004,493	971,734	32,759	97%
DMGV57B	20	1,026,295	992,673	33,622	97%
DMGV51B	18	966,456	934,814	31,642	97%
	Average read count	**998,656**	**963,897**	**34,759**	**96%**

* Females are highlighted

The autosome gene reads were calculated by excluding any reads derived from the X-chromosome, in order to eliminate any dosage effect. The average gene reads represent 13% of the total reads and 22% of the unique DNA sequence, figures that are similar to the bait ratio of 14% of the total reads but which deviate from the bait ratio of unique DNA

sequence ~17 %. The autosome gene reads represent 97 % of the total gene reads mapped. The number of genes reads mapped to chromosome X is double in females (due to the dosage effect). Then, it follows that the sex of the sample tested can be predicted from Table 4.18 and DMGV133 and DMGV92 are females whereas the other samples were males, which is true.

4.5.6.1 The effect of age at sampling on the DNA repair genes

The gene reads were used to calculate the telomere age shortening and the gene count is expected to vary randomly between individuals, with no age effect. The effects of age at sampling on gene reads were calculated (Figure 4.33). The age at sampling has no significant effect on either the total DNA repair gene ($r^2 = 0.15$ and the P value = 0.054) or on the autosome gene read counts ($r^2 = 0.15$ and the P value = 0.052). However, an age-dependent declining trend was observed in both read counts.

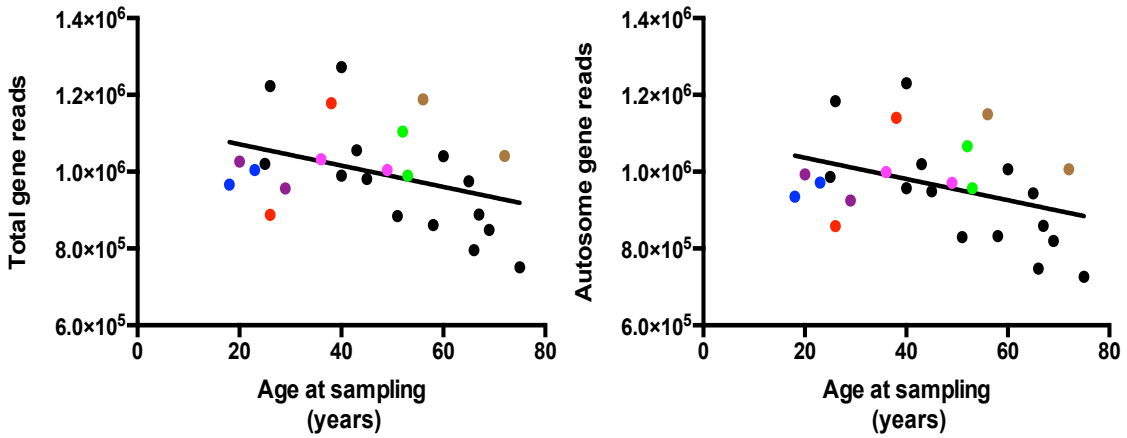


Figure 4.33: The effect of the age at sampling on the total and autosome gene read counts

The effect of sheared DNA size and finished DNA library size were analysed in order to ensure that the declining trend observed in gene read counts is not age related, as represented in Figure 4.34. It can clearly be seen that the sheared DNA size has no significant effect on the total DNA repair gene, as a linear regression showed that $r^2 = 0.00$ and the P value = 0.9. Similarly, the total DNA repair gene reads showed no significant effect on the finished DNA library size; the results obtained from a linear regression were $r^2 = 0.00$ and the P value = 0.7. The same result was demonstrated for the autosome gene read counts.

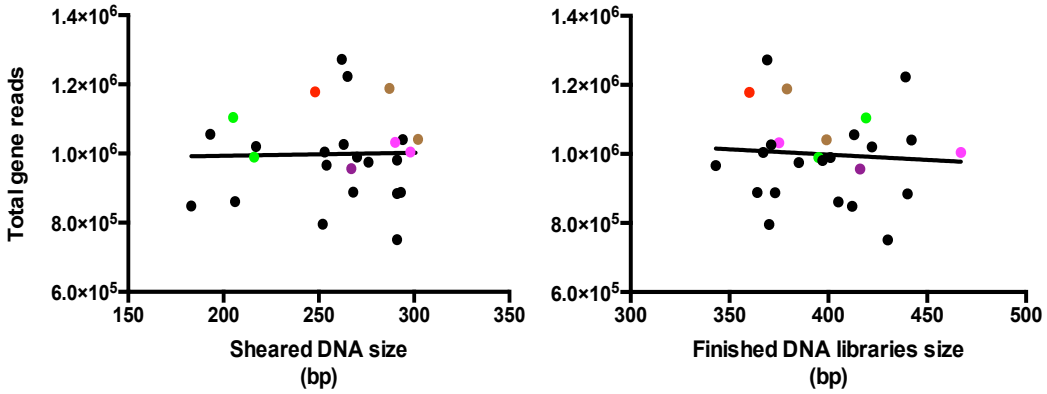


Figure 4.34: Linear regression analysis to investigate the effect of sheared DNA size and finished DNA library size on gene reads.

4.5.7 Human SNP genotyping

In this experiment, we included thousands of human SNPs (n=15,026) that were successfully captured using baits as having a single mismatch to ensure an equal capture for both alleles that could exist at a single SNP. Similar to microsatellites, twelve SNPs were randomly selected and IGV was used to visualise and count the reads mapped to the SNP bait sequence. A genotyping analysis was conducted to investigate the quality of the reads which were generated to examine the efficiency of SNP coverage for genotyping.

The results showed mean 86 times coverage for selected SNPs. The highest mean read-coverage was observed for rs3855951 (with 107 reads) and the lowest mean was observed for rs267193 (with 43 reads), as shown in Figure 4.35.

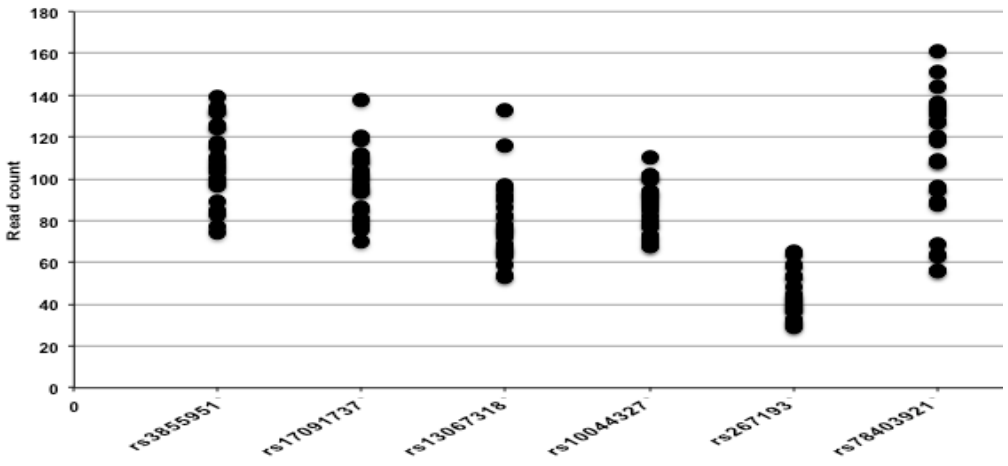


Figure 4.35: SNP read coverage. The read-coverage for the selected SNPs in all 26 samples tested (only six SNPs are shown).

The high coverage of SNP reads facilitates this genotyping. The genotyping analysis was conducted in all 12 SNPs, however, for the sake of brevity, only the analysis of rs17091737 is now discussed.

The bases in the reads associated with the SNPs were counted and the observed total of reads was calculated for each of the samples tested. In homozygotes, all the bases in the reads aligned with the SNP's position and were expected to be a 100% match, whereas in heterozygotes, the base in reads covered the SNPs position was expected to be a 50% match for each allele (see Figure 4.36 and Figure 4.37).

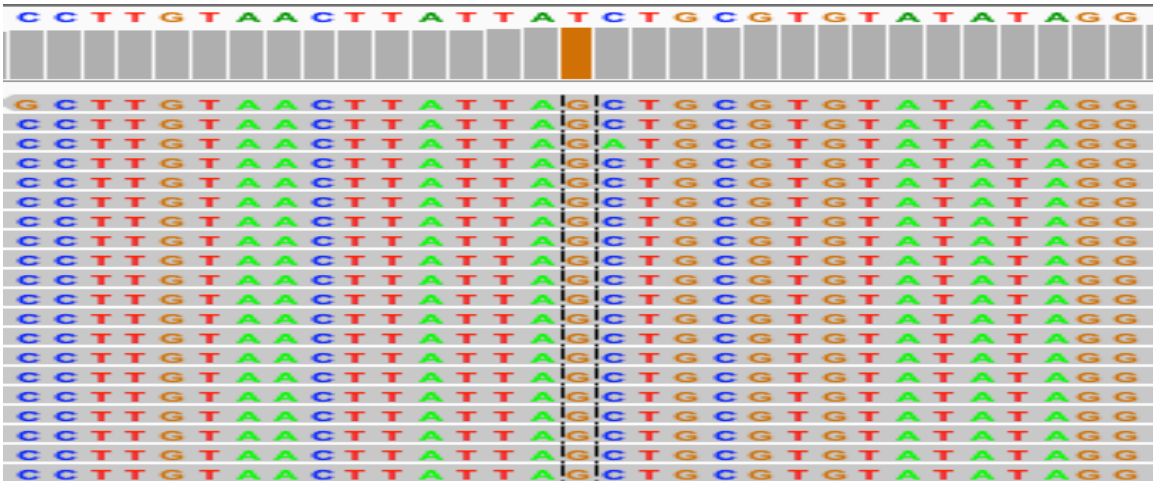


Figure 4.36: Homozygote: Number of bases in reads has 100% match the SNP genotype.

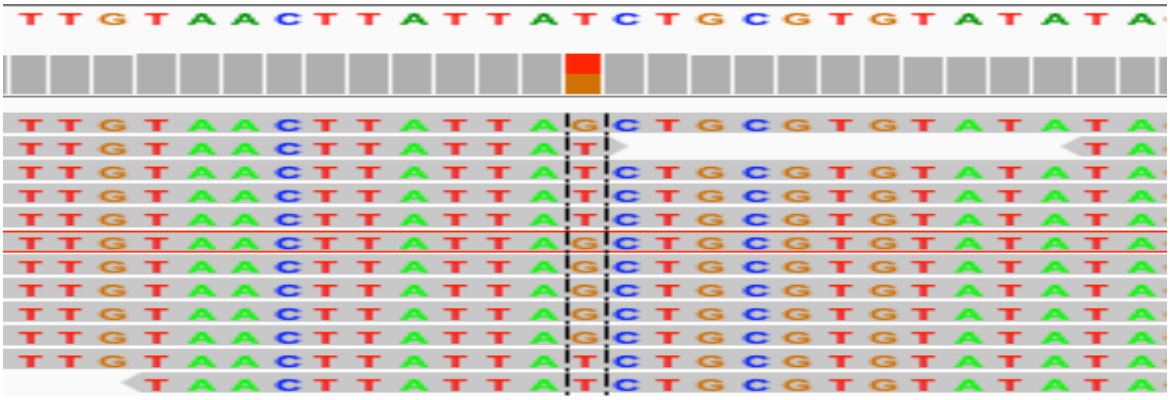


Figure 4.37: Heterozygote: The number of base in reads expected to be 50% match for each allele.

In an ideal scenario, the observed total reads should match the expected total count, but in some samples we observed deviations in the observed reading from the expected count, as shown in Figure 4.38.

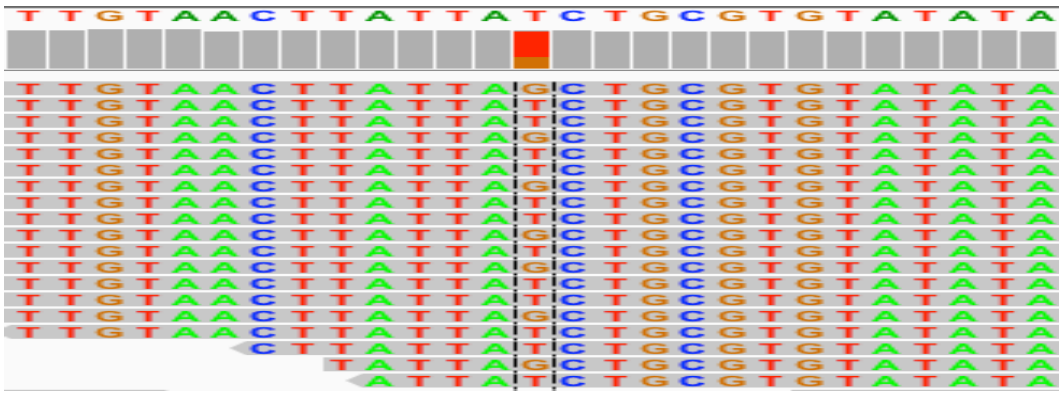


Figure 4.38: Heterozygous SNP (T/G) showed a deviation from the expected value (50%) in the observed reads (70:34).

The null hypothesis states that all subjects are heterozygous (T/G) and there will be no statistically significant difference between the observed and expected reads. When a deviation in reads was observed, a Chi-squared analysis was used to identify the significance of the deviation. The Chi-squared equation was: $\{X^2=\sum(\text{observed - expected})^2 / \text{expected}\}$. In our experiment, only two outcomes are possible (heterozygous) therefore the degree of freedom (total outcome-1) is one. Our critical P value = 0.05 which has a value of 3.841 in Chi-squared table. This means that when we calculate the Chi-squared, and given the final values is >3.84, the null hypothesis will be rejected. Three basic scenarios are expected. The first scenario was that the observed reads in an individual were (T=95;G=2), deviating from the expected reads count (T=50;G=50). The Chi-squared was used to calculate how significant the deviation was ($X^2 = 46.89$) and the associated P value = 7.5×10^{-12}. A highly significant deviation was observed, which is unlikely to be a result of chance and the assumption is therefore that this was caused by sequencing errors and that the true genotype is homozygous (TT). In the second scenario, the observed reads were (T=50; G=52) where G observed reads deviated from the expected (G= (50+52)/2=51 reads. The Chi squared calculation was = 0.08 and the P value = 0.7. No significant deviation was observed from the expected (T=51;G=51) and this individual is therefore deemed to be heterozygous (T/G). In the third scenario, the observed reads (T=90;G=20) deviated from the expected ones (T=55;G=55). The Chi squared analysis showed $X^2 = 18.64$ and the associated P value = 1.5×10^{-5}. A significant deviation was observed, but due to the number of reads mapped to the G allele, this deviation is unlikely to have been due to sequencing errors. The first two scenarios can be explained easily, but the third case requires a more detailed discussion.

The detailed analysis for rs17091737 is shown in Table 4.19. The average read coverage was 97 reads. The highest number of reads observed was a total of 138 reads in sample 18C and the lowest reads observed were 70 reads in 118C. All the bases in reads were counted and used to generate a total read count which was expected to contain only reads in G or T alleles. However, a few reads were observed in unexpected alleles. In most cases, a single base anomaly in the reads was observed, suggesting a sequencing error, therefore these reads were not included in the analysis of the expected read count. The Chi-squared test was conducted and the associated P value was identified. Based on the results obtained, the genotype for the 26 samples tested were identified. In most cases, the deviation in observed reads could be attributed to sequencing errors. Three such cases are now discussed in detail.

In sample 134C, the P value associated with the Chi-squared value (0.03) was found to be marginally significant; as the test was conducted 20 times, this points to a likelihood that the deduced value could be a random chance event.

In sample 70B, the P value = 0.005, which is a significant deviation. This sample contained another pair which showed a P value = 0.25 which is not significant. The sum of the paired reads was G-read = 109 and T-read = 94, and these were used to identify the combined Chi-square and the calculated P value= 0.46 which is not significant and so the deviation which had been observed was probably be due to chance.

The third case was observed in sample 86C, which had a highly significant P value (0.0004). No pair sample exists for this individual to calculate the compound X^2. This observation can be explained in many ways. The presence of sequencing capture artifacts that efficiently capture the T<G allele may have generated such an observation. However, in investigating other heterozygous samples, no evidence was found indicating that the T allele was captured more efficiently. It was also found that the sum of T-reads= 632 and G-reads= 610 had no significant difference and that the p value associated with X^2 (0.66) was not significant. This may indicate the absence of sequence capture bias. Sequencing errors can also generate the same observations. When the G is a natural allele, the sequencing procedure may still erroneously incorporate T. However, by looking to samples with homozygous G as in 22B, 29B, and 29C, no T-reads were found, therefore the error was not found to be due to the sequencing procedure.

To exclude the presence of germline mutation that could exist in individual 86C, IGV was used to identify other polymorphisms located in those regions, with the result that no other

polymorphisms were identified in the flanking sequence of the G allele on either side that could reduce the capacity of capture to such an extent that would generate such a read difference. Such a variant, if present, would be observed in all G alleles.

De novo somatic mutation occurs early in the lifetime of some cells where T is the natural allele and generates the ratio 1:2 observed. TaqMan SNP genotyping assay can be used to confirm this finding. The TaqMan SNP genotyping assay requires forward and reverse PCR primers, and two differently labeled TaqMan binder probes to bind (G/T) alleles. Each allele-specific probe is labelled with a fluorescent reporter dye and is subsequently attached with a fluorescence quencher. In the intact probe, the reporter dye must be quenched. During PCR, Taq DNA polymerase cleaves the reporter dye from the probe by its 5′-nuclease activity, when the probe is completely hybridised to the DNA strand. On separation from the quencher, the reporter dye emits fluorescence. An increase observed in either of the fluorescence dyes points to homozygosity in dye-specific alleles (G:G or T:T), and an increase in the fluorescence of both dyes demonstrate the heterozygosity (G:T) in the given sample. The dye colour can be identified and quantified, and genotypes can be generated (Shen *et al*., 2010). A known heterozygous, such as that in 118C, can be used as a positive control. Sequencing can also be conducted and the obtained result can be used to support or reject the finding.

The presence of small indels in the G allele can result in its misalignment and this can be investigated by amplifying a bigger region and checking for the presence of such variants (as described in Chapter Three).

The ratio observed was 1:2, which may indicate the presence of duplication, but in investigating other SNPs located nearby, no evidence was found in the resulting data to support this hypothesis.

Table 4.19: The detailed analysis for the rs17091737 SNP

	Base numbers in read												
Sample ID	**A**	**C**	**G**	**T**	**Total read count**	**Expected read count**	**Chi-square G**	**Chi-square T**	**Σx^2**	**P value**	**% G/T**	**Genotype**	**Comment**
DMGV119C	0	1	0	62	63	31					0/100	T	
DMGV18B	0	1	1	136	138	68.5	66.51	66.51	133.03	8.91E-31	1/99	T	Sequencing errors
DMGV18C	0	0	1	95	96	48	46.02	46.02	92.04	8.49E-22	1/99	T	Sequencing errors
DMGV99C	0	0	40	36	76	38	0.11	0.11	0.21	0.65	53/47	G/T	Chance
DMGV129C	1	0	39	41	81	40	0.03	0.03	0.05	0.82	48/52	G/T	Chance
DMGV133C	0	0	1	77	78	39	37.03	37.03	74.05	7.61E-18	1/99	G/T	Sequencing errors
DMGV134C	0	0	62	40	102	51	2.37	2.37	4.75	0.03	61/39	G/T	Chance
DMGV22B	0	1	118	0	119	59					100/0	G	
DMGV118C	0	0	35	35	70	35					50/50	G/T	
DMGV13C	0	0	60	60	120	60					50/50	G/T	
DMGV13B	0	0	47	61	108	54	0.91	0.91	1.81	0.18	44/56	G/T	Chance
DMGV106C	0	0	0	109	109	54.5					0/100	T	
DMGV92C	0	1	2	94	97	48	44.08	44.08	88.17	6.02E-21	3/97	T	Sequencing errors
DMGV91C	0	0	51	58	109	54.5	0.22	0.22	0.45	0.50	47/53	G/T	Chance
DMGV48C	0	0	0	95	95	47.5					0/100	T	
DMGV29B	0	0	112	0	112	56					100/0	G	
DMGV29C	1	1	99	0	101	49.5					100/0	G	
DMGV47C	0	1	40	59	100	49.5	1.82	1.82	3.65	0.056	41/59	G/T	
DMGV70B	0	1	60	33	94	46.5	3.92	3.92	7.84	0.005	65/35	G/T	Chance
DMGV70C	0	1	49	61	111	55	0.65	0.65	1.31	0.25	45/55	G/T	Chance
DMGV108C	0	0	93	1	94	47	45.02	45.02	90.04	2.33E-21	99/1	G	Sequencing errors
DMGV57B	0	0	50	36	86	43	1.14	1.14	2.28	0.13	58/42	G/T	Chance
DMGV57C	0	0	43	42	85	42.5	0.01	0.01	0.01	0.91	51/49	G/T	Chance
DMGV86C	0	0	34	70	104	52	6.23	6.23	12.46	0.0004	33/67	G/T	Chance
DMGV51C	0	0	0	89	89						0/100	T	
DMGV51B	0	0	0	81	81						0/100	T	
Total	2	8	1037	1471	97								

*Paired samples: gray shaded *Trios: blue shaded *Mother and son: light green shaded * Query finding: yellow shaded

4.5.7.1 Hardy-Weinberg equilibrium

The number of observed alleles in rs17091737 was counted to identify the allele frequency (observed alleles/total number of alleles) and the results are shown in Table 4.20.

Table 4.20: Allele frequency for the rs17091737 SNP

Alleles	Number of Alleles observed	Allele Frequency observed	1,000 Genome GBR frequency
T	23	0.605	0.65
G	15	0.395	0.35

The genotype frequencies for the individuals tested was calculated using the Hardy-Weinberg equilibrium equation ($p^2 + 2pq + q^2 = 1$). The results obtained for rs17091737 are shown in Table 4.21.

Table 4.21: Hardy-Weinberg equilibrium calculation for the rs17091737 SNP

Genotype	Number of individuals observed	Genotype frequency	Number of individual expected	Chi-square	
TT	7	0.366	8	0.13	
GT	9	0.478	9	0.02	
GG	3	0.156	2	0.21	
Total	19	1.00	Σx^2	0.36	P value= 0.84

The numbers of observed alleles were counted to identify the allele frequency (observed alleles/total number of alleles) and the results were used to calculate the genotype frequency. The allele frequency of the population of British in England and Scotland (GBR) in the 1000-Genomes was used to count the number of individuals expected to have the genotype (Consortium GP, 2010). In this case, there was no significant difference between the number of individuals observed and the number of individuals expected to have the same genotype. The MAF observed (G= 0.395) in our population is similar to the British MAF observed (G= 0.35) by 1,000 Genomes. The samples tested are in the Hardy-Weinberg equilibrium for rs17091737. The same analysis was conducted on other SNPs, and a Chi square was used to determine the significant differences between the observed and expected genotypes. No significant deviation was observed between the observed and expected genotypes in any of the SNPs tested and they all fitted the Hardy-Weinberg equilibrium. The SNP data can be found in Appendix IV.

4.6 Discussion

The main aim of the present study was to conduct a pilot experiment to examine the potential of new technologies to capture both progenitor and low-frequency mutant alleles, thereby achieving the same goals as SP-PCR in a more rapid manner with a higher throughput and which could be used in laboratories where high sample throughput and a rapid turnover are required, such as in a forensic lab. NGS technologies have allowed new approaches to capture and sequence selected DNA sequences. In this experiment, a variety of human DNA sequences including microsatellites, DNA repair genes, telomeric regions, and other sequences which were of interest to our research group were included. A DNA sequencing target enrichment approach was selected in order to capture only the unique DNA sequences that will be easily aligned, since a single copy is present in the reference genome. This experiment is a preliminary one and includes a novel capture strategy. Custom baits, which are complementary to the desired unique DNA sequences selected, were used. The total number of baits was 57,646 covering 7 MB and single bait spans of 120 bp. The cost of the experiment remains relatively expensive and only 20 individuals with an age range at sampling of between 18 and 75 years were included. A range of samples was used that includes trios, a mother and son, and 6 individuals sampled at two different age points, in order to ensure that the experiment's aims are achieved.

The total numbers of sample was 26. The maximum custom bait size available through Agilent was selected. The longer the bait size, the greater the capture which could be expected, as more complementary sequences exist and there will be more opportunity to bind to the targeted sequences. Sonication was recommended by the SureSelectXT Target Enrichment System, which is used by Illumina Paired-End to generate sequencing library (Agilent Technologies, 2010). If the DNA fragment produced does not match the bait sequence, no capture can be expected. The ideal DNA fragment is one that spans the bait sequence. The DNA size which is either too short or too long may reduce the quality of the final sequencing libraries. A DNA shearing step is therefore the most crucial step and its quality was assessed in the present experiment using a DNA Bioanalyser assay to ensure a distribution with a peak size of between 150 to 200 nucleotides. Similarly, the quality of the finished DNA libraries was also assessed and the peak size was found to be between 300 to 400 nucleotides. Finished DNA libraries with a controlled number of reads will be generated. Optimal conditions provide 700K to 900K clusters/mm^2 on the GA IIx. The cluster density is inversely proportional to the concentration of the finished DNA library.

In this experiment, the data generated using the Illumina platform where the Pair Ended (PE) mode was set to produce 151 bp reads. Millions of reads were successfully counted for each sample, with an average read count of 8.0 x10^6. The total number of reads showed a significant correlation with age at sampling. This can be explained by the presence of telomere reads among the total reads that are known to be age dependent. This has also been verified when telomere reads were excluded and the remaining unique DNA sequences showed no significant correlation with age at sampling. This result also gave an indication of DNA sequence capture quality and, as was expected, unique DNA sequence should not be significantly affected by the ageing process. However, with unique DNA sequences, a declining trend in relation with age at sampling was observed. Moreover, the result showed that the telomere reads counted reduced as age increases, indicating the success of our novel approach to capture them (as discussed in Chapter Five).

To identify the quality of the capture for each unique sequence, the Bowtie aligner was chosen to map the reads generated in this experiment to the human genome reference (HG19). Bowtie is commonly used in bioinformatics having showed its effectiveness in previous studies (Langmead and Salzberg, 2012). Since these sequences are unique (*i.e.* they involve a single copy), a single read was forced to map to one location on the human genome reference. The aligned reads that span any bait sequence were then counted and separated to identify each group count. The unique sequence of the bait ratio (~86%) and the unique DNA sequence captured cover 60% of the total reads were captured. The SureSelect capture efficiency was estimated at 60% for exons and 80% for genomic regions (Metzker, 2009).

No significant correlations were found between the total unique DNA sequences captured and age at sampling. No significant correlation was identified either with sheared DNA size, and a declining trend was observed with finished DNA libraries, so separate analysis was conducted on microsatellites and DNA repair genes.

For microsatellite sequences, a novel approach was developed using hybridising baits that targeted only the unique flanking regions of repetitive sequences. Our strategy assumes that a bait bridge will be formed to incorporate the repetitive sequences positioned between the unique flanking that makes the loci arms. As this is experiment was a pilot, only pure mono, di, and tri nucleotides were selected. Our interest was in genotyping polymorphic microsatellites, and the presences of mismatches within the repeat would reduce the degree of polymorphism, so pure microsatellites with longer alleles were chosen as they are more likely to be polymorphic (Ellegren, 2004). 22% of the total number of microsatellites

existing in TRDB were selected when a 60 bp window was screened using RepeatMasker. Nearly 80% were lost using these criteria however, and the selected microsatellites (25,539) were reasonably distributed across all chromosomes.

Unique DNA sequences were mapped to the human genome and reads that matched the microsatellite bait sequences were counted, producing a highly successful approach. The microsatellite bait ratio was that 44% of the total baits used and 51% of unique baits. The microsatellite reads which were successfully mapped represented 40% of the mapped unique reads and 24% of the total captured reads. Our bait bridge strategy was able to capture 97% of the microsatellites selected. The bridge formation was expected to be limited and large repetitive sequences >60 bp might possibly be captured while failing to generate reads that span the full length.

In order to study repeat length variation from the data generated from this experiment, the aligned read must span the microsatellites' full lengths. The alignment process is not simple and full spanning polymorphic loci reads may not be able to align to the reference genome. Moreover the PCR used to generate an NGS read is known to introduce errors when amplified sequences contain microsatellites, especially with respect to mononucleotides and dinucleotides (Clarke *et al*, 2001). SNP surrounding and inside microsatellites can affect both the alignment and genotyping processes. Two members of our research group, Carlos del Ojo Elias and John Cole, observed a high rate of SNPs in the edges of the microsatellites. SNPs that are located at the start of the flanking sequence of the microsatellites lead us to believe that the definition of the start and end of the microsatellites in the bait chromosomes indicates that the consensus sequence (as employed to design the bait) may not be precise in capturing microsatellites. The presence of mutations inside the microsatellites decreases their allele length, whereas if it is present outside them, it may lengthen the microsatellites. Tandem Repeat Finder was used to find microsatellites with pure sequences to generate the bait using only one version of the human genome. However, some of the "pure" microsatellites turned out not to be pure. The SNP present in the reference genome may provide false signals in the TRF analysis microsatellite sequences, thereby producing an inaccurate list of supposedly "pure" microsatellites. Actually, the microsatellites may be longer but interrupted by an SNP. To correct wrong detection, the detection mechanism could be improved by analysing the nearest SNPs and predicting their real size. Polymorphic repeats located in the microsatellites flanking sequences could affect the sequencing process and the aligning process. Both BLAST and RepeatMasker were used to minimise these factors (Brohede, 1999).

Differences were observed in the microsatellites coverage. Some microsatellites have partial left or partial right spanning reads and a high coverage of full spanning reads. In other microsatellites, an inverse pattern was observed. Differences in the microsatellite coverage might be due to variations which are present in the nearby DNA sequences which will prevent accurate mapping. Microsatellites with large allele lengths fail to generate full spanning reads. The total reads quality dropped in 18% of the microsatellite reads especially after or just at the end of the microsatellites. Bridge amplification may cause a frame shift effect in the early stages of the sequencing process (Pompanon *et al*., 2005).

Interestingly, the microsatellite read count showed a significant correlation with age at sampling. The number of microsatellite reads decreased as the age at sampling increased. One explanation for this is that this is a true effect and more mutation is generated in elderly people, which prevent us from capturing these sites leading to a decrease of total read with increasing age. Another explanation is that older individuals have microsatellites with larger allele lengths that fail to generate informative reads. The other possible reason is that the DNA shearing process has compromised the capture process. But, the sheared DNA size has no significant correlation with the total microsatellites reads count. The finished DNA library size is marginally significantly correlated to total microsatellite reads. When the finished DNA library size increases, the total microsatellite reads decrease.

Further, where unique DNA sequence data showed significant or no significant correlation with age at sampling or where declining trends were observed, the same correlation was observed with finished DNA library size. The number of samples needs to be increased in order to verify this observation.

Most of the non-covered microsatellites had a high AT content, which was observed in the microsatellites pattern, or flanking sequences, or both. The presence of non-covered microsatellites can be explained by a misalignment process of these loci, whereby such sequences are commonly spread across the human genome or have been denatured because of the heat generated during the DNA shearing process using sonication. The analysis of the genome data by Illumina has showed an underrepresentation of AT-rich (Dohm *et al*., 2008; Hillier *et al*., 2008, Harismendy *et al*., 2009) and GC-rich regions (Hillier *et al*., 2008; Harismendy *et al*., 2009), which is probably due to an amplification bias during library preparation (Hillier *et al*., 2008).

To investigate how informative the microsatellite reads were, 27 microsatellites were selected and genotyped in 20 individuals using CLC Genomic workbench. The

mononucleotide repeats proved difficult to accurately genotype. In most cases, multiple reads for the same loci were shorter by one or a few repeats than the full-length product. In contrast, dinucleotides were genotyped successfully, but a few samples showed the same pattern which had been observed in mononucleotides; the reads ratio was used to choose the most probable genotype. Reads were observed which differed by one or a few repeats. This can be attributed to PCR slippage during amplification as a part of the sequencing process, making allele length identification difficult (Levinson and Gutman, 1987; Meldgaard and Morling, 1997), especially for heterozygotes with adjacent alleles. Tri- and tetra-nucleotides repeats appear to be significantly less prone to slippage (Edwards *et al*., 1991) and were successfully genotyped. Trios, a mother and son, and paired samples genotypes were matched.

The reads mapped to the SNP position were informative and generated genotypes. Twelve SNPs were successfully genotyped. The average read coverage for the selected SNPs was 97 reads, ensuring a good genotyping. The Hardy-Weinberg equilibrium was calculated and the individuals tested were found to be within its parameters. A Chi-square test was used to identify any significant differences between observed and expected data.

Data were also successfully generated from sequencing 233 DNA repair genes covering 3,118 exons, using 8,252 baits. The reads were aligned to the human genome reference (HG 19) and reads match baits were counted. The reads counted represented 13% of the total reads and the DNA repairs reads showed no significant correlation with age at sampling. Thus, the data can be further used to estimate age-dependent telomere shortening. In the analysed data, we observed a non-significant declining trend with age at sampling, and this observation could be confirmed or corrected by increasing the number of samples.

A linear regression analysis was conducted multiple times in order to analyse the data generated from this experiment. In all of the multiple testing conducted, the assumption was that the distribution of P values (0.05) was significant. This also meant that if the test was conducted 20 times, one test could by chance produce a significant value (1/20 = 0.05). The Bonferroni test is a simple method for correcting the P value when performing independent hypothesis tests. A Bonferroni correction can be obtained by dividing the critical P value by the number of linear regression analysis conducted. The statistical significance of the study is then re-calculated based on this obtained P value. When multiple tests are performed on a single set of data, the Bonferroni correction is used to reduce the chances of obtaining false-positive results (Bland and Altman, 1995). Using the

Bonferroni, the correct critical P value was (0.05/21= 0.0024). Although the Bonferroni correction used to decrease the chances of obtaining false positives, as the number of tests increases it can become very conservative and this increases the risk of generating false negatives.

This experiment can be improved and expanded. Using sonication to shear DNA is a random process and cut DNA sequences at different places, generating different DNA fragments that may or may not include the full length of the query sequences, thereby affecting the outcome. Using other means, such as a biochemical approach using a cocktail of restriction enzymes that cut DNA at specific length or specific sites will ensure that the fragmented DNA includes the full length of the query sequence or the same fragment size and sequence content for each sample tested with a length that can be sequenced with an available sequencing platform. Careful bioinformatics analysis needs to be conducted to identify the enzyme cocktail, the suitable fragment length and the microsatellites that should be captured.

The read length generated by the sequencing platform also contributes to the success of such an approach. There is no doubt that sequencing technology is moving forward to solve such problems. The sequencing technologies which are currently available, such as Illumina GA, Roche 454, ABI SOLiD and Ion Torrent, generate short read lengths and possess methodological limitations that can result in mis-mappings and mis-alignments (Carneiro *et al*., 2012). Significantly longer reads averaging >1 kb can be achieved using Pacific Biosciences technology, which enables more unique seeds for reference alignment. Additionally, the common source of base composition bias is minimised because this approach lacks PCR amplification in the library construction (Carneiro *et al*., 2012).

Alignment and genotyping will be another issue with such long reads. Side by side, the sequencing bioinformatics is also improving. Software such as SatNav, which has been developed by our research group, and other software available in the market will be able to solve these issues. SatNav is a pipeline that incorporates bowtie. It takes a list of STR locations, builds probes for each allele of each STR, uses bowtie to align reads to the probes, filters the reads, genotypes each locus.

LobSTR is another software used to profile STRs in personal genomes. LobSTR joins concepts from signal processing and statistical learning to avoid gapped alignment and to adopt the specific noise patterns in STR calling. LobSTR involves of three steps. The sensing step where informative STR reads are detected and their repeat motif are determined. The alignment step, the STRs' flanking regions are mapped to the reference

finally, the genotyping step where the STR alleles length present at each locus are determined (Gymrek *et.al*, 2012).

In our approach to capturing microsatellites, the bait used was made up of unique flanking sequences to overcome the polymorphic nature of microsatellites and to ensure correct alignment. However, this is true for the capture to avoid capturing unwanted sequences, and to overcome the polymorphic nature of microsatellites. But with regard to alignment, a minimum of 10 bp in flanking is required to ensure accurate mapping. If this is used to capture targeted DNA sequences with shorter flanking DNA sequences, more microsatellites can be added at the same cost.

Identifying full spanning reads that include all microsatellites' lengths and some flanking sequences on both sides that will enable accurate alignment and genotyping remains a challenge. These reads are controlled by three factors. These are now briefly outlined.

First, the sheared DNA must be fragmented to a size that can be sequenced by the selected sequencing platform and its sequence content must be cover the full length including some flanking sequence on both sides. Minimum reads coverage per locus and minimum reads coverage per allele are required too. To overcome the polymorphic nature of the microsatellites, a tested custom DNA reference can be generated and used to replace human genome reference, since the human genome reference contains a single copy for each of the microsatellites and can cover all the different alleles which are present in a population or are generated by errors introduced by PCR.

Finally, in addition to the sequences in question, the target enrichment approach will capture other unwanted sequences that may affect the total reads generated. Moreover, the non-significant declining trend associated with age at sampling observed in some of the data analysed can be also be verified by testing more samples. To minimise the capture of unwanted sequences and verification of non-significant findings, each DNA sequence in question needs to be conducted in a separate experiment with the maximum number of samples possible recently collected. However, due to the high cost of this approach and this experiment's preliminary nature, we conducted all questions in one run.

5 Age estimation based on human telomere shortening

Human telomeres are at the ends of linear chromosomes and consist of multiple arrays of TTAGGG•CCCTAA repeats with an average length of ~10 kb (Neumann *et al*., 2002). They protect the ends of the linear chromosomes and contain special binding proteins that prevent them from being recognised as double-strand DNA breaks (Greider, 1999). Telomeres form large loop structures which are known as telomere loops, or T-loops. Here, the end of the telomere folds back in a long circle and is stabilised by telomere-binding proteins (Griffith *et al*., 1999). As well as potentially activating double-strand break repair, another concern connected with linear chromosomes is the end replication problem where the DNA polymerase is unable to completely replicate the ends of linear DNA (see Figure 5.1) (Harley, 1991). Because of this end replication problem, the telomere regions are potentially shortened with each cell division. However, they can be replaced by telomerase using an RNA template for subsequent amplification (Harley, 1991). Telomerase is not normally expressed in somatic cells, so every time the cell divides, the telomere becomes shorter (Meyerson, 2000). The "cellular or replicative senescence" machinery is used by normal cells which undergo a finite number of cell divisions to stop dividing (Kufe *et al*., 2003). In 1961, Hayflick and Moorhead observed a limited number of cell divisions in cultured human fibroblasts (Harley, 1991). In culture, cells undergo cellular senescence after approximately 50 cell divisions. With each round of DNA replication, the telomeres are shortened (Boukamp, 2001) and when they reach a critical length, cellular senescence is initiated (Bekaert *et al*., 2004). Even though there is no direct correlation between cellular senescence and ageing (Bekaert *et al*., 2004), ageing is thought to be one of the contributing factors.

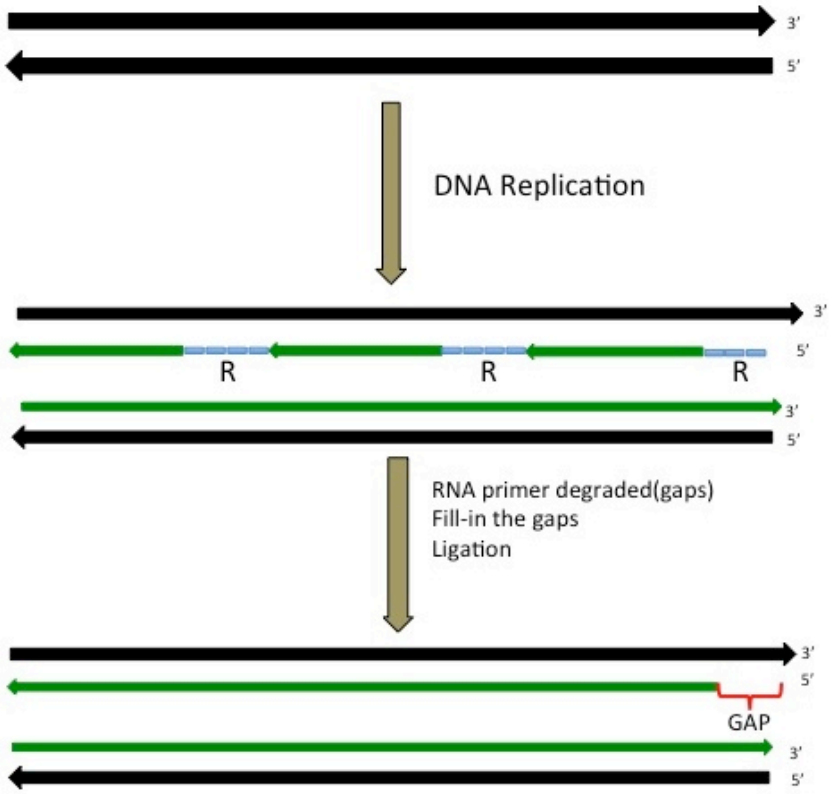


Figure 5.1:The end replication problem: Linear parental strand is shown in black; the newly replicated strands are in green. An RNA primer (blue rectangle (R)) is required to replicate the lagging strand. Removal of the RNA primer generates a gap at the end of one of the two newly synthesised strands (source: modified figure from Meyerson, 2000).

The ageing process has been explained by many theories, but the telomere hypothesis on ageing has been among the favoured ones; it was suggested by Olovkinov as the "marginotomy theory" in 1971(Harley, 1991). The "telomere hypothesis" proposes that the nonstop shortening of telomeres throughout cell division results in ageing (Boukamp, 2001). Cultured human fibroblast telomere analysis was the first direct evidence for telomere loss during cellular growth (Harley *et al*., 1990). Human cells are able to divide 50-70 times before senescence. Age-dependent telomere shortening was observed in studies conducted in both *in vitro* (Harley *et al*., 1990) and *in vivo* (Hastie *et al*., 1990; Lindsey *et al*., 1991). *In vivo*, age-dependent telomere shortening was first demonstrated in blood and colon mucosa (Hastie *et al*., 1990). Subsequently, similar observations were found in tissues such as fibroblasts, liver, thyroid, parathyroid, brain, kidney, vascular tissue, myocardium, spleen and pancreas (Allsopp, *et al*., 1995; Allsopp, *et al*., 1992; Benetos *et al*., 2001; Butler *et al*., 1998; Chang *et al*., 1995; Counter *et al*., 1992; Harley *et al*., 1990; Kammori *et al*., 2002; Lindsey *et al*., 1991; Melk *et al*., 2000; Rufer *et al*., 1999; Takubo *et al*., 2000; Vaziri *et al*., 1993). The risk of death has been observed to increase with shorter telomere length. In a study of the association between telomere length in

blood and mortality, 143 individuals over 60 years of age were investigated and the results obtained indicated that individuals with shorter telomere length were three times more likely to die of heart disease and eight times more likely to die of infectious disease than those with longer telomere lengths (Cawthon *et al*., 2003). It has been also demonstrated that individuals with premature ageing diseases such as Hutchinson-Gilford progeria or Down syndrome have shorter telomere lengths (Harley *et al*., 1990; Vaziri *et al*., 1993). Even Dolly the sheep, the first cloned mammal from adult somatic cells, was found to have shorter telomere lengths, which was believed to be the reason for her premature ageing (Shiels *et al*., 1999). In higher eukaryotes, germ line cells and tumour cells preserve their telomere lengths through expression of the enzyme telomerase (Rattner, 1995). In addition, in some telomerase-negative SV40 tumour cell lines, telomere length is maintained at a constant by a mechanism known as alternative lengthening of telomeres (ALT) (Neumann *et al*., 2002). It has been suggested that a recombination mediated process involving the telomere strand of one molecule serves as a template for the synthesis of DNA on another strand (Nemann *et al*., 2002). In humans, four out of 57 primary tumours tested were also telomerase negative, yet they still showed a significant increase in the average telomere length (Neumann *et al*., 2002).

Similarity in the length of the telomere in monozygotic twins suggests that the length of the telomere is inherited. Moreover, when a comparison in the rate of telomere shortening was made between the homologous chromosomes of both monozygotic twins at a set point in time, no significant difference was observed (Graakjaer *et al*., 2004). An investigation of the telomere length shortening in a homogenous Amish population revealed that, consistent with the similarity in the average life span of the males and females in the population, their telomere lengths also match. In addition to this, a strong association has been observed between paternal telomere length and the telomere length of fathers' offspring (Njajou *et al*., 2007).

The role of parental age at conception in telomere length has been investigated. Three studies, conducted by Unryn *et al*. (2005), Njajou *et al*. (2007), and Meyer *et al*. (2007) have each revealed the importance of the father's age in determining an individual's telomere length. On average, a child has an increase in telomere length of 22 bp in Unryn *et al*. (2005), 10 bp in Njajou *et al*. (2007) and 17 bp in Meyer *et al*. (2007) for each year that the father is older at the time of conception/birth of the child. In sperm, telomere length has been observed to increase with the age of the individuals, due to the activity of the telomerase enzyme in adult testes (Allsopp *et al*., 1992). However, no significant

correlation was observed between the mother's age at conception and her offspring's telomere length (Njajou *et al*., 2007); indeed, in connection with this, no telomerase activity has been demonstrated in the female germline (Unryn *et al*., 2005). Nonetheless, females still transmit telomeres with varying lengths, contributing to individual variation in telomere length.

5.1.1 Methods used to measure telomere length

5.1.1.1 Telomere restriction fragment (TRF) analysis of telomere lengths

TRF is the standard approach used to measure telomere length in humans (Chang *et al*., 1995; Harley *et al*., 1990; Takasaki *et al*., 2003; Tsuji *et al*., 2002). Genomic DNA is digested using restriction enzymes that do not cut the telomere segments and it therefore generates terminal restriction fragments, which are resolved by gel electrophoresis. Subsequently, the resolved fragments are blotted to a membrane. At the end of the process, hybridisation using a labelled telomere probe is used to visualise the telomere fragments (Harley *et al*., 1990).

One major disadvantage of the TRF method is that it requires a large amount of DNA (0.5-5 μg/individual), and can also be time consuming (typically taking 3-5 days to perform) (Cawthon, 2002). The overall sensitivity of this method is considered to be low, and is has also been found to be biased toward detecting longer telomeres (O'Callaghan *et al*., 2008). In TRF, labelled telomere repeat probes hybridise to the fragmented telomeric sequences, and shorter telomeres produce lower signals because they hybridise to fewer copies of the probe. Thus, a telomere length detection threshold exists. Short telomeres below the TRF detection threshold will not be detected and this limits ageing research involving somatic tissues, as the shortest telomeres are also considered to be the most critical in initiating cell replicative senescence (Baird, 2003). Sometimes, subtelomeric regions are also included in the fragments (Baird *et al*., 2006; Nordfjäll *et al*., 2005), raising the question as to whether the results obtained from the TRF length analysis represent the real telomere lengths or not. Furthermore, the auto-radiographic smears produced also mean that telomere length analysis is at least partially subjective (Nakagawa *et al*., 2004). This method is also restriction enzyme dependent, as the different enzymes used result in restriction fragments that may differ by up to 5% (Baird *et al*., 2006; Cawthon 2002). Additional differences can also be caused by the presence of polymorphisms in restriction sites near the telomeric and subtelomeric regions in different chromosomes (Cawthon, 2002; Nakagawa *et al*., 2004).

5.1.1.2 Real time PCR analysis of telomere length

In most cases, low amounts of DNA are extracted from the samples collected from crime scenes. Using a PCR-based approach would overcome this problem. Cawthon has previously used quantitative real time PCR to measure telomere lengths (Cawthon, 2002). Low quantities of DNA (35 ng) are required and the analysis is rapid (it can be finished in a few hours). Specific primers were used by Cawthon to specifically bind to telomere regions and ensure that no primer dimer is formed (Cawthon, 2002). Primers are specially designed to have mismatches even with the target sequences, as when the two primers form a dimer, more mismatches are generated and the 3′-terminus base of each primer cannot form a stable base pair with the base opposite it, thus blocking the addition of new bases by DNA polymerase (Cawthon, 2002), as shown in Figure 5.2. No subtelomeric regions are amplified and only telomeric (TTAGGG) repeats are amplified (Nordfjäll *et al*., 2005). In quantitative real time PCR, a single-copy gene that is assumed to be constant is used in comparison to the amount of DNA in a sample (Cawthon, 2002).

Using the quantitative real time PCR analysis of telomere length, both telomere repeats and the single-copy gene concentration are calculated for the same sample. The concentration of these two parameters is determined using an external standard curve. Then, the total concentration of the telomere repeat (T) is divided by the total concentration of the single-copy gene (S), resulting in the (T/S ratio) value that is further divided by the T/S ratio obtained for the control DNA. The resulting ratio represents the number of telomere repeats, known as the relative telomere length (Cawthon, 2002).

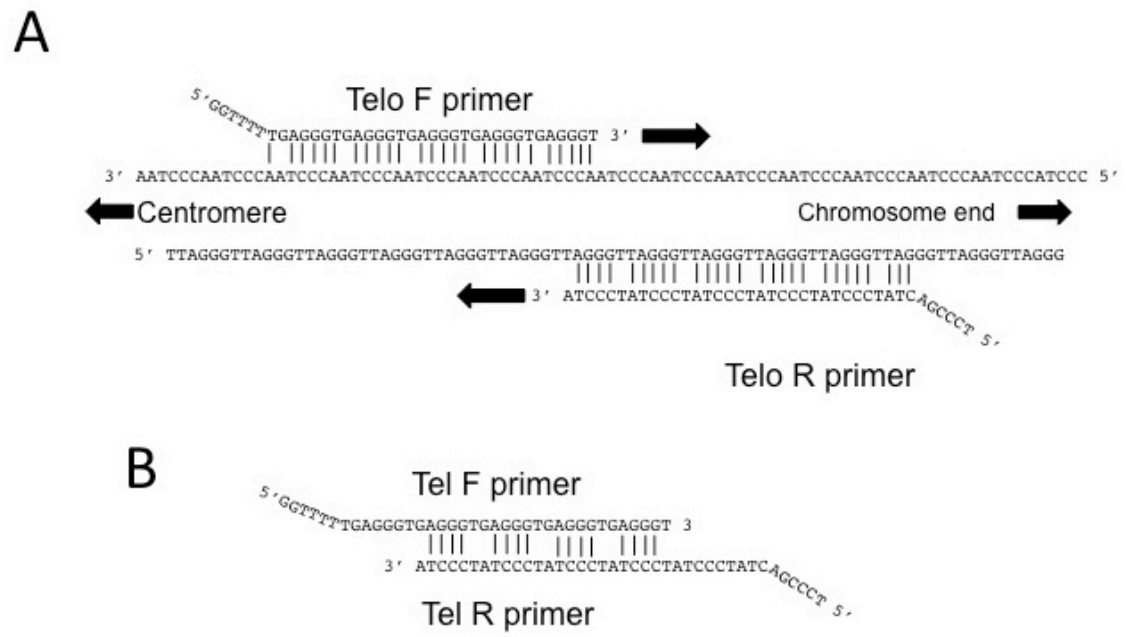


Figure 5.2: Telomere measurement using quantitative real time PCR: (A) primers annealing to the genomic DNA. Telo F primer (37 bp) with 31 bp of telomere complementary sequence anneals to the telomeric region and oriented 5' ➔ 3' toward the centromere. Telo R primer (39 bp) with 33 bp of telomere complementary sequence anneals to the telomeric region and oriented 5' ➔ 3' toward the end of the chromosome. In both primers a mismatch is present at every sixth base. (B) Primer dimer. To form the dimer, repeated patterns are created of six bases comprising four uninterrupted paired bases followed by two mismatch bases. The last base at 3' end of each primer cannot form a stable base pair with the base opposite it, preventing the addition of bases by the DNA polymerase (Cawthon, 2002).

5.1.2 Factors that may affect telomere lengths

Both lifestyle and environmental factors have been observed to have an impact on telomere lengths, through free radicals and antioxidants generated in the body (Lin *et al*., 2012). Other studies to estimate telomere shortening in smokers have also been conducted (Morla *et al*., 2006). In comparison with non-smokers, smokers' telomere length decreased significantly faster with age. A dose-response relationship has been observed between the cumulative long-life exposure to tobacco smoking (pack-yrs) and telomere length. Moreover, obesity and cigarette smoking have also been shown to significantly decrease telomere length blood lymphocytes in women (Valdes, 2005).

5.2 Results

5.2.1 Measurement of fragment telomere length - the theory

DNA target enrichment and next-generation sequencing will be used together in the present study to estimate age-dependent telomere shortening. The target enrichment approach will be used to capture DNA fragments containing a telomeric repetitive sequence (T) and unique DNA sequences (U) using custom baits. The captured DNA sequences will be sequenced using the Illumina platform. The short reads generated will then be counted, in order to identify the number of telomere reads (T), and the number of unique sequence counts. The data obtained will be used to calculate the relative telomere length for each sample by dividing the number of telomere repeats (T) by the number of reads covering unique DNA sequences (U). The resulting value (represented by the T/U ratio) will be used to estimate age-dependent telomere shortening. The obtained value represents the quantity of telomere repeats, known as the relative telomere length, that is analogous to the Cawthon analysis (Cawthon, 2002).

5.2.2 Experimental approach (telomere region capture)

The aim of this novel experiment is to capture and estimate the telomere length using a target enrichment NGS approach. Our strategy was to use unique probes (120 bp) to capture unique DNA sequences. Therefore, a single bait was used to capture a single unique sequence, and a 120 bp probe consisting of pure telomeric sequence $(TTAGGG)_{20}$ was used to capture the telomere regions. This novel approach is required, as the telomere regions consist of a multi-array of TTAGGG sequences with the average size in an individual estimated to be ~10 kb for each chromosome end. Each chromosome end was equivalent to ~83 baits (10,000/120). The 92 telomeric ends of 46 chromosomes were therefore covered using ~7,667 baits ((10,000 x 2 x 46)/120) generated to capture telomere regions for both arms (p,q) and representing 13.3% of the total baits ordered. The total telomere bait capture size was ~920 kb for each individual tested.

5.2.3 Library preparation

DNA samples were obtained from 20 individuals with an age range of 18-75 years. The library preparation method was discussed in Chapter Four.

5.3 Telomere length estimation in the general population

The Illumina sequencer generated million of reads for each individual. Using CLC Genomic Workbench, some NGS reads were successfully aligned to the telomeric regions on the human reference genome, as shown in Figure 5.3.

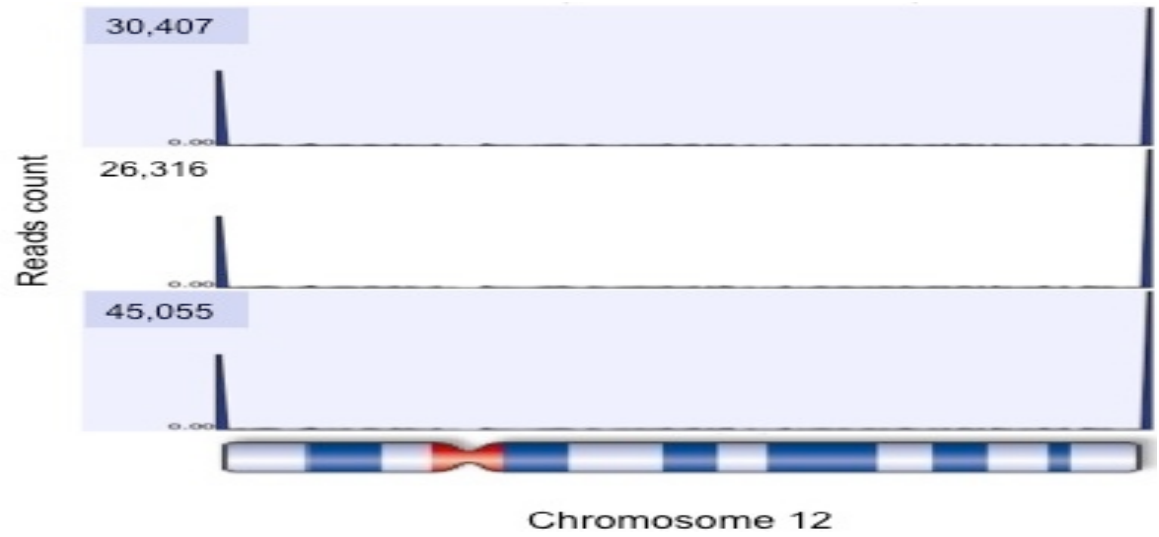


Figure 5.3: Mapped NGS telomeric reads: a CLC Genomic Workbench was used to align reads to human reference genome and huge number of reads were successfully mapped to both q and p arm of each chromosomes. (3 samples and reads mapped to chromosome 12 are shown here).

The CLC Genomic workbench was used to count the reads containing telomeric sequences. A custom reference $(TTAGGG)_{25}$ was generated and the reads were mapped at a 50% similarity level. However, the telomeric reads generated include short reads (<150 bp), variants present in telomere repeats such as TTGGGG, TTAGGG, and TGAGGG (Allshire *et al*, 1989), and sequencing errors. These mismatches might cause a decrease in the telomeric read count. Therefore, with aid of Dr. Graham Hamilton, another count was conducted: it counted the presence of three pure continuous telomeric repeats, ≥ $(TTAGGG)_3$, or more within a single read (*i.e* 151 bp) and was called First count (F-telomere count).

Using the novel NGS approach, telomeric regions were captured successfully in all of the samples tested and then counted. Both counts were conducted on the generated data, resulting in the read counts shown in Figure 5.4. The F-telomere count registered between 500,000 and 1,100,000 reads. The CLC-telomere count ranged between 400,000 and 1,000,000 reads. As expected, the use of CLC count resulted in lower read counts. A detailed analysis for each count was carried out.

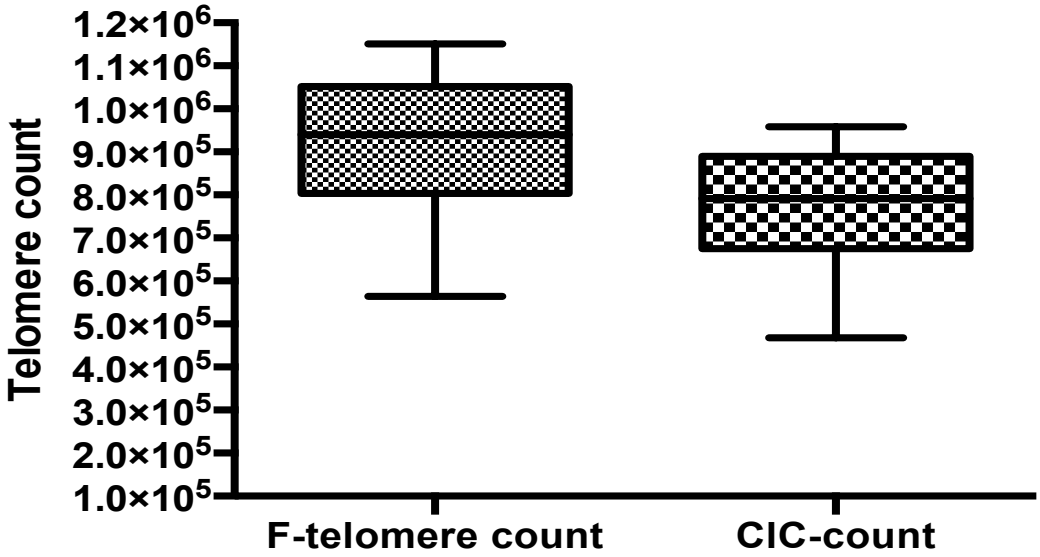


Figure 5.4: Average telomere reads count. The Illumina sequencer has generated millions of reads for each individual. The data show both the F and CLC telomeric counts**.**

5.4 Telomere count

The detailed count per sample for both the F-telomere and CLC-telomere counts are shown in Table 5.1. The average F-telomere count totalled 923,462 reads, representing ~12% of the total read count, which is similar to the bait ratio used to capture the telomere region (13.3%). The average CLC-telomere count totalled 776,599, or ~10% of the total read count. Read coverage was calculated by dividing the average read count by the number of baits used to capture telomere regions (*i.e.* 923,462/7,667). The read coverage was ~120 for the F-count and ~101 for the CLC- count. The percentage of average difference between the F and CLC telomere reads was ~16% observed and this was approximately the same across most of the samples.

Table 5.1: Total telomere reads count.

Sample ID	*Age at sampling*	*F-telomere read count*	*CLC-telomere read count*	*Telomere read difference*	*Read difference/F-telo (%)*
DMGV119C	75	690,948	580,793	110,155	16
DMGV18C	72	805,418	680,804	124,614	15
DMGV99C	69	785,496	663,381	122,115	16
DMGV129C	67	688,994	564,381	124,613	18
DMGV133C	66	564,474	468,475	95,999	17
DMGV22B	65	822,142	678,665	143,477	17
DMGV106C	60	820,108	690,720	129,388	16
DMGV118C	58	715,020	601,446	113,574	16
DMGV18B	56	1,059,334	882,002	177,332	17
DMGV13C	53	912,632	767,220	145,412	16
DMGV13B	52	1,067,082	913,823	153,259	14
DMGV92C	51	955,368	815,312	140,056	15
DMGV29C	49	798,736	662,102	136,634	17
DMGV48C	45	1,034,746	869,824	164,922	16
DMGV47C	43	1,025,382	876,868	148,514	14
DMGV108C	40	1,141,128	958,051	183,077	16
DMGV134C	40	1,011,128	837,064	174,064	17
DMGV70C	38	925,140	760,851	164,289	18
DMGV29B	36	875,106	722,003	153,103	17
DMGV57C	29	882,282	741,240	141,042	16
DMGV70B	26	1,117,424	930,837	186,587	17
DMGV91C	26	1,068,920	902,513	166,407	16
DMGV86C	25	1,024,374	883,251	141,123	14
DMGV51C	23	1,050,036	885,102	164,934	16
DMGV57B	20	1,018,023	857,267	160,756	16
DMGV51B	18	1,150,570	927,786	222,784	19
	Average	**923,462**	**776,599**	**146,863**	**16**
	Average read coverage	**120**	**101**		
	Percent of total read (%)	**12%**	**10.2%**		

5.4.1 Age-dependent telomere shortening

The total read count showed a significant correlation with age at sampling (as discussed in Chapter Four). This finding can be explained by the presence of telomere reads among the total read counts and the trend of decreasing telomere reads with increasing age. To clarify this, the effect of age at sampling on the total uncorrected telomere read counts using F-telomere count was analysed, and the results are presented in Figure 5.5. The linear regression analysis showed a highly significant correlation between age at sampling and the uncorrected F-telomere count (P value < 0.0001 and $r^2 = 0.55$). A similar finding was obtained with the uncorrected CLC-telomere count, which also showed a highly significant correlation with age at sampling (Figure 5.6) (P value < 0.0001 and $r^2 = 0.53$).

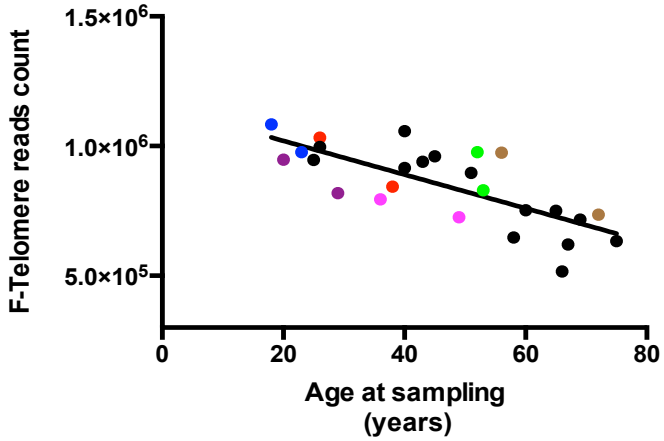


Figure 5.5: Age-dependent telomere shortening using F-telomere count.

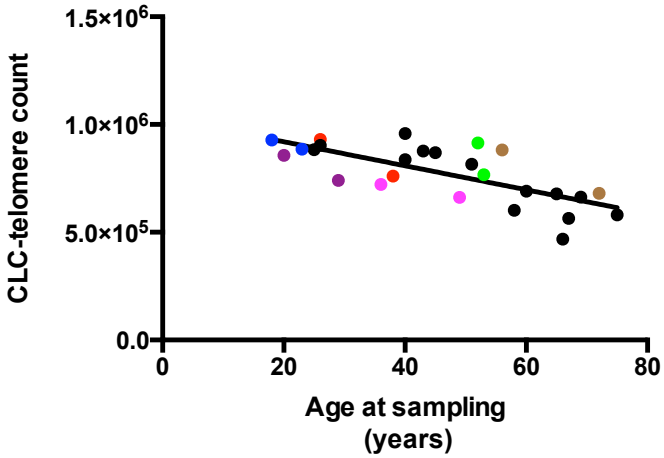


Figure 5.6: Age-dependent telomere shortening using CLC-telomere count

5.4.2 Comparison between F and CLC telomere count

Both F and CLC telomere reads showed a significant correlation with age at sampling. The difference between the two reads was ~16% and they were shown to be highly correlated with each other (P value < 0.0001 and the $r^2 = 0.99$) (Figure 5.7). As both counts provide a similar measure and are highly correlated, only one measure was used for additional analysis. The CLC-count was selected as this measure.

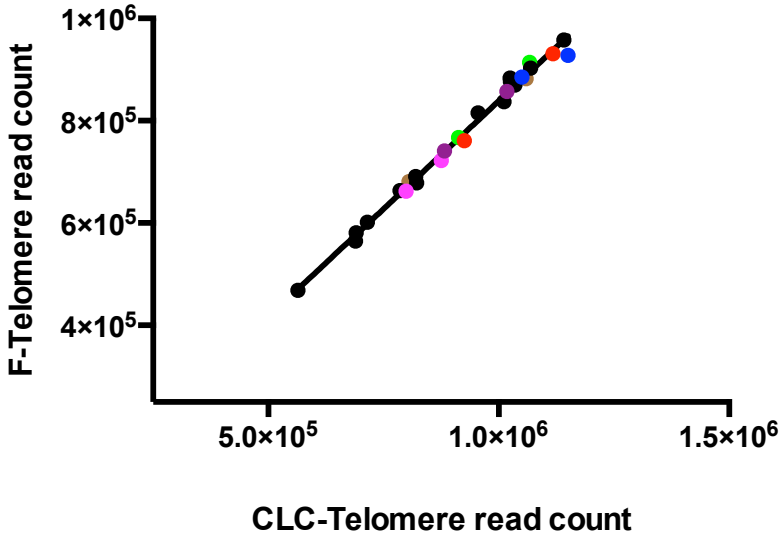


Figure 5.7: Linear regression analysis: correlation between F and CLC- telomere read count

5.4.3 Using the B or C-samples of pairs to estimate the telomere length

The telomere count was previously used to show the correlation with age at sampling among six individuals whose blood samples were collected at two different ages. The B-sample was the first to be collected and the C-sample was the more recent sample. To exclude any inflation in the correlation that could arise from these dependent samples in the previous analysis, the CLC-telomere count for either the B-samples or C-samples were used independently to calculate the effect of age at sampling (Figure 5.8). Examination of the C-samples showed a significant correlation between age at sampling and the difference between telomere counts conducted. The P value was < 0.0001 and $r^2 = 0.61$. Similarly, the B-samples were also used separately to conduct the same analysis (P value < 0.0001 and $r^2 = 0.62$) (again, as shown in Figure 5.8). A linear regression analysis using B- or C-samples independently showed a marginally higher significant correlation with age at sampling.

Moreover, all six pairs go down with significant P value = 0.02. Therefore, B- and C-samples will be used together in further analysis.

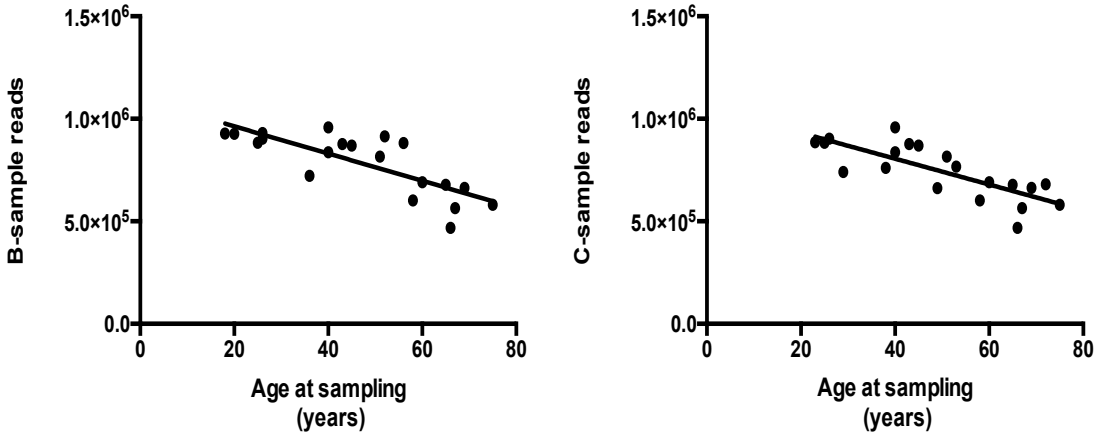


Figure 5.8: Separate examinations of C-samples and B-samples to measure the effect of age at sampling.

5.4.4 Telomere length analysis using corrected telomere reads

The first aim of this experiment, to capture the telomere regions, was successfully achieved. Our strategy was to use the captured telomeric reads (T) and the reads generated from sequencing 233 DNA repair genes (U) to calculate the relative telomere length (T/U). Reads that were derived from X-Y chromosomes were excluded so only autosome gene reads were used. The relative telomere length (T/U) was obtained by dividing the CLC-telomere reads by the unique autosome gene reads, which showed a marginally significant correlation with age at sampling (r^2 = 0.15 and the P value = 0.052). The obtained result was used to estimate age-dependent telomere shortening. A linear regression analysis using corrected telomere reads (T/U) showed a significant correlation with age at sampling (Figure 5.9) (P value = 0.006 and r^2 = 0.27). The corrected telomere reads generated using NGS showed a significant correlation with age at sampling as expected.

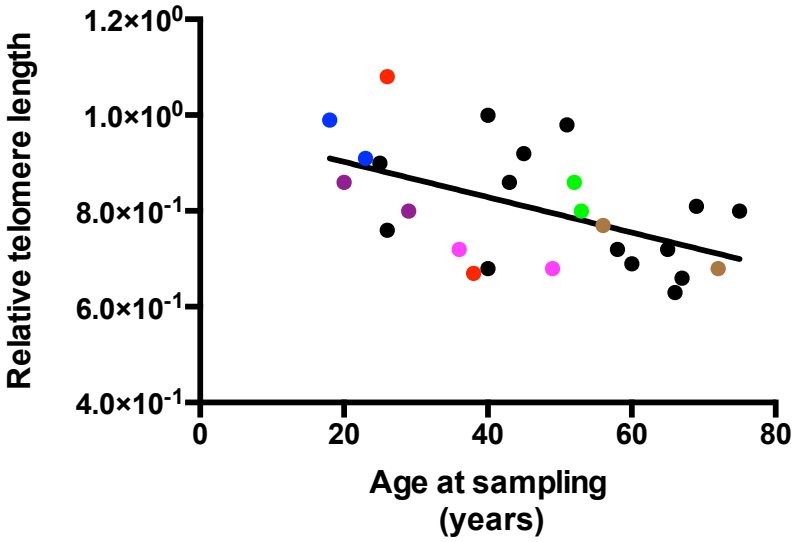


Figure 5.9: Linear regression analysis. Age-dependent telomere shortening using calculated relative telomere length (T/U).

Ideally, the r^2 value would be high, as would be the level of accuracy of age prediction. In forensics, one aim of age estimation is to minimise the number of suspects and the number of tests which need to be carried out to identify them. However, in the present case, the r^2 coefficient of the correlation value was relatively low. Therefore, in order to determine how useful it might be, we investigated the predictive power of our analysis. Firstly, we investigated the relative confidence in the regression line to determine the 95% confidence interval (Figure 5.10). The 95% confidence interval gives insight into the accuracy of the mean relationship in the variables across the population. The results indicated that the confidence interval is relatively low.

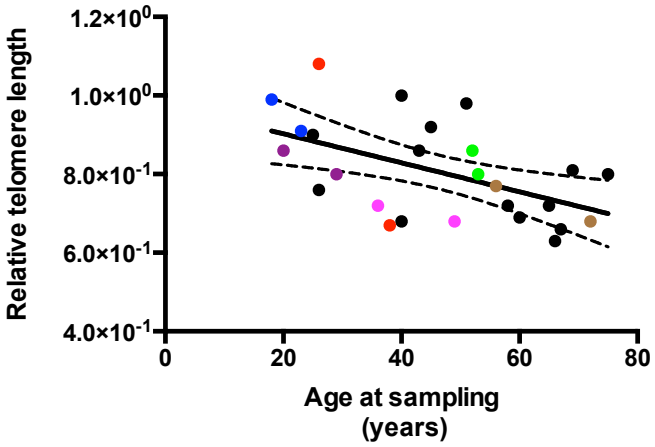


Figure 5.10: 95% confidence interval analysis. The relative telomere data were used to determine the confidence interval.

Because the r^2 value is not at 100%, variation still exists in the population which has not yet been revealed. Using an age range would be the best way to estimate rather than using a point age estimate. A point age estimation can be calculated from the relative telomere length linear regression analysis (Y = -74.33*X + 106.1). The 95% prediction interval (PI) is used to estimate where future observations will fit within the range interval calculated based on previously observed data and can be used to give the best estimate of age range for samples collected from a crime scene (Figure 5.11). The calculated age range from a 95% PI is very wide. For instance, at relative telomere ratio 0.8, it was 47 ± 27 years.

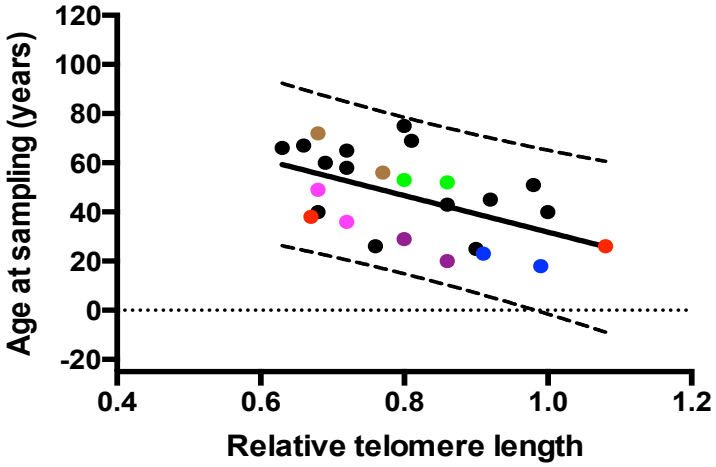


Figure 5.11: 95% prediction interval. The relative telomere data used to determine the prediction interval to give the best estimate.

Both of the aims of this experiment were achieved. Normalising telomere reads using autosome gene reads would be expected to improve the correlation between telomere reads and age at sampling, but the results were unexpected and the significant correlation remained, but with decreased values. Therefore, remodelling using data from Chapter Four was carried out using R language. The data generated from our NGS experiments was used to develop a simple linear statistical model, which is shown in Table 5.2.

Table 5.2: Data generated by NGS used to create linear modelling

Sample ID	*Age at sampling (years)*	*Total reads*	*CLC-telomere count*	*Autosome gene reads*	*Microsatellite reads*
DMG119C	75	5,745,730	580,793	726,045	1,395,726
DMG18C	72	8,010,376	680,804	1,006,365	1,983,579
DMG99C	69	6,527,096	663,381	819,907	1,492,667
DGM129C	67	6,672,786	564,381	859,047	1,680,458
DGM133C	66	5,776,950	468,475	747,431	1,453,850
DMG22B	65	7,328,262	678,665	943,397	1,824,668
DMG106C	60	7,693,758	690,720	1,006,294	1,874,588
DMG118C	58	6,574,880	601,446	832,304	1,552,375
DMG18B	56	9,043,646	882,002	1,149,794	2,260,023
DMG13C	53	7,572,124	767,220	956,709	1,794,679
DMG13B	52	8,814,984	913,823	1,066,055	1,943,490
DGM92C	51	6,849,546	815,312	829,697	1,643,340
DMG29C	49	7,838,174	662,102	971,075	1,904,067
DMG48C	45	7,794,014	869,824	948,691	1,859,796
DMG47C	43	8,293,476	876,868	1,019,652	1,890,750
DMG108C	40	7,752,890	958,051	957,069	1,938,405
DGM134C	40	9,484,008	837,064	1,230,412	2,344,811
DMG70C	38	8,632,068	760,851	1,140,108	2,202,607
DMG29B	36	7,815,304	722,003	999,015	1,970,763
DMG57C	29	7,142,628	741,240	924,856	1,806,298
DMG70B	26	9,164,438	930,837	858,211	2,237,907
DGM91C	26	7,095,050	902,513	1,183,857	1,709,950
DMG86C	25	7,955,916	883,251	986,108	1,784,445
DMG51C	23	7,675,374	885,102	971,734	1,909,609
DMG57B	20	7,768,158	857,267	992,673	1,924,012
DMG51B	18	7,742,960	927,786	934,814	2,053,284

5.5 Linear model of age at sampling and telomere reads

Linear modelling between telomere reads and age at sampling was carried out to obtain a formula to predict the age of a person. The r^2 value from the linear regression analysis normally indicates how good the fit to the line of regression is.

5.5.1 Model formulation

A linear regression model relates a dependent variable, *y,* to a set of independent variables, $x_1, x_2, \ldots, x_{p-1}$ as:

$y = a_0 + a_1x_1 + a_2x_2 + \ldots a_{p-1}x_{p-1} + \ldots,$ where a_0 is the intercept, a_i; $i = 1, 2, \ldots,$ p-1 are the coefficients of the regression and p denotes the number of parameters in the model. ε is the error term.

Fitting a linear model took place in the R language (an environment for statistical computing and graphics), using the lm function together with the following model form:

reg < -lm (Age at sampling (years) ~ P1+ P2 + P3 + P_n).

Six models were developed using R (Table 5.3). Two of these models will be discussed in more detail.

5.5.1.1 Modelling age-dependent telomere shortening

The lm function was used together with telomere reads and age at sampling with the following model form:

reg1 <-lm (Age at sampling (years) ~ Telomere count)

The results obtained from the regression analysis of this model are shown in Table 5.3.

Table 5.3: Linear regression model of telomere reads

Parameters	Estimated	Std. Error	t value	P value
Intercept	1.196e+02	1.438e+01	8.318	1.57e-08 ***
Telomere count	-9.479e-05	1.832e-05	-5.175	2.67e-05 ***

Note: Significant codes: 0 *** 0.001 ** 0.01 * 0.05,0.1,1

The results showed a significant correlation to exist between age at sampling and telomere count, similar to the results obtained from a simple linear regression for the uncorrected telomere reads. The results showed the adjusted $r^2 = 0.51$ and the P value = 2.67e-05. The results in the table can therefore be expressed as: $Y = 1.196e+02 - 9.479e\text{-}05X_{\text{telomere}}$

5.5.1.2 Modelling telomere, microsatellite and autosome gene reads

Models that include the microsatellite, autosome, and telomere reads were developed using the lm function with the following model formula:

reg4 < lm (Age at sampling (years) ~Telomere count + Autosome gene reads + Microsatellite reads)

The results obtained from the regression analysis of this model are shown in Table 5.4.

Table 5.4: Linear regression model of telomere reads and autosome genes

Parameters	Estimated	Std. Error	t value	Pr(>|t|)
Intercept	1.209e+02	2.136e+01	5.658	1.09e-05***
Telomere count	-9.250e-05	2.440e-05	-3.791	0.001***
Microsatellite reads	-5.597e-06	1.640e-05	-0.341	0.736
Autosome genes	7.652e-06	3.012e-05	0.254	0.802

Note: Significant codes: 0 *** 0.001 ** 0.01 * 0.05.0.1.1

In this model, only the telomere reads showed a highly significant correlation with age at sampling and neither microsatellite nor autosome gene reads showed significant correlations. The results showed the multiple $r^2 = 0.53$, the adjusted $r^2 = 0.47$ and the model P value = 0.0007 for the model. The results in the table can therefore be expressed as:

$$Y = 1.209e+02 \ -9.250e-05\ X_{telomere} \ -5.597e-06 X_{microsatellites} + 7.652e-06\ X_{autosome}$$

5.5.2 Model checking

In developing the model and using it to draw conclusions, it is necessary to check that it does fit the data. Therefore, different ways of checking the assumptions of the model are shown in Figure 5.12. The top-left plot, of residual versus fitted (predicted) values, indicates whether there is any random variation, without any systematic patterns in either location or scale. The top-right plot, which is a scale-location plot, checks for any possible changes in scale using the square root of the residuals and in doing so, no discernable pattern to the plot should be observed. The bottom left plot, a normal quantile-quantile plot (Q-Q plot), checks whether the residuals of the data are normally distributed or not. The bottom-right plot, when calculating the samples average, checks that each observed value contributes evenly (*i.e.* with the same weight) to determine the average value. However, in a regression analysis, each observation has an important role. A Cook's distances plot measures the significance of each observation to the regression. Observations with smaller distances have little effect on the regression results when they are removed. Observations

with distances larger than 1 are suspicious and suggest the presence of a possible outlier or a poor model. See Figure 5.12.

The results show that there is a random variation and no trend in the mean of the residuals, and the Q-Q plot indicates that the residuals follow a normal distribution, as the data approximately produce the straight line of a normal distribution. Finally, the scale-location plot shows that the residual magnitudes seem consistent with constant variance. To sum up, the top-left, top-right and bottom-left panels confirm that our assumptions, that the telomere count decrease with age at sampling and the data obtained from NGS appear to be valid.

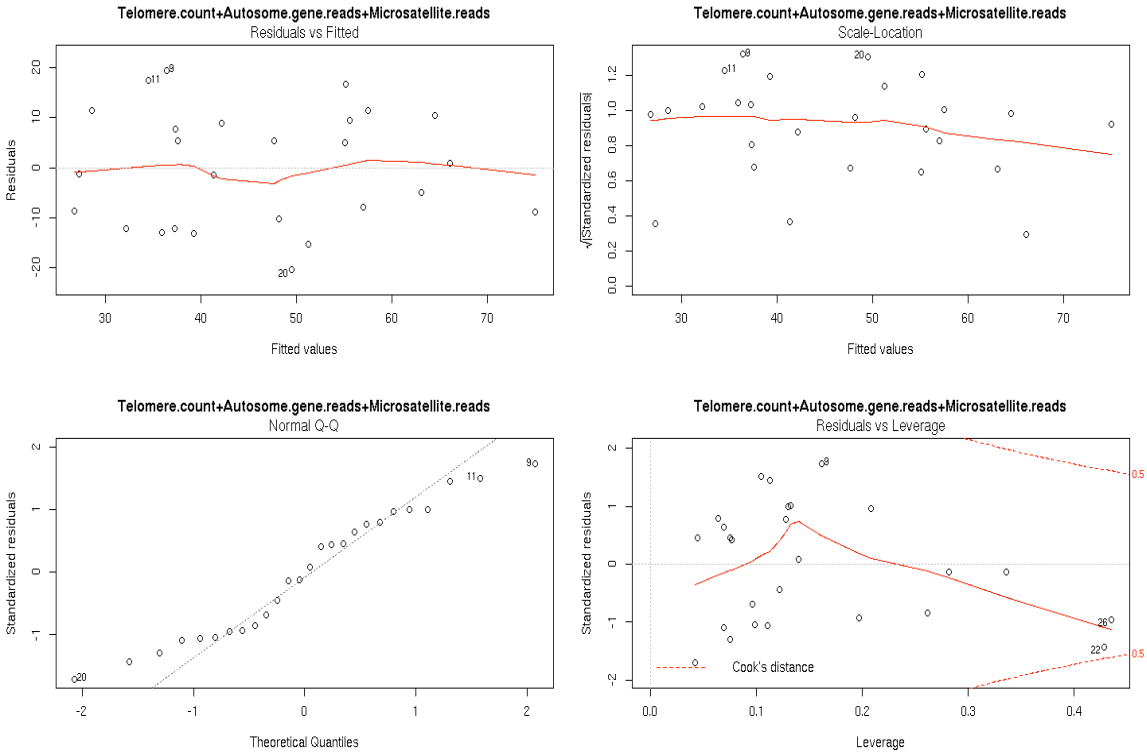


Figure 5.12: Telomere reads, microsatellite and autosome gene reads model checking. Top-left plot: Residuals against the fitted values; Top-right: Square root of abs values of residuals against the fitted values (scale-location plot); Bottom-left: Normal probability plot; Bottom-right: Cook's distance plot.

Relative telomere length was also used to develop similar linear models using R (Table 5.5). The simple linear regression (model 1) using relative telomere showed that $r^2 = 0.27$ and the model P value = 0.006. When relative telomere reads and an autosome gene read count were used to develop linear model 2, an improvement was evident in the adjusted $r^2 = 0.47$ and significant model P value = 0.0002 (Table 5.5) and both parameters were significant. A significant model 3 was obtained when the relative telomere length and microsatellite reads were used. In this case, the adjusted $r^2 = 0.38$ and the significant model P value = 0.002 (Table 5.5). When the relative telomere length, autosome gene reads and

microsatellite reads were used to generate model 4, the model was significant (adjusted r^2 = 0.46 and model P value = 0.001) however only the relative telomere length was significant, while autosome gene reads parameters were marginally significant (Table 5.5). The microsatellite reads P value was insignificant, indicating that any value which they had added had not, in fact, improved the model; the autosome added similar but significant value. Clearly, the data from all the models generated produced the same results which had previously been obtained from the uncorrected telomere reads (adjusted r^2 = 0.51). This means that the internal parameters used to generate these models cannot be used to explain what led to these results. Another factor that may contribute to the observed results is the finished DNA library size. Previously, the autosome reads showed a declining trend with age at sampling and with the finished DNA library size. Correcting the relative telomere length (r^2 = 0.24; P value = 0.006) by dividing its value by the finished DNA library size to exclude any effect generated on autosome genes reads leads to the unexpected result of corrected telomere reads with age at sampling. This analysis has improved the correlation value. A linear regression analysis showed a P value = 0.002 and r^2 = 0.30 which may partially explain some of the results observed when using relative telomere length to estimate age-dependent telomere shortening (Figure 5.13).

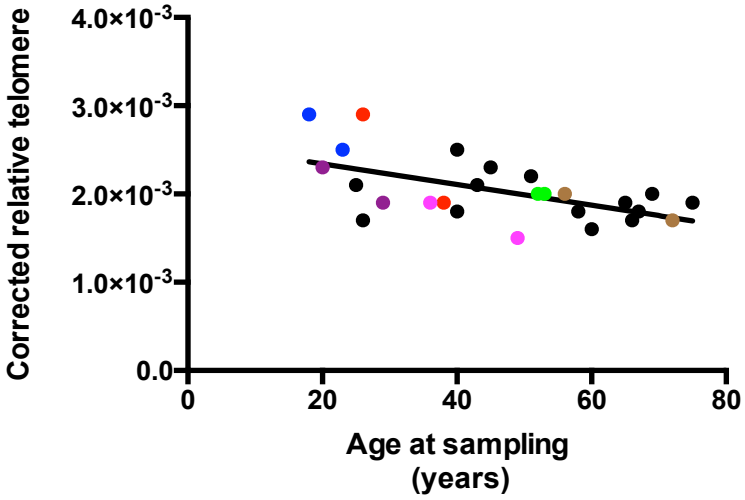


Figure 5.13: Corrected relative telomere length using finished DNA library size

Table 5.5: Linear models generated using R language

Model 1	**Age at sampling years and relative telomere length**			
	Multiple r-squared	**Adjusted r-squared**	**P value**	**Comment**
Parameters	0.27	0.24	0.006	Significant
Relative telomere length			0.006	Significant
Model 2	**Age at sampling years, relative telomere length, and autosome gene reads**			
	Multiple R-squared	**Adjusted R-squared**	**P value**	**Comment**
Parameters	0.52	0.47	0.0002	Significant
Relative telomere length			0.0004	Significant
Autosome gene reads			0.003	Significant
Model 3	**Age at sampling years, relative telomere length, and microsatellite reads**			
	Multiple R-squared	**Adjusted R-squared**	**P value**	**Comment**
Parameters	0.43	0.38	0.001	Significant
Relative telomere length			0.007	Significant
Microsatellite reads			0.01	Significant
Model 4	**Age at sampling years, relative telomere length, autosome gene and microsatellite reads**			
	Multiple R-squared	**Adjusted R-squared**	**P value**	**Comment**
Parameters	0.51	0.45	0.001	Significant
Relative telomere length			0.001	Significant
Autosome gene reads			0.06	Not significant
Microsatellite reads			0.8	not significant

The difference between the observed and predicted values (*i.e.* the residuals) of autosome gene reads and age and CLC-telomere reads and age was used to calculate the relative telomere length using R. The residuals were found to be correlated to each other, and a significant correlation was observed ($r^2 = 0.16$; P value = 0.04), as shown in Figure 5.14. This indicates that both reads were affected by age in a similar pattern; a decreasing count was noticed aligned with increasing age at sampling.

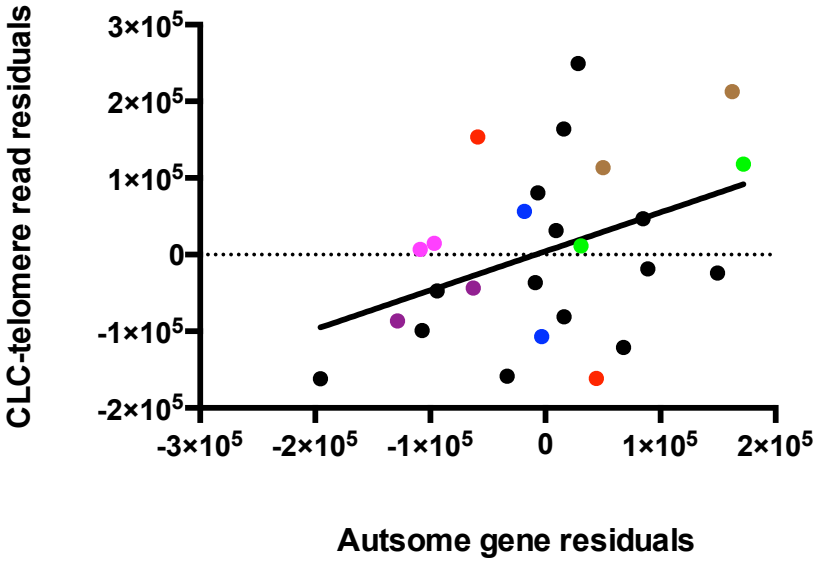


Figure 5.14: Residuals correlation with age at sampling: Calculated CLC-telomere reads residuals and autosome reads residuals.

5.5.3 : Human telomere variant repeats

Human telomeres do not consist of pure $(TTAGGG)_n$ repeats through their entire length but instead encompass a minimum of three motifs (Allshire *et al*,1989; Allshire and Hastie, 1989). These motifs are TTAGGG, TTGGGG, and TGAGGG. TTGGGG and TGAGGG are reportedly located at the end of the chromosome and at the start of the telomeric region and to have a similar sequence arrangement at the end of most human chromosomes (Allshire and Hastie, 1989). Telomere sequence variance has been used to develop a PCR-based approach (TVR-PCR) to map the distribution of the telomere (TTAGGG) and variant repeats (TGAGGG and TCAGGG) at the proximal end of the telomere repeat array (Baird *et al*., 1995). These motifs are differing by one base, which enable research to develop a method determining the distribution of telomere variant repeats. Briefly, a radioactively labelled allele-specific flanking primer together with one of three 'tagged' telomere or variant repeats primers were used to amplify genomic DNA. Polyacrylamide gel was used to resolve the PCR product and the detection was carried out by

autoradiography. The scatter pattern of the telomere motif can be determined (Baird *et al*., 1995).

Baits were used to capture the telomeric regions and only consists of TTAGGG repeats. The telomere read count was firstly analysed using CLC Genomic Workbench to see if the reads captured contained all three of these motifs, and the results are shown in Figure 5.15.

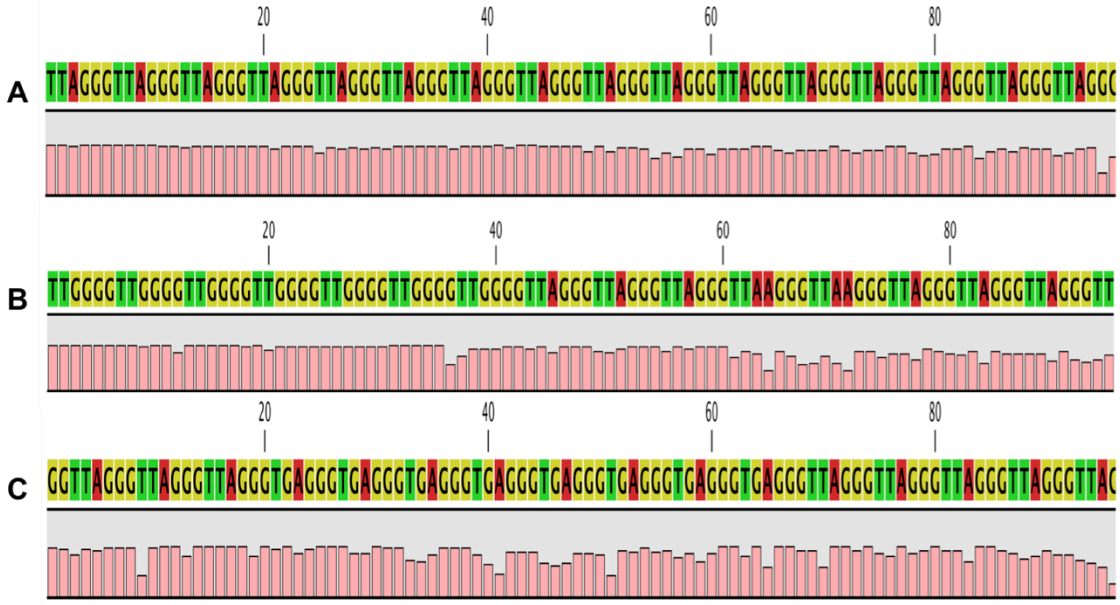


Figure 5.15: Telomere motifs. Three telomere motifs were observed A): TTAGGG; B): TTGGGG and C): TGAGGG.

The data analysis showed the presence of the three patterns. Long tract reads containing pure TTAGGG repeats were the most commonly observed. TTAGGG repeats interrupted with TTGGGG or TGAGGG were also observed. A bioinformatics analysis by Dr. Graham Hamilton was conducted to quantify the number of reads comprising these motifs using Bowtie in each sample (Table 5.6). Any NGS read containing ≥3 repeats was counted. The count was first conducted for TTAGGG and was then repeated for the other two motifs. Reads containing TTAGGG repeats represented 99% of the total telomere captured. Reads comprising TTGGGG represented 14% and reads including TGAGGG represented 7% of the total telomere reads counted. The data obtained showed an overlap between reads, as some telomere reads contained more than one motif. Thus, a new count approach was adopted. Using F-telomere reads, the number of repeats of each motif was counted rather than the number of reads containing these motifs, and the resulting data are shown in Table 5.7. Both TTGGGG and TGAGGG were reported to be located at the chromosome termini and far from the attrition region (Allshire *et al*., 1989). So, their numbers are expected to be constant throughout a person's lifetime, unlike the TTAGGG number, which should decrease with age. The three patterns were identified, and a linear

regression analysis was conducted to investigate how the numbers behaved with ageing. The results for each pattern are discussed below.

Table 5.6: Bioinformatics analysis to quantity reads containing three telomere motifs

Sample ID	*Total telomere count*	*TTAGGG*	*%*	*TTGGGG*	*%*	*TGAGGG*	*%*
DMGV119C	690,948	682572	99%	100404	15%	42604	6%
DMGV18C	805,418	793366	99%	119528	15%	54310	7%
DMGV99C	785,496	775768	99%	108112	14%	55768	7%
DMGV129C	688,994	677342	98%	105830	15%	52480	8%
DMGV133C	564,474	555518	98%	83444	15%	38062	7%
DMGV22B	822,142	810672	99%	105348	13%	52676	6%
DMGV106C	820,108	808796	99%	125482	15%	48018	6%
DMGV118C	715,020	705010	99%	112794	16%	53118	7%
DMGV18B	1,059,334	1044398	99%	140500	13%	68762	6%
DMGV13C	912,632	899412	99%	121548	13%	62214	7%
DMGV13B	1,067,082	1054014	99%	143454	13%	74958	7%
DMGV92C	955,368	945398	99%	127614	13%	59878	6%
DMGV29C	798,736	786670	98%	118840	15%	72596	9%
DMGV48C	1,034,746	1023474	99%	138726	13%	75568	7%
DMGV47C	1,025,382	1012958	99%	143454	14%	68284	7%
DMGV108C	1,141,128	1126806	99%	150482	13%	67170	6%
DMGV134C	1,011,128	995750	98%	155950	15%	78680	8%
DMGV70C	925,140	910058	98%	130252	14%	56118	6%
DMGV29B	875,106	861748	98%	122496	14%	65534	7%
DMGV57C	882,282	870478	99%	129512	15%	58160	7%
DMGV70B	1,117,424	1103262	99%	135800	12%	64144	6%
DMGV91C	1,068,920	1057996	99%	150282	14%	69724	7%
DMGV86C	1,024,374	1013246	99%	138500	14%	68270	7%
DMGV51C	1,050,036	1036244	99%	139882	13%	66166	6%
DMGV57B	1,018,023	1005568	99%	140227	14%	69605	7%
DMGV51B	1,150,570	1136094	99%	137500	12%	68558	6%
Average	**923,462**	**911,254.54**	**99%**	**127922**	**14**	**61978**	**7%**

Table 5.7: Telomere variant count in F-telomere reads

	Repeat count				
Sample ID	**TTAGGG**	**TTGGGG**	**TGAGGG**	**Total**	**F- read count**
DMGV119C	10,052,128	541,804	361,461	10,955,393	690,948
DMGV18C	11,826,479	628,540	427,239	12,882,258	805,418
DMGV99C	11,685,472	590,072	440,258	12,715,802	785,496
DMGV129C	9,787,762	547,690	401,833	10,737,285	688,994
DMGV133C	8,233,912	436,067	298,958	8,968,937	564,474
DMGV22B	11,941,685	588,278	440,484	12,970,447	822,142
DMGV106C	11,877,029	661,457	404,345	12,942,831	820,108
DMGV118C	10,334,988	586,688	423,633	11,345,309	715,020
DMGV18B	15,509,923	758,951	544,681	16,813,555	1,059,334
DMGV13C	13,723,773	660,993	493,646	14,878,412	912,632
DMGV13B	16,159,870	769,311	579,594	17,508,775	1,067,082
DMGV92C	14,018,943	725,252	524,591	15,268,786	955,368
DMGV29C	10,733,766	641,430	541,000	11,916,196	798,736
DMGV48C	14,352,478	783,893	612,544	15,748,915	1,034,746
DMGV47C	15,467,483	771,059	550,826	16,789,368	1,025,382
DMGV108C	17,143,918	840,260	585,565	18,569,743	1,141,128
DMGV134C	14,470,579	798,971	583,912	15,853,462	1,011,128
DMGV70C	13,484,748	695,963	475,086	14,655,797	925,140
DMGV29B	12,571,407	657,170	512,802	13,741,379	875,106
DMGV57C	12,980,290	678,681	462,588	14,121,559	882,282
DMGV70B	16,595,290	774,299	565,458	17,935,047	1,117,424
DMGV91C	15,359,192	837,727	598,245	16,795,164	1,068,920
DMGV86C	15,807,599	744,574	530,792	17,082,965	1,024,374
DMGV51C	15,500,121	772,265	544,964	16,817,350	1,050,036
DMGV57B	15,131,390	738,186	532,632	16,402,208	1,018,023
DMGV51B	17,040,536	781,865	535,901	18,358,302	1,150,570
Average	13,530,414	692,748	498,963	14,722,125	923,462
% of total repeat	**92**	**5**	**3**	**100**	

5.5.3.1 TTAGGG age-dependent shortening

The TTAGGG repeats represent 92% of the total telomere repeats counted. The correlation with age was calculated for uncorrected ($r^2 = 0.53$ and $P < 0.0001$) and corrected counts, using a linear regression analysis ($r^2 = 0.29$ and P value = 0.005), as Figure 5.16 shows.

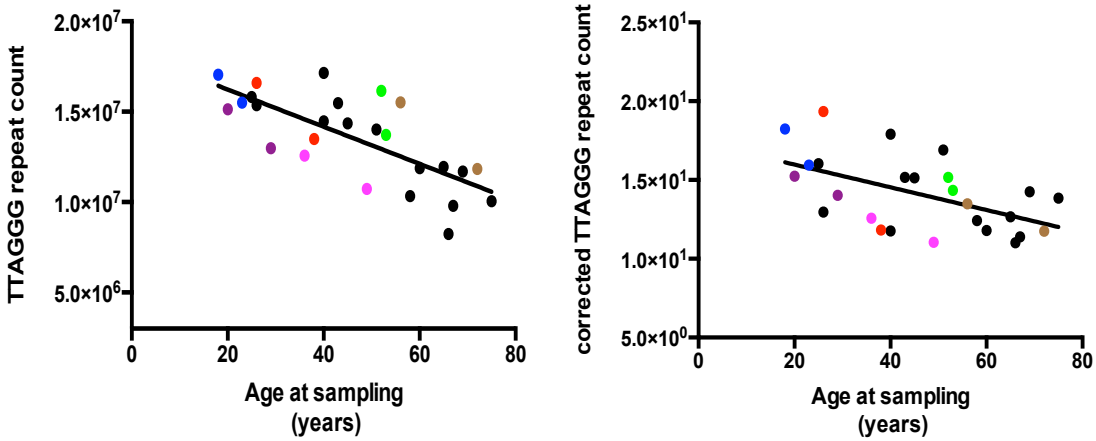


Figure 5.16: TTAGGG-repeats age-dependent shortening.

5.5.3.2 TTGGGG age-dependent shortening

TTGGGG repeats constitute 5% of the total telomere repeats counted. Similarly, the correlation with age at sampling was calculated for uncorrected and corrected TTGGGG - telomere repeats. As TTAGGG, the linear regression analysis showed highly significant correlation between uncorrected TTGGGG repeats and age at sampling ($r^2 = 0.52$ and $P < 0.0001$). The autosome gene reads were used to correct the TTGGGG repeats and the corrected count also showed significant correlation with age at sampling ($r^2 = 0.24$ and P value = 0.01) (Figure 5.17).

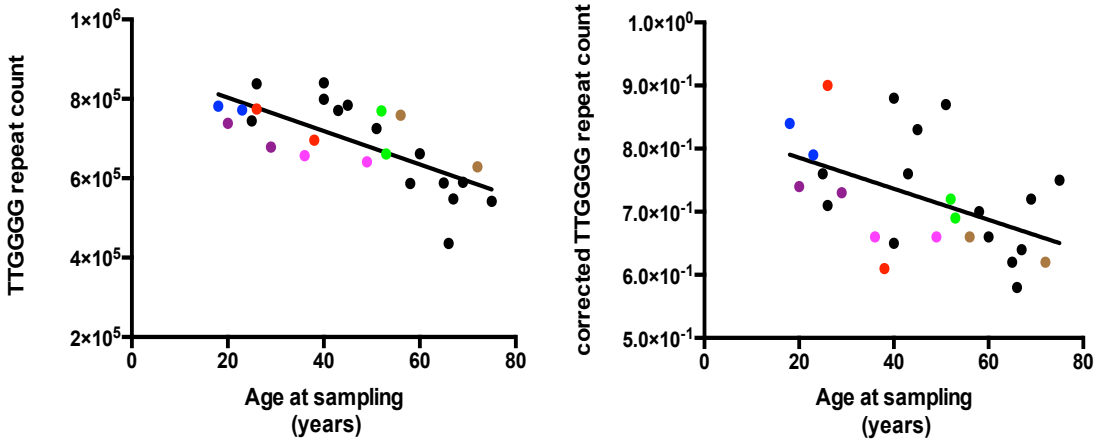


Figure 5.17: TTGGGG-repeats age-dependent shortening

5.5.3.3 TGAGGG age-dependent shortening

Finally, TGAGGG-telomere repeats, which represent only 3% of the total telomere repeats recorded, were also analysed. The correlation with age at sampling was calculated for both a corrected and an uncorrected count. A linear regression analysis showed a significant correlation between the uncorrected TGAGGG repeats count and the age at sampling (r^2 = 0.44 and P value = 0.0002). The corrected repeats showed a significant correlation with age at sampling (r^2 = 0.20 and P value = 0.02) as shown in Figure 5.18.

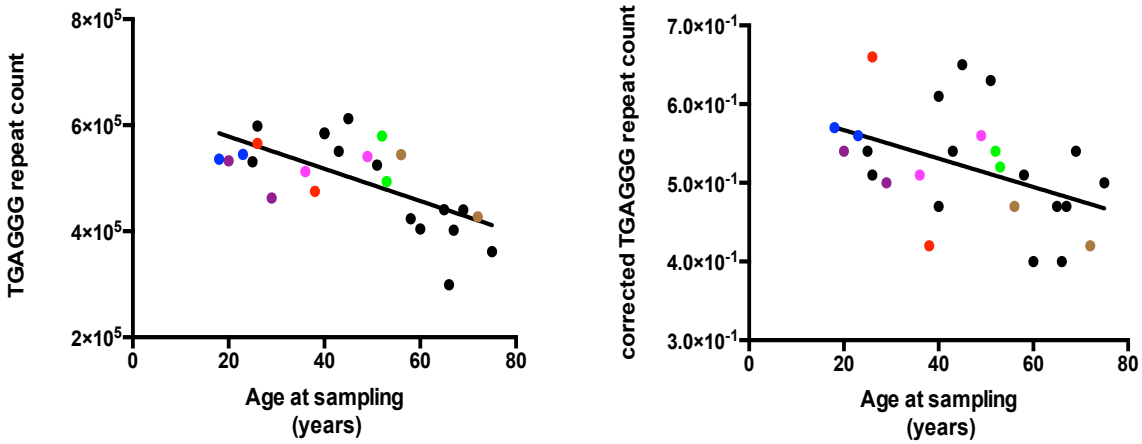


Figure 5.18: TGAGGG-repeats age-dependent shortening.

The TGAGGG and TTGGGG repeats only differ from the most common telomere repeat (TTAGGG) by a single base, where T > G or A>G. To rule out the possibility of sequencing errors resulting in such variations between repeats, CLC Genomic workbench was used to quantify the presence of pure TTAGGG repeats and other variants. Six custom references, each containing ten repeats, were generated, of which the first custom reference included pure TTAGGG repeats. In the other five references, only the centre repeat was replaced by one variant repeat. Sample 119C was selected and the NGS reads were mapped to the custom references (Table 5.8). Using high stringency parameters, the number of reads mapped to the six custom references was 8279. Reads mapped to the reference containing pure TTAGGG repeats were the most observed repeats, with 4,593 reads (56% of the total reads mapped). The reads containing TTGGGG repeats mapped to the custom reference totalled 2,160 reads (~ 26% of the total reads mapped), and the reads mapped to the custom reference containing TAAGGG totalled 1445 reads (~ 18% of the total reads mapped). Seventy-three reads were mapped containing TGAGGG repeats (0.009% of the total reads mapped) whereas only three and five reads containing TTAGGC and TTACGG reads, respectively, were mapped to the custom references (0.0004 and

0.0006% of the total reads mapped, respectively). The data showed that TTAGGG, TTGGGG, and TGAGGG are likely to be true telomere variants, a finding which is consistent with earlier studies (e.g. Allshire *et al*., 1989). The reads containing TTACGG and TTAGGC were very low and may reflect sequencing errors. Moreover, the data revealed that 18% of the reads mapped contained TAAGGG repeats; a search of the literature shows that this sequence variant has previously also been reported to exist in human telomeres (Baird *et al*., 1995).

Table 5.8: Quantification analysis of telomere motifs to rule out sequencing errors

Custom reference	TTAGGG	TTGGGG	TGAGGG	TAAGGG	TTAGGC	TTACGG	Total
Read count	4,593	2,160	73	1,445	3	5	8,279
% of total count	56	26	0.009	18	0.0004	0.0006	

5.5.4 Telomere shortening in paired samples

Paired samples were used to estimate the telomere shortening for each year (Table 5.9). Sample 18 will now be discussed in detail. The total telomere repeats count was used to conduct this calculation. In all paired samples, the B-samples (which were collected at early ages) showed higher repeat counts than did the C-samples. For example, in sample 18, the B-sample was collected at 56 years of age with 16,813,555 repeats, whereas the C-sample was collected more recently at an age at sampling of 72 years, with 12,882,258 repeats. The read coverage was calculated by dividing the total number of autosome DNA repair gene reads counted, by the total number of DNA baits (8,252). The results showed that the read coverage of 18B was 138 (1,149,794/8,252) and 122 for 18C (1,006,365/8,252).

The results obtained were then used to calculate the number of repeats for a single read coverage and the results of this calculation were 120,670 repeats for 18B and 105,632 repeats for 18C. The age difference for each paired sample was calculated (C-B) and the results showed 16 years' difference (72-56) between 18B and 18C. The repeats difference per single coverage was also calculated between each paired samples (B-C) and showed 15,038 repeats (120,670 - 105,632). These results were used to calculate the number of repeat shortening for each chromosome end (15,038 / (92/2) = 327 repeats). Finally, the numbers generated were used to calculate the number of repeat shortening per year

(327/16) and the result was ~20 repeats shortening per year for sample 18. In order to estimate telomere shortening for each paired sample in bp, the result obtained was multiplied by 6 bp, the repeat size, producing a figure of ~120 bp shortening /year for sample 18. The same analysis was conducted for all the other paired samples. The estimated amount of telomere shortening was also calculated for the present experiment's data using the linear regression formula (Y = -110053*X + 1.981e+007; X=age) to predict the repeat count for the oldest (75 years) and youngest (18 years) individuals. The repeats obtained for them were 11,556,025 and 17,829,046 repeats, respectively. A similar analysis was conducted to identify the number of repeats /year. The results which were obtained showed that the estimated telomere shortening was 60 bp /year on average in this population.

Table 5.9: The estimation of telomere shortening in paired samples

Sample ID	Age at sampling	Total telomere repeat count	Autosome read count	Read coverage	Number of repeat/ coverage	Age difference (C-B)	Repeats difference (B-C)	Number of repeats/ Chromosome ends	Number of repeats/Year	Bp
DMGV18C	72	12,882,258	1,006,365	122	105,632	16	15,038	327	20	120
DMGV18B	56	16,813,555	1,149,794	139	120,670					
DMGV13C	53	14,878,412	956,709	116	128,332	1	7,198	156	156	936
DMGV13B	52	17,508,775	1,066,055	129	135,530					
DMGV29C	49	11,916,196	971,075	118	101,261	13	12,244	266	20	120
DMGV29B	36	13,741,379	999,015	121	113,506					
DMGV70C	38	14,655,797	1,140,108	138	106,077	12	66,374	1443	120	720
DMGV70B	26	17,935,047	858,211	104	172,452					
DMGV57C	29	14,121,559	924,856	112	125,999	9	10,351	225	25	150
DMGV57B	20	16,402,208	992,673	120	136,350					
DMGV51C	23	16,817,350	971,734	118	142,814	5	19,243	418	84	504
DMGV51B	18	18,358,302	934,814	113	162,057					
		The estimation of telomere shortening in data								
DMGV119C	75	11,556,025	934,814	88	131,318	57	26,461	575	10	60
DMGV51B	18	17,829,046	726,045	113	157,779					

5.6 Discussion

The main aim of this experiment was to use NGS to develop a new and novel method to capture telomeric regions and to use the generated data to measure the relative telomere length, in order to predict the age of a person. A DNA target enrichment approach was selected. Therefore, biotinylated RNA custom baits were generated that solely include telomeric repetitive sequences with an average size of 120 bp $(TTAGGG)_{20}$ which will be easily captured onto streptavidin-labelled magnetic beads. We used 7,667 bait copies to capture telomeric regions in 26 samples, with an age range of between 18 and 75 years, representing 13.3% of the total number of baits used.

The DNA target enrichment and NGS approach successfully generated reads that mapped to the telomeric regions on the human reference genome (see Figure 5.2, above). Two read count approaches were employed, to ensure that short telomeric reads and reads including variants present in telomere repeats would not be missed, nor any errors generated by sequencing (as described above). Both count approaches generated telomeric reads with an average ~923,462/sample. The first aim of this experiment was achieved.

Telomere length is known to decrease with age and the age of a person can be estimated from their telomere length. The data generated from our approach were used to achieve our second goal. Firstly, the total telomere repeat count was used to investigate the age-dependent telomere shortening and a highly significant correlation with age at sampling was identified for both F and CLC counts ($r^2 = 0.53$; P value <0.0001). Both counts were highly correlated and had similar measures (P value =< 0.0001 and the $r^2= 0.98$). Therefore, the CLC-count was selected for the conducting of further analysis.

The samples used to produce these data included dependent samples from 6/20 individuals whose blood samples were collected at two different ages (B and C). The CLC-count using either B- or C-samples was used to study the effect of these dependent samples. A linear regression analysis carried out using independent samples showed a marginally higher significant correlation with age at sampling. All the paired samples significantly decrease with age (P> 0.0156). Excluding B- or C- samples may result in excluding useful data for the analysis that may lead to an overestimation of the effect of age at sampling on telomere length. Thus, all the data generated from this experiment were used to conduct additional analysis.

The uncorrected telomere reads showed a significant correlation with age at sampling (r^2 = 0.53), similar to a TRF conducted by Takasaki *et al*. in 2003, which produced a result of r^2 = 0.562. However, our strategy was to use the relative telomere length to predict human age. The autosome DNA repair gene reads (U) counted showed no significant correlation with age at sampling and the generated telomere read (T) was used to calculate the relative telomere length (T/U); this showed a significant correlation with age at sampling (P value = 0.006 and r^2 = 0.27). The obtained r^2 was low; this means it may not have any practical use in estimating age in forensic investigations. The r^2 value is normally used to indicate the suitability to fit to a line of regression and would reflect how good the age prediction is. Using the data generated to estimate human age, our model generated a formula (Y = -74.33*X + 106.1) where Y is the predicted age depending on X which is the size of the relative telomere length, observed from the samples which have been collected as evidence in crime.

A range of years would be appropriate for use in criminal investigations, as it would minimise the number of suspects who would have to be considered by investigators, and the number of tests to be conducted, thus reducing the time and cost required. So, the 95% prediction interval was calculated for model 1 (relative telomere length) to estimate the age range. For example, if the telomere reads obtained from a crime scene revealed that the relative telomere length (T/U) identified was 0.8, the predicted age would be calculated using model 1 in which the linear regression formula is (Y = -74.33*X + 106.1= ~47 years). The age range for this suspect would therefore be (47 ± 27 years). In criminal forensics, when trying to identify suspects, most of their ages fit between 20-60 years which in turn means that the vast majority of suspects fit within our age range and in this case the result has little power to define a narrow enough range to be useful. Nonetheless, in the case of an unknown victim where the age range is between 0 and 100 years, our model may have some application.

We expected to generate a more significant correlation by conducting an NGS that is known to be a highly quantitative method, especially from TRF conducted by Takasaki *et al*. (2003), in which r^2 = 0.562, and Tsuji *et al*. (2002), in which r^2 = 0.692, and which is a subjective approach. The most likely explanation for this observation is that there was a non-significant declining trend in autosome gene reads with age at sampling (r^2 = 0.15; P value = 0.052), as has previously been mentioned in Chapter Four, that was used to calculate the relative telomere length (T/U) and was probably related to the observed non-significant increase in finished DNA library size and age (r^2 = 0.06; P value = 0.23). This

means that the relative number of telomere reads among older people are overestimated. The data were corrected for the finished DNA library size and the linear regression then showed a marginal improvement ($r^2 = 0.33$; P value = 0.002).

Remodelling using R was then conducted, and all models generated using different parameters including telomere count, relative telomeres count, autosome reads, and microsatellite reads. The modelling showed most of the value added was related to the uncorrected telomere count ($r^2 = 0.53$; P value <0.0001).

The bait $(TTAGGG)_{20}$ was designed to capture pure TTAGGG repeats. However, the data showed three telomeric repeat motifs. A bioinformatics count which was conducted revealed that TTAGGG repeats were the most dominant ones and occupied 92% of the total telomere repeats captured. Both TTGGGG and TGAGGG were also present and represented 5% and 3% respectively, and were located within TTAGGG repeats tracts. Illumina pair reads estimated <0.99% error rates. We used six custom references to distinguish the true telomeric variants from variants which could be generated by sequencing errors. The results suggested that where G>C, TTAGGC and TTACGG patterns were generated by a sequencing error. However, the results revealed a new pattern (TAAGGG) which was located in telomere regions, which had previously been identified as rarely being present in human telomere (Baird *et al*., 1995).

TTGGGG and TGAGGG were found to be located at the start of the telomeric region at the chromosome ends and were not expected to change due to age (Allshire *et al*., 1989). Therefore, a linear regression analysis was carried out to investigate the age correlation with each motif. It was decided to use the repeats count rather than the reads count because the reads count showed an overlap between motifs. The majority of telomere reads contain TTAGGG, at a level of 99%. Only 14% of the telomere reads contain TTGGGG repeats and 7% include TGAGGG repeats. TTAGGG repeats showed a highly significant correlation with age at sampling (P value =< 0.0001 and the $r^2 = 0.53$). Our data also showed that TTGGGG repeats and TGAGGG repeats showed significant correlations with age at sampling (P value = 0.0002 and the $r^2 = 0.44$). This may suggest that these repeats are scattered in the telomere region but are fewer in number. Allshire *et al*., 1989, used synthesised labelled oligonucleotides and a Southern blot approach to identify these variants. As TTAGGG is the most dominant pattern, cross-hybridisation may occur, which might mask other patterns that may exist (such as TTGGGG repeats and TGAGGG) and they could be underestimated as a consequence. Restriction enzymes that cut GAGG and GGTGA were used to determine the position of each pattern, which may be affected by the

presence of any variant at restriction sites. Only by sequencing the entire human telomere (10-15 kb) can the patterns and arrangements of variant repeats be established.

Moreover, the paired samples were analysed to estimate the amount of telomere shortening per year. The results showed different rates of telomere shortening. The population average estimate for telomere shortening was found to be 60 bp/year. The age-dependent telomere shortening value calculated in the experiment was consistent with prior studies conducted by Leteurtre *et al*. (1999) and Iwama *et al*. (1998). Leteurtre *et al*. measured the telomere length in blood collected from 42 healthy individuals aged between 1 and 55 years, and their results showed that the rate of telomere shortening was 60 bp/year. Iwama *et al*. found the rate of telomere shortening to be 67 bp/year in 124 healthy individuals aged between 4 and 95 years.

It was hoped that by using NGS to target only the telomeric region that does not include the sub-telomeric region as in the TRF method (Cawthon, 2002), an improved correlation may be observed between telomere length and age. The unique DNA sequences showed a declining trend with age at sampling, which might contribute to the results observed. In addition, other factors exist which influence telomere length, such as sex and heritability. Moreover, variation is present in the activity of telomerase and other telomere proteins in germ line cells of the parents of individuals. Oxidative damage could also affect telomere length. Variations in lifestyles such as smoking and drinking could affect oxidative damage. The level of anti-oxidants could also be affected by dietary differences, and there could be differences in the rates of repair of oxidative damage.

So the measurement of telomere length may not, by itself, be appropriate to predict the age of a human. Longitudinal studies of telomere lengths including factors that may affect them which are carried out by observing individuals at different life stages may help to build our understanding of what controls telomere length and to clarify the relationship between telomere lengths and age. Likewise, the mode of inheritance of telomere lengths can be investigated in multi-generational studies in matched populations for ethnicity, lifestyle, and sex. Genes also control 25-30% of longevity variations (Deelen *et al*., 2011; Deelen *et al*., 2013). Five genes (*TERC, TERT, NAF1, OBFC1 and RTEL1*) are recognised to be associated with telomere biology (Codd *et al*, 2013). Inter-individual variation in leucocyte telomere length has recently found to be associated with genetic variants (Codd *et al*., 2013). It would therefore be useful to study the genetic variants that contribute to telomere length variation, in order to determine individual genotypes, as telomere shortening is influenced by an individual's genotype (Codd *et al*., 2013).

6 Discussion and conclusion

6.1 Discussion

The first aim of this study was to identify microsatellites with a relatively high somatic mutation frequency in the general population. TRDB revealed 23 pure microsatellites with pattern size ≥ 2 and ≤ 10 bp and repeat copy number ≥50. These microsatellites showed copy number variation among the samples tested. A bulk PCR was conducted to identify individuals (age >40 years) with allele length > 50 repeats. Out of 23 loci, expanded alleles (≥50) were not detected at four loci (19-AT-431, 7-AT-232, 16-TTTC-505, and 1-CTCCCT-151). The failure to observe expanded alleles at these loci might because they are rare in the population. Alternatively the expansion detected in the reference genome may be an artefact due to sequencing or cloning errors. An increase in the sample size may further demonstrate the possible expansion of these loci, or the samples used for the construction of the human reference genomes could be reanalysed. Three loci (10-CA-994, 20-CA-417 and 13-AAG-102) showed allele lengths > 100 repeats. SP-PCR of microsatellites with a repeat copy number ≥50 showed only a low level of somatic instability in the general population. The present work was conducted at the time when the NGS read length was 150 bp which was an obstacle to investigate the 23 expanded microsatellites. Now, alternative platforms of high throughput sequencing could be used. For example, the Ion PGM system can achieve 400 bp read lengths (Loman *et al.,* 2012), Illumina MiSeq up to 600 bp (Juneman *et al.,* 2013), and PacBio provides long reads with lengths in excess of 20 kb (http://files.pacb.com/pdf/PacBio_RS_II_Brochure.pdf). Some (not all) of the expanded microsatellites (di, tri, tetra, penta) identified could possibly be examined using the Ion PGM system. However, as multiple PCRs are the prerequisite of the Ion PGM system, this may introduce a lot of errors due to PCR slippage. Owing to the high read length of the PacBio system and limited use of PCR, all the microsatellites examined in this study could be theoretically evaluated using this platform. However, the platform is very expensive and only a limited number of machines are available. Additionally, absolute output of this platform is relatively low (75,000 reads), therefore limiting the number of samples, and/or loci that could be included at high read coverage. The recently developed Illumina MiSeq platform provides good advantages in terms of read length and depth, thus could be potentially used effectively for analysing the expanded microsatellites. Unpublished data from our research group show that MiSeq is more reliable than Ion PGM system when sequencing expanded loci. The Ion PGM system is unable to sequence up to 20 CTG compared to at least 70 CTG with MiSeq and at least

600 CTG with PacBio. SP-PCR using traditional technologies was found to lack the resolution power to delineate the variability of the examined expanded microsatellite due to the relatively low somatic mutation frequency of the loci. Alternatively, it is possible that the somatic mutation frequency at these loci is high, but that the changes are mostly very small and not detectable on agarose gels. The mentioned NGS platforms can distinguish single repeat length difference and might provide a more sensitive platform for investigating these loci.

We proved that by searching the genome for large expanded repeats revealed their existence and they were found to be expanded in some individuals but not in all. This suggests there are other expanded microsatellites in the human genome as the expanded loci in the human reference (HG19 and Venter) investigated did not overlap. To identify these (unexplored) expanded microsatellites the stringency of the selection criteria could be lowered (*e.g.* copy number > 30). Based on the mentioned criteria, 563 pure microsatellites with copy number >30 were identified in HG19 whereas 482 pure microsatellites with similar range of copy number were found in Venter genome reference. As the total numbers of microsatellites observed are much higher under these relaxed criteria, examination of all those microsatellites sequentially through bulk PCR may be impractical. Alternatively, Ion AmpliSeq custom panels could be used to investigate allele length variation. Ion AmpliSeq is based on multiplex PCR amplification. The total primer pool size is up to 3,072 primer pairs per pool (lifetechnologies.com/ampliseqcustom).

The newly developed 1000 Genomes database was not used in this study, as the read length used to construct the 1000 Genome database is limited to ~150 bp (The 1000 Genomes Project Consortium, 2010), which is not suitable for examining expanded microsatellites.

The human genome reference sequence was used to design the flanking primers but not all microsatellites were easily amplified. Our data highlight the structural complexity of the human genome and revealed the presence of structural variants in 13-AAG-102, 2-CA-181 and 21-TCCT-409. We need to be aware of such complexity that may exist in flanking regions such as deletion, insertion and, duplication and handling these is still problematic. Such additional changes in the flanking region of microsatellites complicate alignment with the reference genomes. Standard alignment pipelines are often unable to align these changes appropriately as evident by the prediction of multiple deletions spanning both microsatellites and their flanking DNA. In this study no deletions were identified of the same magnitude as predicted in the database. Only three microsatellites with structural

variants were examined in this study which was relatively easy, but very time consuming. It would be of interest to explore many microsatellites across the genome in similar depth. However, such a task would be extensive and demand considerable amounts of experimental time and computational resources.

SP-PCR as an approach is laborious, time consuming and involves usage of radioactive isotopes. One of the project's aims was to investigate the potential of new technologies to achieve the same goals in a rapid manner using high throughput analysis. A target enrichment sequencing approach was selected to capture only the sequences in question. Custom bait (probes) design is a necessary step where the baits are complementary to the desired unique DNA sequences. The Illumina GA IIx platform was used for Illumina Paired-End Sequencing approach. Baits were designed using the Agilent SureSelect^{XT} Target Enrichment System. Several approaches were used to design the baits depending on the targeted regions. The custom baits were used to generate a finished DNA library that included microsatellites, DNA repair genes, telomere sequences, and others.

Microsatellites with only pure (100% match) mono, di, and trinucleotides with unique flanking sequence and copy number ≥5 were investigated. Using TRF and RepeatMasker®, 25,539 microsatellites with unique 60 bp flanking sequence from both sides were selected. The novel approach was conducted to capture microsatellite sequences using baits only consisting of unique flanking sequences. Repetitive microsatellite patterns were removed, except if they were present in the unique flanking sequence. The same tools using different bait designing approach were used to design baits at the same size (120 bp) for telomeric region, DNA repair genes, the *DMPK* region and SNPs.

Sequencing of these different genomic regions from the 26 DM1 samples using extensive coverage of baits generated millions of reads. The subsequent analyses showed a significant correlation with age at sampling ($r^2 = 0.19$ and the P value = 0.03), irrespective to the time points of the data. This is due to the presence of telomere reads among the total read count which are known to shorten with age.

Microsatellites were successfully captured using baits consisting of unique flanking sequences. Microsatellite mapping errors were reduced due to the existence of a unique flanking sequence on both sides. The captured microsatellites were 4,491 mono-, 18,924 di-, and 1,234 trinucleotides. The alignment process provided an average coverage driven by the allele length due to the inverse correlation between allele length and reads spanning the full microsatellite and a part of the flanking region. Therefore, shorter loci have more

spanning reads compared to larger loci. The full spanning reads were informative and could be used to produce reliable genotypes.

Although the read coverage could be improved by using microsatellites with short allele length, these loci are typically less variable and will be less informative. Alternatively, more informative microsatellites with longer alleles could be analysed using sequencing platforms (*e.g.* MiSeq) that generate longer reads.

Based on earlier genome-wide SNP genotyping studies, 40-50X coverage is adequate for genotyping (Lomas, 2013). However, my preliminary analysis of 12 SNPs (randomly chosen), suggests that coverage between 35-40X can be used reliably for genotyping. However, microsatellites need higher read coverage > 50X as they are prone to PCR slippage and sequencing errors that may effect their allele distribution. For ageing studies where low mutation frequencies are expected, especially in shorter microsatellites, much higher coverage may be required.

Due to the possibility of PCR slippage, genotyping using microsatellites could be questionable as it may produce artefacts. For example mononucleotides showed a high rate of PCR slippage and failed to be genotyped and should not be included in future analysis. Dinucleotides also showed PCR slippage, but we were able to generate reliable genotypes. However, these may be also not the best candidates to study mutation frequencies. Trinucleotides showed a lower rate of PCR slippage and produced reliable genotypes. Tetra, penta and hexanucleotide were not included in this study because of the technological limitations, but would be expected to be even less susceptible to PCR slippage errors. However with the innovation in the high throughput sequencing methodologies it may be now possible to conduct mutation frequency analysis using these microsatellites. One such method developed is “Safe Sequencing System” by tagging the template molecule with a unique sequence identifier prior to amplification and sequencing (Kinde et al., 2011). Briefly, a unique identifier (UID) is an artificially synthesised nucleotide sequence ranging between 12 or 14 random bp and incorporated between DNA specific primer sequence and universal primer at the 5′ end of the forward amplicon specific primer. These primers yield up to 268 million different UIDs to ensure that the number of UIDs exceeds the number of original DNA template molecules and two different original DNA template do not link to the same UID (Kinde et al., 2011). The first amplification links the unique identifier to each strand of original DNA template using two PCR cycles. Then the generated DNA template will be copied repeatedly in the subsequent amplification using twenty-five PCR cycles. This generates different families of UID

sequences that can be resolved following sequencing of these templates. Prior to second round of PCR the unincorporated UID-containing primers must be removed by exonuclease. Universal primers are tagged with phosphoramidite to protect against exonuclease activity. NGS reads with high quality are grouped into different UID families. A true mutation is distinguished from the PCR errors by the frequency of the variation observed in the sequence of amplicons. True variants were found to be consistently present in ≥95% of the sequences/template molecule and referred to as "Super mutant". By contrast artefacts do not show that high frequency of sequence variation (Kinde et al., 2011).

More recently, another method has been developed referred to as "Duplex sequencing" where two different random adaptors are attached to both ends of the dsDNA template prior to the PCR. The adaptor is composed of two strands of DNA, to one of which random nucleotide sequence is attached while opposite strand with A overhang was created with DNA polymerase. The overhang facilitates the attachment of the adaptor to the T-tailed DNA fragments. Subsequent amplification was conducted using Illumina flow cell compatible primers to generate families of amplicons. Then sequencing of amplicons was conducted to classify them into two categories derived from each of the two single parental strands based on the orientation of the adaptors. Finally, the frequency of the variety differentiates true mutations from PCR errors. If the variation is present in both groups of PCR amplicons, it is considered as a true mutation, whereas PCR error at the first round of amplification is present in only one class of amplicons, while other errors during PCR or sequencing are randomly distributed (Kennedy et al., 2013; Schmitt et al., 2012). Duplex sequencing is more accurate as both strands of the duplex DNA are tagged with random complementary double-strand nucleotide sequence where in "Safe Sequencing System" only single DNA strands were tagged with UIDs from one end.

In our experiment, SureSelect sequence capture was used which is an expensive approach. In future experiments, panels like AmpliSeq could be used which are more cost effective. Secondly, to gain more coverage and accuracy, only those microsatellites would be used which have shown variation among the tested population.

By capturing the DNA repair genes using target enrichment approach we were able to predict the sex of the sample tested, as the number of gene reads mapped to chromosome X is double in females (the dosage effect). The linear regression analysis showed no significant effect on either the total DNA repair gene or on the autosome gene read counts with age at sampling. However, a declining trend was observed in relation to increase age

in both reads. They were used to calculate the relative telomere length. Nonetheless, these genes can be included as control genes and SNPs located in these genes can be investigated in future to test for their association with somatic instability with the age. This will distinguish between age-dependent somatic instability and somatic instability consequent of genetic variation in these genes.

In addition, thousands of human SNPs were successfully captured using the designed baits, confirming an equal capture for both alleles that could exist at a single SNP. SNPs associated with telomere length (*trans*-acting modifier) may be included in future experiments. Telomere shortening is influenced by an individual's genotype (Codd *et al*., 2013). Therefore, it would be useful to study the genetic variants that contribute to telomere length variation, in order to determine individual genotypes.

Target enrichment of the fragmented telomeric regions and sequencing is in line with earlier studies demonstrating age-dependent telomere shortening. It would be of great interest if the intact telomeric region could be captured. In this regard a method developed by Baird et al (2003) referred to as single telomere length analysis (STELA) could be used to capture the full-length sequence of telomere. Briefly, telomere sequences were amplified using the consensus CCCTAA sequence. The consensus sequence is followed by non-complementary 20 nucleotides (linker) to the G-rich 3′ over hang. This linker will ligate to complementary sequence at the 5′end. Subsequently PCR can be conducted using an upstream chromosome-specific primer located in the subtelomeric region and a primer identical to the 20 nucleotides in the linker. Then, the total length of the telomere is estimated using SP-PCR. The STELA methodology has originally been designed for the sex chromosomes (X or Y). However, given the availability of the human genome sequences, the sequence of chromosomal specific region that is adjacent to telomere (for the autosomes) has been explored with reasonable accuracy, it is possible to implement the same methodology to other chromosomes. Given the potential long telomeric repeat lengths, the PacBio platform or Nanopore could be used for examining telomere length. In addition, the variation in the data within the population could possibly be minimised by analysing multiple chromosomes instead of one chromosome end.

In the present study DNA repair genes were used to normalise the telomere read count, however the comparative ratio is very low (1/7667) and hundreds of thousands of reads are required to generate a generate a good measure of telomere length. Alternatively, the Cawthon Q-PCR primers could be used to amplify the telomere region which then could be sequenced and counted. However, it is a challenge to find a control sequence that could

equate the telomere region length. If the previously unique DNA sequences used, we expect to observe the same ratio (1/7667) and estimation of age-dependent telomere shortening would be impractical. One such target is Alu repeats, comprising ~10% of the human genome (Batzer and Deininger, 2002). However, these sequences are relatively more common (300 Mb) than telomere region (300/1) thereby could not be used to normalise telomere length. A 1/1 ratio would more appropriate and a set of primers with some mismatches could be designed that amplify Alu repeats less efficiently which then could reliably be used to calculate relative telomere length. Alternatively, a family of Alu repeats (~ 1 Mb) could be identified that could equate the telomere repeats length or another type of dispersed repeat with a relatively high frequency in the genome.

6.2 Conclusion

The primary aim of this study to estimate human age by investigating the level of somatic instability of microsatellites was not achieved. Nevertheless, the expanded microsatellites showed a low level of somatic instability in general population; if the data set were expanded it is possible that some degree of association could be established between age and somatic instability. Application of a high throughput approach was found to be successful for genotyping; it provides insights into capturing microsatellites using unique flanking sequences. The relative telomere lengths data analysis showed that our novel approach was capable to capture telomeric region and generate significant correlation with age. With the advent of new sequencing platforms and bioinformatics it is reasonable to assume that in near future the caveats present in this study will be overcome and new insights in relation to the aims will be gained.

Bibliography

Abuin A, Zhang H and Bradley A. (2000) Genetic analysis of mouse embryonic stem cells bearing *Msh3* and *Msh2* single and compound mutations. *Molecular and Cellular Biology* **20**: 149-157.

Alexeyev MF. (2009) Is there more to aging than mitochondrial DNA and reactive oxygen species? *Febs Journal* **276**: 5768-5787.

Alkass K, Buchholz BA, Ohtani S, Yamamoto T, Druid H and Spalding KL. (2010) Age Estimation in Forensic Sciences Application of combined aspartic acid racemization and radiocarbon analysis. *Molecular and Cellular Proteomics* **9**: 1022-1030.

Allen HL, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S and Raychaudhuri S. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**: 832-838.

Allshire RC, Dempster M and Hastie ND. (1989) Human telomeres contain at least three types of G–rich repeat distributed non-randomly. *Nucleic Acids Research* **17**: 4611-4627.

Allsopp RC and Harley CB. (1995) Evidence for a critical telomere length in senescent human fibroblasts. *Experimental Cell Research* **219**: 130-136.

Allsopp RC, Vaziri H, Patterson C, Goldstein S, Younglai EV, Futcher AB, Greider CW and Harley CB. (1992) Telomere length predicts replicative capacity of human fibroblasts. *Proceedings of the National Academy of Sciences,USA* **89**: 10114-10118.

Alvarez M and Ballantyne J. (2006) The identification of newborns using messenger RNA profiling analysis. *Analytical Biochemistry* **357**: 21-34.

Anvret M, Ahlberg G, Grandell U, Hedberg B, Johnson K and Edstrōm L. (1993) Larger expansions of the CTG repeat in muscle compared to lymphocytes from patients with myotonic dystrophy. *Human Molecular Genetics* **2**: 1397-1400.

Apfeld J, O'Connor G, McDonagh T, DiStefano PS and Curtis R. (2004) The AMP-activated protein kinase AAK-2 links energy levels and insulin-like signals to lifespan in *C. elegans*. *Genes and Development* **18**: 3004-3009.

Aquilina G and Bignami M. (2001) Mismatch repair in correction of replication errors and processing of DNA damage. *Journal of Cellular Physiology* **187**: 145-154.

Ashizawa T, Dubel JR and Harati Y. (1993) Somatic instability of CTG repeat in myotonic dystrophy. *Neurology* **43**: 2674-2674.

Ashizawa T, Dunne P, Ward P, Seltzer W and Richards C. (1994) Effects of the sex of myotonic dystrophy patients on the unstable triplet repeat in their affected offspring. *Neurology* **44**: 120-120.

Ashley Jr CT and Warren ST. (1995) Trinucleotide repeat expansion and human disease. *Annual Review of Genetics* **29**: 703-728.

Aslanidis C, Jansen G, Amemiya C, Shutler G, Mahadevan M, Tsilfidis C, Chen C, Alleman J, Wormskamp NG and Vooijs M. (1992) Cloning of the essential myotonic dystrophy region and mapping of the putative defect. *Nature* **355**: 548-551.

Aubert G and Lansdorp PM. (2008) Telomeres and aging. *Physiological Reviews* **88**: 557-579.

Baird D, Jeffreys A and Royle N. (1995) Mechanisms underlying telomere repeat turnover, revealed by hypervariable variant repeat distribution patterns in the human Xp/Yp telomere. *The EMBO Journal* **14**: 5433.

Baird DM, Britt-Compton B, Rowson J, Amso NN, Gregory L and Kipling D. (2006) Telomere instability in the male germline. *Human Molecular Genetics* **15**: 45-51.

Baird DM, Rowson J, Wynford-Thomas D and Kipling D. (2003) Extensive allelic variation and ultrashort telomeres in senescent human cells. *Nature Genetics* **33**: 203-207.

Baker SJ, Markowitz S, Fearon ER, Willson J and Vogelstein B. (1990) Suppression of human colorectal carcinoma cell growth by wild-type p53. *Science* **249**: 912-915.

Bartke A. (2005) Minireview: role of the growth hormone/insulin-like growth factor system in mammalian aging. *Endocrinology* **146**: 3718-3723.

Batzer MA and Deininger PL. (2002) Alu repeats and human genomic diversity. *Nature Reviews Genetics* **3**: 370-379.

Beckman KB and Ames BN. (1998) The free radical theory of aging matures. *Physiological Reviews* **78**: 547-581.

Beers MH, Jones TV, Berkwits M, Kaplan J and Porter R. (2004) *The Merck Manual of Health and Aging*: Merck Research Laboratories Whitehouse Station.

Bekaert S, Derradji H and Baatout S. (2004) Telomere biology in mammalian germ cells and during development. *Developmental Biology* **274**: 15-30.

Bellacosa A. (2001) Functional interactions and signaling properties of mammalian DNA mismatch repair proteins. *Cell Death and Differentiation* **8**: 1076-1092.

Benetos A, Okuda K, Lajemi M, Kimura M, Thomas F, Skurnick J, Labat C, Bean K and Aviv A. (2001) Telomere length as an indicator of biological aging the gender effect and relation with pulse pressure and pulse wave velocity. *Hypertension* **37**: 381-385.

Benson G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**: 573.

Bidichandani SI, Ashizawa T and Patel PI. (1998) The GAA triplet-repeat expansion in Friedreich ataxia interferes with transcription and may be associated with an unusual DNA structure. *The American Journal of Human Genetics* **62**: 111-121.

Bland JM and Altman DG. (1995) Multiple significance tests: the Bonferroni method. *BMJ: British Medical Journal* **310**: 170.

Boby T, Patch A-M and Aves S. (2005) TRbase: a database relating tandem repeats to disease genes for the human genome. *Bioinformatics* **21**: 811-816.

Boukamp P. (2001) Ageing mechanisms: the role of telomere loss. *Clinical and Experimental Dermatology* **26**: 562-565.

Braida C, Stefanatos RK, Adam B, Mahajan N, Smeets HJ, Niel F, Goizet C, Arveiler B, Koenig M and Lagier-Tourenne C. (2010) Variant CCG and GGC repeats within the CTG expansion dramatically modify mutational dynamics and likely contribute toward unusual symptoms in some myotonic dystrophy type 1 patients. *Human Molecular Genetics* **19**: 1399-1412.

Brierley EJ, Johnson MA, James OF and Turnbull DM. (1997) Mitochondrial involvement in the ageing process. Facts and controversies. *Molecular and Cellular Biochemistry* **174**: 325-328.

Brinkmann B, Klintschar M, Neuhuber F, Hühne J and Rolf B. (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *The American Journal of Human Genetics* **62**: 1408-1415.

Brinkmann B, Sajantila A, Goedde H, Matsumoto H, Nishi K and Wiegand P. (1996) Population genetic comparisons among eight populations using allele frequency and sequence data from three microsatellite loci. *European Journal of Human Genetics: EJHG* **4**: 175.

Brock GJ, Anderson NH and Monckton DG. (1999) *Cis*-acting modifiers of expanded CAG/CTG triplet repeat expandability: associations with flanking GC content and proximity to CpG islands. *Human Molecular Genetics* **8**: 1061-1067.

Brohede J and Ellegren H. (1999) Microsatellite evolution: polarity of substitutions within repeats and neutrality of flanking sequences. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **266**: 825-833.

Brook JD, McCurrach ME, Harley HG, Buckler AJ, Church D, Aburatani H, Hunter K, Stanton VP, Thirion J-P and Hudson T. (1992) Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. *Cell* **68**: 799-808.

Brunner H, Brüggenwirth H, Nillesen W, Jansen G, Hamel B, Hoppe R, de Die C, Höweler C, Van Oost B and Wieringa B. (1993) Influence of sex of the transmitting parent as well as of parental allele site on the CTG expansion in myotonic dystrophy (DM). *American Journal of Human Genetics* **53**: 1016.

Budowle B, Moretti TR, Baumstark AL, Defenbaugh DA and Keys KM. (1999) Population Data on the Thirteen CODIS Core Short Tandem Repeat Loci in African-Americans, US Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians. *Journal of Forensic Sciences* **44**: 1277-1286.

Butler JM. (2005) *Forensic DNA typing: biology, technology, and genetics of STR markers*: Academic Press.

Butler JM. (2006) Genetics and genomics of core short tandem repeat loci used in human identity testing. *Journal of Forensic Sciences* **51**: 253-265.

Butler JM and Levin BC. (1998) Forensic applications of mitochondrial DNA. *Trends in Biotechnology* **16**: 158-162.

Buxton J, Shelbourne P, Davies J, Jones C, Van Tongeren T, Aslanidis C, de Jong P, Jansen G, Anvret M and Riley B. (1992) Detection of an unstable fragment of DNA specific to individuals with myotonic dystrophy. *Nature* **355**: 547-548.

Callahan JL, Andrews KJ, Zakian VA and Freudenreich CH. (2003) Mutations in yeast replication proteins that increase CAG/CTG expansions also increase repeat fragility. *Molecular and Cellular Biology* **23**: 7849-7860.

Carneiro M, Russ C, Ross M, Gabriel S, Nusbaum C and DePristo M. (2012) Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC genomics* **13**: 375.

Cawthon RM. (2002) Telomere measurement by quantitative PCR. *Nucleic Acids Research* **30**: e47-e47.

Cawthon RM, Smith KR, O'Brien E, Sivatchenko A and Kerber RA. (2003) Association between telomere length in blood and mortality in people aged 60 years or older. *The Lancet* **361**: 393-395.

Chang E and Harley CB. (1995) Telomere length and replicative aging in human vascular tissues. *Proceedings of the National Academy of Sciences* **92**: 11190-11194.

Chen X, Santhana Mariappan S, Moyzis RK, Bradbury EM and Gupta G. (1998) Hairpin induced slippage and hyper-methylation of the fragile X DNA triplets. *Journal of Biomolecular Structure and Dynamics* **15**: 745-756.

Clarke LA, Rebelo CS, Goncalves J, Boavida MG and Jordan P. (2001) PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. *Molecular Pathology* **54**: 351-353.

Cleary J and Pearson C. (2003) The contribution of *cis*-elements to disease-associated repeat instability: clinical and experimental evidence. *Cytogenetic and Genome Research* **100**: 25-55.

Codd V, Nelson CP, Albrecht E, Mangino M, Deelen J, Buxton JL, Hottenga JJ, Fischer K, Esko T and Surakka I. (2013) Identification of seven loci affecting mean telomere length and their association with disease. *Nature Genetics* **45**: 422-427.

Collins JR, Stephens RM, Gold B, Long B, Dean M and Burt SK. (2003) An exhaustive DNA micro-satellite map of the human genome using high performance computing. *Genomics* **82**: 10-19.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C and Campbell P. (2009) Origins and functional impact of copy number variation in the human genome. *Nature* **464**: 704-712.

Consortium 1000 GP. (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061-1073.

Coolbaugh-Murphy MI, Xu J, Ramagli LS, Brown BW and Siciliano MJ. (2005) Microsatellite instability (MSI) increases with age in normal somatic cells. *Mechanisms of Ageing and Development* **126**: 1051-1059.

Cooper GM, Zerr T, Kidd JM, Eichler EE and Nickerson DA. (2008) Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nature Genetics* **40**: 1199-1203.

Corte-Real F. (2004) Forensic DNA databases. *Forensic Science International* **146**: S143-S144.

Counter CM, Avilion AA, LeFeuvre CE, Stewart NG, Greider CW, Harley CB and Bacchetti S. (1992) Telomere shortening associated with chromosome instability is arrested in immortal cells which express telomerase activity. *The EMBO Journal* **11**: 1921.

Cutler RG. (1984) Urate and ascorbate: their possible roles as antioxidants in determining longevity of mammalian species. *Archives of Gerontology and Geriatrics* **3**: 321-348.

Darlow JM and Leach DR. (1998) Secondary structures in d (CGG) and d (CCG) repeat tracts. *Journal of Molecular Biology* **275**: 3-16.

Database, N C I D 2013, NCIDD - National Criminal Investigation DNA Database, 22/04/2013, (http://www.crimtrac.gov.au/systems_projects/NationalCriminal Invest gationNADatabaseNCIDD.html).

De Meyer T, Rietzschel ER, De Buyzere ML, De Bacquer D, Van Criekinge W, De Backer GG, Gillebert TC, Van Oostveldt P and Bekaert S. (2007) Paternal age at birth is an important determinant of offspring telomere length. *Human Molecular Genetics* **16**: 3097-3102.

de Wind N, Dekker M, Berns A, Radman M and te Riele H. (1995) Inactivation of the mouse *Msh*2 gene results in mismatch repair deficiency, methylation tolerance, hyperrecombination, and predisposition to cancer. *Cell* **82**: 321-330.

Dean RG, Socher SH and Cutler RG. (1985) Dysdifferentiative nature of aging: age-dependent expression of mouse mammary tumor virus and casein genes in brain and liver tissues of the C57BL/6J mouse strain. *Archives of Gerontology and Geriatrics* **4**: 43-51.

Deelen J, Beekman M, Uh HW, Helmer Q, Kuningas M, Christiansen L, Kremer D, van der Breggen R, Suchiman HED and Lakenberg N. (2011) Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited. *Aging Cell* **10**: 686-698.

Deelen J, Uh H-W, Monajemi R, van Heemst D, Thijssen PE, Böhringer S, van den Akker EB, de Craen AJM, Rivadeneira F and Uitterlinden AG. (2013) Gene set analysis of GWAS data for human longevity highlights the relevance of the insulin/IGF-1 signaling and telomere maintenance pathways. *Age* **35**: 235-249.

Dohm JC, Lottaz C, Borodina T and Himmelbauer H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research* **36**: e105-e105.

Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden DW, Smith DR and Lander ES. (1987) A genetic linkage map of the human genome. *Cell* **51**: 319-337.

Drenos F and Kirkwood TB. (2005) Modelling the disposable soma theory of ageing. *Mechanisms of Ageing and Development* **126**: 99-103.

Dumont P, Burton M, Chen QM, Gonos ES, Frippiat C, Mazarati J-B, Eliaers F, Remacle J and Toussaint O. (2000) Induction of replicative senescence biomarkers by sublethal oxidative stresses in normal human fibroblast. *Free Radical Biology and Medicine* **28**: 361-373.

Edwards A, Civitello A, Hammond HA and Caskey CT. (1991) DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *American Journal of Human Genetics* **49**: 746.

Ellegren H. (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* **5**: 435-445.

Feser J and Tyler J. (2011) Chromatin structure as a mediator of aging. *FEBS Letters* **585**: 2041-2048.

Fleisher B. (1918) Über myotonische Dystrophie mit Katarakt: Eine hereditäre, familiare Degeneration. *Arch Opthalmol* **96**: 91-133.

Flurkey K, Papaconstantinou J, Miller RA and Harrison DE. (2001) Lifespan extension and delayed immune and collagen aging in mutant mice with defects in growth hormone production. *Proceedings of the National Academy of Sciences* **98**: 6736-6741.

Foiry L, Dong L, Savouret C, Hubert L, te Riele H, Junien C and Gourdon G. (2006) *Msh3* is a limiting factor in the formation of intergenerational CTG expansions in DM1 transgenic mice. *Human Genetics* **119**: 520-526.

Fortune MT, Vassilopoulos C, Coolbaugh MI, Siciliano MJ and Monckton DG. (2000) Dramatic, expansion-biased, age-dependent, tissue-specific somatic mosaicism in a transgenic mouse model of triplet repeat instability. *Human Molecular Genetics* **9**: 439-445.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P and Leal SM. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851-861.

Freudenreich CH, Kantrow SM and Zakian VA. (1998) Expansion and length-dependent fragility of CTG repeats in yeast. *Science* **279**: 853-856.

Freudenreich CH, Stavenhagen JB and Zakian VA. (1997) Stability of a CTG/CAG trinucleotide repeat in yeast is dependent on its orientation in the genome. *Molecular and Cellular Biology* **17**: 2090-2098.

Friedman AL. (1999) Forensic DNA profiling in the 21st century. *International Journal of Offender Therapy and Comparative Criminology* **43**: 168-179.

Frudakis T. (2010) *Molecular photofitting: predicting ancestry and phenotype using DNA*: Access Online via Elsevier.

Fu Y-H, Friedman DL, Richards S, Pearlman JA, Gibbs RA, Pizzuti A, Ashizawa T, Perryman MB, Scarlato G and Fenwick R. (1993) Decreased expression of myotonin-protein kinase messenger RNA and protein in adult form of myotonic dystrophy. *Science* **260**: 235-238.

Fu Y-H, Kuhl D, Pizzuti A, Pieretti M, Sutcliffe JS, Richards S, Verkert AJ, Holden JJ, Fenwick Jr RG and Warren ST. (1991) Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell* **67**: 1047-1058.

Gacy AM and McMurray CT. (1998) Influence of hairpins on template reannealing at trinucleotide repeat duplexes: a model for slipped DNA. *Biochemistry* **37**: 9426-9434.

Galindo C, McIver L, McCormick J, Skinner M, Xie Y, Gelhausen R, Ng K, Kumar N and Garner H. (2009) Global microsatellite content distinguishes humans, primates, animals, and plants. *Molecular Biology and Evolution* **26**: 2809-2819.

Garamendi P, Landa M, Ballesteros J and Solano M. (2005) Reliability of the methods applied to assess age minority in living subjects around 18 years old: A survey on a Moroccan origin population. *Forensic Science International* **154**: 3-12.

Geiersbach KB and Samowitz WS. (2011) Microsatellite instability and colorectal cancer. *Archives of Pathology and Laboratory Medicine* **135**: 1269-1277.

Gelfand Y, Rodriguez A and Benson G. (2007) TRDB—the tandem repeats database. *Nucleic Acids Research* **35**: D80-D87.

Giannini G, Rinaldi C, Ristori E, Ambrosini MI, Cerignoli F, Viel A, Bidoli E, Berni S, D'Amati G and Scambia G. (2004) Mutations of an intronic repeat induce impaired *MRE11* expression in primary human cancer with microsatellite instability. *Oncogene* **23**: 2640-2647.

Giles RE, Blanc H, Cann HM and Wallace DC. (1980) Maternal inheritance of human mitochondrial DNA. *Proceedings of the National Academy of Sciences* **77**: 6715-6719.

Gimenez LE, Ghildyal P, Fischer KE, Hu H, Ja WW, Eaton BA, Wu Y, Austad SN and Ranjan R. (2013) Modulation of methuselah expression targeted to *Drosophila* insulin producing cells extends life and enhances oxidative stress resistance. *Aging Cell* **12**: 121-129.

Gomes-Pereira M, Foiry L, Nicole A, Huguet A, Junien C, Munnich A and Gourdon G. (2007) CTG trinucleotide repeat "big jumps": large expansions, small mice. *PLoS Genetics* **3**: e52.

Gomes-Pereira M, Fortune MT, Ingram L, McAbney JP and Monckton DG. (2004) Pms2 is a genetic enhancer of trinucleotide CAG· CTG repeat somatic mosaicism: implications for the mechanism of triplet repeat expansion. *Human Molecular Genetics* **13**: 1815-1825.

Gomes-Pereira M, Fortune MT and Monckton DG. (2001) Mouse tissue culture models of unstable triplet repeats: in vitro selection for larger alleles, mutational expansion bias and tissue specificity, but no association with cell division rates. *Human Molecular Genetics* **10**: 845-854.

Gomes-Pereira M and Monckton DG. (2006) Chemical modifiers of unstable expanded simple sequence repeats: what goes up, could come down. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **598**: 15-34.

Goodwin W, Linacre A and Hadi S. (2011) *An Introduction to Forensic Genetics*: Wiley. com.

Graakjaer J, Pascoe L, Der Sarkissian H, Thomas G, Kolvraa S, Christensen K and LondoñoVallejo JA. (2004) The relative lengths of individual telomeres are defined in the zygote and strictly maintained during life. *Aging Cell* **3**: 97-102.

Graham EA. (2008) DNA reviews: predicting phenotype. *Forensic Science, Medicine, and Pathology* **4**: 196-199.

Greaves LC and Turnbull DM. (2009) Mitochondrial DNA mutations and ageing. *Biochimica et Biophysica Acta (BBA)-General Subjects* **1790**: 1015-1020.

Greider CW. (1999) Telomeres do D-loop–T-loop. *Cell* **97**: 419-422.

Griffin R, Chamberlain A, Hotz G, Penkman K and Collins M. (2009) Age estimation of archaeological remains using amino acid racemization in dental enamel: A comparison of morphological, biochemical, and known ages at death. *American Journal of Physical Anthropology* **140**: 244-252.

Griffith JD, Comeau L, Rosenfield S, Stansel RM, Bianchi A, Moss H and de Lange T. (1999) Mammalian telomeres end in a large duplex loop. *Cell* **97**: 503-514.

Gulcher J. (2012) Microsatellite markers for linkage and association studies. *Cold Spring Harbor Protocols* **2012**: pdb. top068510.

Gymrek M, Golan D, Rosset S and Erlich Y. (2012) lobSTR: a short tandem repeat profiler for personal genomes. *Genome Research* **22**: 1154-1162.

Hanna CW, Bretherick KL, Gair JL, Fluker MR, Stephenson MD and Robinson WP. (2009) Telomere length and reproductive aging. *Human Reproduction* **24**: 1206-1211.

Harfe BD and Jinks-Robertson S. (2000) DNA mismatch repair and genetic instability. *Annual Review of Genetics* **34**: 359-399.

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ and Levy S. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* **10**: R32.

Harley CB. (1991) Telomere loss: mitotic clock or genetic time bomb? *Mutation Research/DNAging* **256**: 271-282.

Harley CB, Futcher AB and Greider CW. (1990) Telomeres shorten during ageing of human fibroblasts.

Harley HG, Brook JD, Rundle SA, Crow S, Reardon W, Buckler AJ, Harper PS, Housman DE and Shaw DJ. (1992) Expansion of an unstable DNA region and phenotypic variation in myotonic dystrophy. *Nature* **355**: 545-546.

Harley HG, Rundle S, MacMillan J, Myring J, Brook J, Crow S, Reardon W, Fenton I, Shaw D and Harper P. (1993) Size of the unstable CTG repeat sequence in relation to phenotype and parental transmission in myotonic dystrophy. *American Journal of Human Genetics* **52**: 1164.

Harper P. (2001) Myotonic dystrophy.

Hastie ND and Allshire RC. (1989) Human telomeres: fusion and interstitial sites. *Trends in Genetics* **5**: 326-330.

Hastie ND, Dempster M, Dunlop MG, Thompson AM, Green DK and Allshire RC. (1990) Telomere reduction in human colorectal carcinoma and with ageing. *Nature* **346**: 866-868.

Hattori K, Tanaka M, Sugiyama S, Obayashi T, Ito T, Satake T, Hanaki Y, Asai J, Nagano M and Ozawa T. (1991) Age-dependent increase in deleted mitochondrial DNA in the human heart: possible contributory factor to presbycardia. *American Heart Journal* **121**: 1735-1742.

Hegan DC, Narayanan L, Jirik FR, Edelmann W, Liskay RM and Glazer PM. (2006) Differing patterns of genetic instability in mice deficient in the mismatch repair genes *Pms2, Mlh1*, *Msh*2, *Msh*3 and *Msh*6. *Carcinogenesis* **27**: 2402-2408.

Henshaw PS. (1957) Genetic transition as a determinant of physiologic and radiologic aging and other conditions. *Radiology* **69**: 30-36.

Herbst A, Pak JW, McKenzie D, Bua E, Bassiouni M and Aiken JM. (2007) Accumulation of mitochondrial DNA deletion mutations in aged muscle fibers: evidence for a causal role in muscle fiber loss. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* **62**: 235-245.

Higham CF and Monckton DG. (2013) Modelling and inference reveal nonlinear length-dependent suppression of somatic instability for small disease associated alleles in myotonic dystrophy type 1 and Huntington disease. *Journal of The Royal Society Interface* **10**: 20130605.

Highnam G, Franck C, Martin A, Stephens C, Puthige A and Mittelman D. (2013) Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Research* **41**: e32-e32.

Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M and Huang W. (2008) Whole-genome sequencing and variant discovery in *C. elegans*. *Nature Methods* **5**: 183-188.

Holliday R and Tarrant G. (1972) Altered enzymes in ageing human fibroblasts. *Nature* **238**: 26.

Home Office, U K 2013, The national DNA database, 21/11/2013, (http://www.homeoffice.gov.uk).

Höweler CJ, Busch H, Geraedts J, Niermeijer M and Staal A. (1989) Anticipation in myotonic dystrophy: fact or fiction? *Brain* **112**: 779-797.

Iwama H, Ohyashiki K, Ohyashiki JH, Hayashi S, Yahata N, Ando K, Toyama K, Hoshika A, Takasaki M and Mori M. (1998) Telomeric length and telomerase activity vary with age in peripheral blood cells obtained from normal individuals. *Human Genetics* **102**: 397-402.

Jaenisch R and Bird A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics* **33**: 245-254.

Jakupciak JP and Wells RD. (1999) Genetic instabilities in (CTG· CAG) repeats occur by recombination. *Journal of Biological Chemistry* **274**: 23468-23479.

Jakupciak JP and Wells RD. (2000a) Gene conversion (recombination) mediates expansions of CTG· CAG repeats. *Journal of Biological Chemistry* **275**: 40003-40013.

Jakupciak JP and Wells RD. (2000b) Genetic instabilities of triplet repeat sequences by recombination. *IUBMB Life* **50**: 355-359.

Jankowski C, Nasar F and Nag DK. (2000) Meiotic instability of CAG repeat tracts occurs by double-strand break repair in yeast. *Proceedings of the National Academy of Sciences* **97**: 2134-2139.

Jansen G, Willems P, Coerwinkel M, Nillesen W, Smeets H, Vits L, Höweler C, Brunner H and Wieringa B. (1994) Gonosomal mosaicism in myotonic dystrophy patients: involvement of mitotic events in (CTG) n repeat variation and selection against extreme expansion in sperm. *American Journal of Human Genetics* **54**: 575.

Jeffreys AJ, Neumann R and Wilson V. (1990) Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell* **60**: 473-485.

Jeffreys AJ, Wilson V and Thein SL. (1985) Hypervariable'minisatellite' regions in human DNA. *Nature* **314**: 67-73.

Jobling MA and Gill P. (2004) Encoded evidence: DNA in forensic analysis. *Nature Reviews Genetics* **5**: 739-751.

John JCS, Facucho-Oliveira J, Jiang Y, Kelly R and Salah R. (2010) Mitochondrial DNA transmission, replication and inheritance: a journey from the gamete through the embryo and into offspring and embryonic stem cells. *Human Reproduction Update* **16**: 488-509

Jünemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, Kalinowski J, Mellmann A, Goesmann A, von Haeseler A and Stoye J. (2013) Updating benchtop sequencing performance comparison. *Nature Biotechnology* **31**: 294-296.

Kajander OA, Karhunen PJ and Jacobs HT. (2002) The relationship between somatic mtDNA rearrangements, human heart disease and aging. *Human Molecular Genetics* **11**: 317-324.

Kammori M, Nakamura K-I, Kawahara M, Mimura Y, Kaminishi M and Takubo K. (2002) Telomere shortening with aging in human thyroid and parathyroid tissue. *Experimental Gerontology* **37**: 513-521.

Kang S, Jaworski A, Ohshima K and Wells RD. (1995) Expansion and deletion of CTG repeats from human disease genes are determined by the direction of replication in *E. coli*. *Nature Genetics* **10**: 213-218.

Kanungo M and Gandhi BS. (1972) Induction of malate dehydrogenase isoenzymes in livers of young and old rats. *Proceedings of the National Academy of Sciences,USA* **69**: 2035-2038.

Kanungo MS. (2005) *Genes and aging*: Cambridge University Press.

Kapiteijn E, Liefers G, Los L, Klein Kranenbarg E, Hermans J, Tollenaar R, Moriya Y, van de Velde C and van Krieken J. (2001) Mechanisms of oncogenesis in colon versus rectal cancer. *The Journal of Pathology* **195**: 171-178.

Kator K, Cristofalo V, Charpentier R and Cutler R. (1985) Dysdifferentiative nature of aging: passage number dependency of globin gene expression in normal human diploid cells grown in tissue culture. *Gerontology* **31**: 355-361.

Kayser M and Schneider PM. (2009) DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations. *Forensic Science International: Genetics* **3**: 154-161.

Kennedy L, Evans E, Chen C-M, Craven L, Detloff PJ, Ennis M and Shelbourne PF. (2003) Dramatic tissue-specific mutation length increases are an early molecular

event in Huntington disease pathogenesis. *Human Molecular Genetics* **12**: 3359-3367.

Kennedy L and Shelbourne PF. (2000) Dramatic mutation instability in HD mouse striatum: does polyglutamine load contribute to cell-specific vulnerability in Huntington's disease? *Human Molecular Genetics* **9**: 2539-2544.

Kennedy SR, Salk JJ, Schmitt MW and Loeb LA. (2013) Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. *PLoS Genetics* **9**: e1003794.

Kerley ER. (1965) The microscopic determination of age in human bone. *American Journal of Physical Anthropology* **23**: 149-163.

Khrapko K, Bodyak N, Thilly WG, van Orsouw NJ, Zhang X, Coller HA, Perls TT, Upton M, Vijg J and Wei JY. (1999) Cell-by-cell scanning of whole mitochondrial genomes in aged human heart reveals a significant fraction of myocytes with clonally expanded deletions. *Nucleic Acids Research* **27**: 2434-2441.

Kibbe WA. (2007) OligoCalc: an online oligonucleotide properties calculator. *Nucleic acids research* **35**: W43-W46.

Kidd KK, Pakstis A, Speed W and Kidd J. (2004) Understanding human DNA sequence variation. *Journal of Heredity* **95**: 406-420.

Kinde I, Wu J, Papadopoulos N, Kinzler KW and Vogelstein B. (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences* **108**: 9530-9535.

Kirkwood T. (1993) The disposable soma theory: evidence and implications. *Netherlands Journal of Zoology* **43**: 359-363.

Kirkwood TB, Kapahi P and Shanley DP. (2000) Evolution, stress, and longevity. *Journal of Anatomy* **197**: 587-590.

Kirkwood TBL. (2005) Understanding the odd science of aging. *Cell* **120**: 437-447.

Kloosterman A and Kersbergen P. (2003) Efficacy and limits of genotyping low copy number DNA samples by multiplex PCR of STR loci. *International Congress Series.* Elsevier, 795-798.

Knudson AG. (1971) Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences* **68**: 820-823.

Koboldt DC, Ding L, Mardis ER and Wilson RK. (2010) Challenges of sequencing human genomes. *Briefings in Bioinformatics* **11**: 484-498.

Koops B-J and Schellekens M. (2007) Forensic DNA phenotyping: regulatory issues.

Kopsidas G, Kovalenko SA, Kelso JM and Linnane AW. (1998) An age-associated correlation between cellular bioenergy decline and mtDNA rearrangements in human skeletal muscle. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **421**: 27-36.

Kovtun IV and McMurray CT. (2001) Trinucleotide expansion in haploid germ cells by gap repair. *Nature Genetics* **27**: 407-411.

Kremer B, Almqvist E, Theilmann J, Spence N, Telenius H, Goldberg Y and Hayden M. (1995) Sex-dependent mechanisms for expansions and contractions of the CAG repeat on affected Huntington disease chromosomes. *American Journal of Human Genetics* **57**: 343.

Kufe DW, Pollock RE, Weichselbaum RR, Bast RC, Gansler TS, Holland JF and Frei E. (2003) Holland-Frei cancer medicine.

Lahue RS and Slater DL. (2003) DNA repair and trinucleotide repeat instability. *Frontiers in Bioscience: a Journal and Virtual Library* **8**: s653-665.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M and FitzHugh W. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Langmead B and Salzberg SL. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357-359.

Larizza A, Pesole G, Reyes A, Sbisà E and Saccone C. (2002) Lineage specificity of the evolutionary dynamics of the mtDNA D-loop region in rodents. *Journal of Molecular Evolution* **54**: 145-155.

Larsson N and Clayton DA. (1995) Molecular genetic aspects of human mitochondrial disorders. *Annual Review of Genetics* **29**: 151-178.

Lavedan C, Hofmann-Radvanyi H, Shelbourne P, Rabes J, Duros C, Savoy D, Dehaupas I, Luce S, Johnson K and Junien C. (1993) Myotonic dystrophy: size-and sex-dependent dynamics of CTG meiotic instability, and somatic mosaicism. *American Journal of Human Genetics* **52**: 875.

Lawes D, SenGupta S and Boulos P. (2003) The clinical importance and prognostic implications of microsatellite instability in sporadic cancer. *European Journal of Surgical Oncology (EJSO)* **29**: 201-212.

Leteurtre F, Li X, Guardiola P, Le Roux G, Sergère JC, Richard P, Carosella ED and Gluckman E. (1999) Accelerated telomere shortening and telomerase activation in Fanconi's anaemia. *British Journal of Haematology* **105**: 883-893.

Levinson G and Gutman GA. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution* **4**: 203-221.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF and Denisov G. (2007) The diploid genome sequence of an individual human. *PLoS Biology* **5**: e254.

Lewin B. (2004) *Genes Viii*: Pearson Prentice Hall Upper Saddle River.

Lewis R. (1998) Telomere tales. *BioScience* **48**: 981-985.

Li YC, Korol AB, Fahima T, Beiles A and Nevo E. (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology* **11**: 2453-2465.

Lia A-S, Seznec H, Hofmann-Radvanyi H, Radvanyi F, Duros C, Saquet C, Blanche M, Junien C and Gourdon G. (1998) Somatic instability of the CTG repeat in mice transgenic for the myotonic dystrophy region is age dependent but not correlated to the relative intertissue transcription levels and proliferative capacities. *Human Molecular Genetics* **7**: 1285-1291.

Lin J, Epel E and Blackburn E. (2012) Telomeres and lifestyle factors: roles in cellular aging. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **730**: 85-89.

Lindsey J, McGill NI, Lindsey LA, Green DK and Cooke HJ. (1991) In vivo loss of telomeric repeats with age in humans. *Mutation Research/DNAging* **256**: 45-48.

Litt M and Luty JA. (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American Journal of Human Genetics* **44**: 397.

Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D and Darnell J. (2000) Organelle DNAs.

Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J and Pallen MJ. (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology* **30**: 434-439.

Lomas D. (2013) UCL Dean of Medical Science.

Lovejoy CO, Meindl RS, Mensforth RP and Barton TJ. (1985) Multifactorial determination of skeletal age at death: a method and blind tests of its accuracy. *American Journal of Physical Anthropology* **68**: 1-14.

Ma Y-S, Wu S-B, Lee W-Y, Cheng J-S and Wei Y-H. (2009) Response to the increase of oxidative stress and mutation of mitochondrial DNA in aging. *Biochimica et Biophysica Acta (BBA)-General Subjects* **1790**: 1021-1029.

Madisen L, Hoar DI, Holroyd CD, Crisp M, Hodes ME and Reynolds JF. (1987) The effects of storage of blood and isolated DNA on the integrity of DNA. *American Journal of Medical Genetics* **27**: 379-390.

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J and Turner DJ. (2010) Target-enrichment strategies for next-generation sequencing. *Nature Methods* **7**: 111-118.

Manley K, Shirley TL, Flaherty L and Messer A. (1999) *Msh*2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. *Nature Genetics* **23**: 471-473.

Marcadier JL and Pearson CE. (2003) Fidelity of Primate Cell Repair of a Double-strand Break within a (CTG)·(CAG) Tract effect of slipped DNA structures. *Journal of Biological Chemistry* **278**: 33848-33856.

Martorell L, Monckton DG, Gamez J, Johnson KJ, Gich I, de Munain AL and Baiget M. (1998) Progression of somatic CTG repeat length heterogeneity in the blood cells of myotonic dystrophy patients. *Human Molecular Genetics* **7**: 307-312.

McMurray CT. (1999) DNA secondary structure: a common and causative factor for expansion in human disease. *Proceedings of the National Academy of Sciences* **96**: 1823-1825.

McMurray CT. (2010) Mechanisms of trinucleotide repeat instability during human development. *Nature Reviews Genetics* **11**: 786-799.

Megha T, Ferrari F, Benvenuto A, Bellan C, Lalinga A, Lazzi S, Bartolommei S, Cevenini G, Leoncini L and Tosi P. (2002) *p53* mutation in breast cancer. Correlation with cell kinetics and cell of origin. *Journal of Clinical Pathology* **55**: 461-466.

Meindl RS and Lovejoy CO. (1985) Ectocranial suture closure: A revised method for the determination of skeletal age at death based on the lateral anterior sutures. *American Journal of Physical Anthropology* **68**: 57-66.

Meissner C, Bruse P and Oehmichen M. (2006) Tissue-specific deletion patterns of the mitochondrial genome with advancing age. *Experimental Gerontology* **41**: 518-524.

Meissner C, von Wurmb N, Schimansky B and Oehmichen M. (1999) Estimation of age at death based on quantitation of the 4977-bp deletion of human mitochondrial DNA in skeletal muscle. *Forensic Science International* **105**: 115-124.

Meldgaard M and Morling N. (1997) Detection and quantitative characterization of artificial extra peaks following polymerase chain reaction amplification of 14 short tandem repeat systems used in forensic investigations. *Electrophoresis* **18**: 1928-1935.

Melk A, Ramassar V, Helms LMH, Moore RON, Rayner D, Solez KIM and Halloran PF. (2000) Telomere shortening in kidneys with age. *Journal of the American Society of Nephrology* **11**: 444-453.

Metzker ML. (2009) Sequencing technologies—the next generation. *Nature Reviews Genetics* **11**: 31-46.

Meyer E, Wiegand P, Rand S, Kuhlmann D, Brack M and Brinkmann B. (1995) Microsatellite polymorphisms reveal phylogenetic relationships in primates. *Journal of Molecular Evolution* **41**: 10-14.

Meyerson M. (2000) Role of telomerase in normal and cancer cells. *Journal of Clinical Oncology* **18**: 2626-2634.

Michikawa Y, Mazzucchelli F, Bresolin N, Scarlato G and Attardi G. (1999) Aging-dependent large accumulation of point mutations in the human mtDNA control region for replication. *Science* **286**: 774-779.

Miller FJ, Rosenfeldt FL, Zhang C, Linnane AW and Nagley P. (2003) Precise determination of mitochondrial DNA copy number in human skeletal and cardiac muscle by a PCR based assay: lack of change of copy number with age. *Nucleic Acids Research* **31**: e61-e61.

Miret JJ, Pessoa-Brandao L and Lahue RS. (1997) Instability of CAG and CTG trinucleotide repeats in Saccharomyces cerevisiae. *Molecular and Cellular Biology* **17**: 3382-3387.

Monckton DG, Wong L-JC, Ashizawa T and Caskey CT. (1995) Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses. *Human Molecular Genetics* **4**: 1-8.

Morales F, Couto JM, Higham CF, Hogg G, Cuenca P, Braida C, Wilson RH, Adam B, del Valle G and Brian R. (2012) Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity. *Human Molecular Genetics* **21**: 3558-3567.

Morla M, Busquets X, Pons J, Sauleda J, MacNee W and Agusti AGN. (2006) Telomere shortening in smokers with and without COPD. *European Respiratory Journal* **27**: 525-528.

Mullis KB. (1990) The unusual origin of the polymerase chain reaction. *Scientific American* **262**: 56-61.

Nakagawa S, Gemmell NJ and Burke T. (2004) Measuring vertebrate telomeres: applications and limitations. *Molecular Ecology* **13**: 2523-2533.

Napierala M, Parniewski P, Pluciennik A and Wells RD. (2002) Long CTG· CAG repeat sequences markedly stimulate intramolecular recombination. *Journal of Biological Chemistry* **277**: 34087-34100.

Nawrot TS, Staessen JA, Gardner JP and Aviv A. (2004) Telomere length and possible link to X chromosome. *The Lancet* **363**: 507-510.

Nestor CE and Monckton DG. (2011) Correlation of inter-locus polyglutamine toxicity with CAG• CTG triplet repeat expandability and flanking genomic DNA GC content. *PloS One* **6**: e28260.

Neumann AA and Reddel RR. (2002) Telomere maintenance and cancer? look, no telomerase. *Nature Reviews Cancer* **2**: 879-884.

Njajou OT, Cawthon RM, Damcott CM, Wu S-H, Ott S, Garant MJ, Blackburn EH, Mitchell BD, Shuldiner AR and Hsueh W-C. (2007) Telomere length is paternally inherited and is associated with parental lifespan. *Proceedings of the National Academy of Sciences* **104**: 12135-12139.

Nolin SL and Lewis F. (1996) Familial transmission of the FMR1 CGG repeat. *American Journal of Human Genetics* **59**: 1252.

Nordfjäll K, Larefalk Å, Lindgren P, Holmberg D and Roos G. (2005) Telomere length and heredity: Indications of paternal inheritance. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 16374-16378.

O'Callaghan Nathan J. VSD, Philip Thomas, and Michael Fenech. (2008) A quantitative real-time PCR method for absolute telomere length. *Biotechniques* **44**: 807-809.

Ohshima K and Wells RD. (1997) Hairpin Formation during DNA Synthesis Primer Realignmentin Vitro in Triplet Repeat Sequences from Human Hereditary Disease Genes. *Journal of Biological Chemistry* **272**: 16798-16806.

Okuda K, Bardeguez A, Gardner JP, Rodriguez P, Ganesh V, Kimura M, Skurnick J, Awad G and Aviv A. (2002) Telomere length in the newborn. *Pediatric Research* **52**: 377-381.

Olivier M, Hollstein M and Hainaut P. (2010) TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harbor perspectives in biology* **2**.

Orgel L. (1970) PNAS (USA) 49: 517 (1963). Orgel, LE. *PNAS (USA)* **67**: 1476.

Orgel LE. (1963) The maintenance of the accuracy of protein synthesis and its relevance to aging. *Science of Aging Knowledge Environment* **1963**: 8.

Ottini L, Falchetti M, Saieva C, De Marco M, Masala G, Zanna I, Paglierani M, Giannini G, Gulino A and Nesi G. (2004) *MRE11* expression is impaired in gastric cancer with microsatellite instability. *Carcinogenesis* **25**: 2337-2343.

Ou X, Zhao H, Sun H, Yang Z, Xie B, Shi Y and Wu X. (2011) Detection and quantification of the age-related sjTREC decline in human peripheral blood. *International Journal of Legal Medicine* **125**: 603-608.

Ozturk S, Sozen B and Demir N. (2014) Telomere length and telomerase activity during oocyte maturation and early embryo development in mammalian species. *Molecular Human Reproduction* **20**: 15-30.

Panchbhai A. (2011) Dental radiographic indicators, a key to age estimation. *Dentomaxillofacial Radiology* **40**: 199-212.

Pearson CE, Edamura KN and Cleary JD. (2005) Repeat instability: mechanisms of dynamic mutations. *Nature Reviews Genetics* **6**: 729-742.

Pearson CE, Ewel A, Acharya S, Fishel RA and Sinden RR. (1997) Human *MSH*2 binds to trinucleotide repeat DNA structures associated with neurodegenerative diseases. *Human Molecular Genetics* **6**: 1117-1123.

Pearson CE and Sinden RR. (1996) Alternative structures in duplex DNA formed within the trinucleotide repeats of the myotonic dystrophy and fragile X loci. *Biochemistry* **35**: 5041-5053.

Pearson CE and Sinden RR. (1998) Trinucleotide repeat DNA structures: dynamic mutations from dynamic DNA. *Current Opinion in Structural Biology* **8**: 321-330.

Pelletier R, Krasilnikova MM, Samadashwily GM, Lahue R and Mirkin SM. (2003) Replication and expansion of trinucleotide repeats in yeast. *Molecular and Cellular Biology* **23**: 1349-1357.

Peltomäki P. (2001) Deficient DNA mismatch repair: a common etiologic factor for colon cancer. *Human Molecular Genetics* **10**: 735-740.

Peltomäki P. (2005) Lynch syndrome genes. *Familial Cancer* **4**: 227-232.

Petruska J, Arnheim N and Goodman MF. (1996) Stability of intrastrand hairpin structures formed by the CAG/CTG class of DNA triplet repeats associated with neurological diseases. *Nucleic Acids Research* **24**: 1992.

Phillips C, Fernandez-Formoso L, Garcia-Magarinos M, Porras L, Tvedebrink T, Amigo J, Fondevila M, Gomez-Tato A, Alvarez-Dios J and Freire-Aradas A. (2011) Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel. *Forensic Science International: Genetics* **5**: 155-169.

Pinto D, Marshall C, Feuk L and Scherer SW. (2007) Copy-number variation in control population cohorts. *Human Molecular Genetics* **16**: R168-R173.

Pisano S, Galati A and Cacchione S. (2008) Telomeric nucleosomes: Forgotten players at chromosome ends. *Cellular and Molecular Life Sciences* **65**: 3553-3563.

Pompanon F, Bonin A, Bellemain E and Taberlet P. (2005) Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics* **6**: 847-846.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP and Gu Y. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**: 341.

Rattner J. (1995) Centromeres and telomeres. *Principles of Medical Biology* **2**: 93-120.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR and Chen W. (2006) Global variation in copy number in the human genome. *Nature* **444**: 444-454.

Ren F, Li C, Xi H, Wen Y and Huang K. (2009) Estimation of human age according to telomere shortening in peripheral blood leukocytes of Tibetan. *The American Journal of Forensic Medicine and Pathology* **30**: 252-255.

Richard G-F, Dujon B and Haber J. (1999) Double-strand break repair can lead to high frequencies of deletions within short CAG/CTG trinucleotide repeats. *Molecular and General Genetics MGG* **261**: 871-882.

Richard G-F, Goellner GM, McMurray CT and Haber JE. (2000) Recombination-induced CAG trinucleotide repeat expansions in yeast involve the *MRE11–RAD50–XRS2* complex. *The EMBO Journal* **19**: 2381-2390.

Richards RI. (2001) Dynamic mutations: a decade of unstable expanded repeats in human genetic disease. *Human Molecular Genetics* **10**: 2187-2194.

Richards RI and Sutherland GR. (1994) Simple repeat DNA is not replicated simply. *Nature Genetics* **6**: 114-116.

Rufer N, Brümmendorf TH, Kolvraa S, Bischoff C, Christensen K, Wadsworth L, Schulzer M and Lansdorp PM. (1999) Telomere fluorescence measurements in granulocytes and T lymphocyte subsets point to a high turnover of hematopoietic stem cells and memory T cells in early childhood. *The Journal of Experimental Medicine* **190**: 157-168.

Sacher GA. (1956) On the statistical nature of mortality, with especial reference to chronic radiation mortality. *Radiology* **67**: 250.

Samadashwily GM, Raca G and Mirkin SM. (1997) Trinucleotide repeats affect DNA replication in vivo. *Nature Genetics* **17**: 298-304.

Sarkar PS, Chang H-C, Boudi FB and Reddy S. (1998) CTG Repeats Show Bimodal Amplification in *E. coli Cell* **95**: 531-540.

Savouret C, Brisson E, Essers J, Kanaar R, Pastink A, te Riele H, Junien C and Gourdon G. (2003) CTG repeat instability and size variation timing in DNA repair-deficient mice. *The EMBO Journal* **22**: 2264-2273.

Savouret C, Garcia-Cordier C, Megret J, te Riele H, Junien C and Gourdon G. (2004) *MSH*2-dependent germinal CTG repeat expansions are produced continuously in spermatogonia from DM1 transgenic mice. *Molecular and Cellular Biology* **24**: 629-637.

Scheffler IE. (2007) *Mitochondria*: John Wiley and Sons.

Scherzinger E, Lurz R, Turmaine M, Mangiarini L, Hollenbach B, Hasenbank R, Bates GP, Davies SW, Lehrach H and Wanker EE. (1997) Huntingtin-encoded polyglutamine expansions form amyloid-like protein aggregates in vitro and in vivo. *Cell* **90**: 549-558.

Schmeling A, Olze A, Reisinger W, König M and Geserick G. (2003) Statistical analysis and verification of forensic age estimation of living persons in the Institute of Legal Medicine of the Berlin University Hospital Charite. *Legal Medicine* **5**: S367-S371.

Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB and Loeb LA. (2012) Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences* **109**: 14508-14513.

Schofield MJ and Hsieh P. (2003) DNA MISMATCH REPAIR: Molecular Mechanisms and Biological Function*. *Annual Reviews in Microbiology* **57**: 579-608.

Schranz D. (1959) Age determination from the internal structure of the humerus. *American Journal of Physical Anthropology* **17**: 273-277.

Schug M, Wetterstrand K, Gaudette M, Lim R, Hutter C and Aquadro C. (1998) The distribution and frequency of microsatellite loci in *Drosophila* melanogaster. *Molecular Ecology* **7**: 57-70.

Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, Liu Y, Weinstock GM, Wheeler DA and Gibbs RA. (2010) A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Research* **20**: 273-280.

Sherman S, Jacobs P, Morton N, Froster-Iskenius U, Howard-Peebles P, Nielsen K, Partington M, Sutherland G, Turner G and Watson M. (1985) Further segregation analysis of the fragile X syndrome with special reference to transmitting males. *Human Genetics* **69**: 289-299.

Sherman S, Morton N, Jacobs P and Turner G. (1984) The marker (X) syndrome: a cytogenetic and genetic analysis. *Annals of Human Genetics* **48**: 21-37.

Sherr CJ. (2004) Principles of tumor suppression. *Cell* **116**: 235-246.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM and Sirotkin K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29**: 308-311.

Shiels PG, Kind AJ, Campbell KHS, Waddington D, Wilmut I, Colman A and Schnieke AE. (1999) Analysis of telomere lengths in cloned sheep. *Nature* **399**: 316-317.

Shoubridge EA. (1994) Mitochondrial DNA diseases: histological and cellular studies. *Journal of Bioenergetics and Biomembranes* **26**: 301-310.

Simbolo M, Gottardi M, Corbo V, Fassan M, Mafficini A, Malpeli G, Lawlor RT and Scarpa A. (2013) DNA Qualification Workflow for Next Generation Sequencing of Histopathological Samples. *PloS One* **8**: e62692.

Sinden RR, Potaman VN, Oussatcheva EA, Pearson CE, Lyubchenko YL and Shlyakhtenko LS. (2002) Triplet repeat DNA structures and human genetic disease: dynamic mutations from dynamic DNA. *Journal of Biosciences* **27**: 53-65.

Solomon EP, Berg LR, Martin DW and Villee C. (2008) *Biology*: Thomson Brooks/Cole.

Soto AM and Sonnenschein C. (2004) The somatic mutation theory of cancer: growing problems with the paradigm? *BioEssays* **26**: 1097-1107.

Spivakov M and Fisher AG. (2007) Epigenetic signatures of stem-cell identity. *Nature Reviews Genetics* **8**: 263-271.

Strachan T and Read AP. (1999) Human Molecular Genetics. *An overview of Mutation, Polymorphism, and DNA Repair*.

Sukernik R, Derbeneva O, Starikovskaya E, Volodko N, Mikhailovskaya I, Bychkov IY, Lott M, Brown M and Wallace D. (2002) The mitochondrial genome and human mitochondrial diseases. *Russian Journal of Genetics* **38**: 105-113.

Sun JX, Helgason A, Masson G, Ebenesersdóttir SS, Li H, Mallick S, Gnerre S, Patterson N, Kong A and Reich D. (2012) A direct characterization of human mutation based on microsatellites. *Nature Genetics* **44**: 1161-1165.

Swami M, Hendricks AE, Gillis T, Massood T, Mysore J, Myers RH and Wheeler VC. (2009) Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. *Human Molecular Genetics* **18**: 3039-3047.

Szilard L. (1959) On the nature of the aging process. *Proceedings of the National Academy of Sciences of the United States of America* **45**: 30.

Takasaki T, Tsuji A, Ikeda N and Ohishi M. (2003) Age estimation in dental pulp DNA based on human telomere shortening. *International Journal of Legal Medicine* **117**: 232-234.

Takubo K, Nakamura K-I, Izumiyama N, Furugori E, Sawabe M, Arai T, Esaki Y, Mafune K-I, Kammori M and Fujiwara M. (2000) Telomere shortening with aging in human liver. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* **55**: B533-B536.

Teschendorff AE, West J and Beck S. (2013) Age-associated epigenetic drift: implications, and a case of epigenetic thrift? *Human Molecular Genetics* **22**: R7-R15.

Thompson R, Zoppis S and McCord B. (2012) An overview of DNA typing methods for human identification: past, present, and future. *DNA Electrophoresis Protocols for Forensic Genetics*. Springer, 3-16.

Thornton CA, Johnson K and Moxley RT. (1994) Myotonic dystrophy patients have larger CTG expansions in skeletal muscle than in leukocytes. *Annals of Neurology* **35**: 104-107.

Thorvaldsdóttir H, Robinson JT and Mesirov JP. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**: 178-192.

Toussaint O, Royer V, Salmon M and Remacle J. (2002) Stress-induced premature senescence and tissue ageing. *Biochemical Pharmacology* **64**: 1007-1009.

Tsuji A, Ishiko A, Takasaki T and Ikeda N. (2002) Estimating age of humans based on telomere shortening. *Forensic Science International* **126**: 197-199.

Unryn BM, Cook LS and Riabowol KT. (2005) Paternal age is positively linked to telomere length of children. *Aging Cell* **4**: 97-101.

Valdes AM, Andrew T, Gardner JP, Kimura M, Oelsner E, Cherkas LF, Aviv A and Spector TD. (2005) Obesity, cigarette smoking, and telomere length in women. *The Lancet* **366**: 662-664.

van den Broek WJ, Nelen MR, Wansink DG, Coerwinkel MM, te Riele H, Groenen PJ and Wieringa B. (2002) Somatic expansion behaviour of the (CTG) n repeat in myotonic dystrophy knock-in mice is differentially affected by *Msh3* and *Msh6* mismatch–repair proteins. *Human Molecular Genetics* **11**: 191-198.

Van Oorschot RA and Jones MK. (1997) DNA fingerprints from fingerprints. *Nature-London-*: 767-767.

Vaziri H, Schächter F, Uchida I, Wei L, Zhu X, Effros R, Cohen D and Harley CB. (1993) Loss of telomeric DNA during aging of normal and trisomy 21 human lymphocytes. *American Journal of Human Genetics* **52**: 661.

Verkerk AJ, Pieretti M, Sutcliffe JS, Fu Y-H, Kuhl D, Pizzuti A, Reiner O, Richards S, Victoria MF and Zhang F. (1991) Identification of a gene (*FMR*) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**: 905-914.

Vo AT, Zhu F, Wu X, Yuan F, Gao Y, Gu L, Li G-M, Lee T-H and Her C. (2005) hMRE11 deficiency leads to microsatellite instability and defective DNA mismatch repair. *EMBO Reports* **6**: 438-444.

von Wurmb-Schwark N, Bosinski H and Ritz-Timme S. (2007) What do the X and Y chromosomes tell us about sex and gender in forensic case analysis? *Journal of Forensic and Legal Medicine* **14**: 27-30.

von Wurmb-Schwark N, Higuchi R, Fenech A, Elfstroem C, Meissner C, Oehmichen M and Cortopassi G. (2002) Quantification of human mitochondrial DNA in a real time PCR. *Forensic Science International* **126**: 34-39.

Wambaugh J. (1989) The Blooding: True Story of the Narborough Village Murders. *New York: William Morrow*.

Wambaugh J. (2011) *The blooding*: Open Road Media.

Wang Z, Weber JL, Zhong G and Tanksley S. (1994) Survey of plant short tandem DNA repeats. *Theoretical and Applied Genetics* **88**: 1-6.

Warren S. (1956) Longevity and causes of death from irradiation in physicians. *Journal of the American Medical Association* **162**: 464-468.

Weber JL. (1990) Human DNA polymorphisms and methods of analysis. *Current Opinion in Biotechnology* **1**: 166-171.

Weber JL and May PE. (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics* **44**: 388.

Weber JL and Wong C. (1993) Mutation of human short tandem repeats. *Human Molecular Genetics* **2**: 1123-1128.

Weber-Lehmann J, Schilling E, Gradl G, Richter DC, Wiehler J and Rolf B. (2014) Finding the needle in the haystack: Differentiating "identical" twins in paternity testing and forensics by ultra-deep next generation sequencing. *Forensic Science International: Genetics* **9**: 42-46.

Wells RD, Dere R, Hebert ML, Napierala M and Son LS. (2005) Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucleic Acids Research* **33**: 3785-3798.

Wells RD, Parniewski P, Pluciennik A, Bacolla A, Gellibolian R and Jaworski A. (1998) Small slipped register genetic instabilities in Escherichia coli in triplet repeat

sequences associated with hereditary neurological diseases. *Journal of Biological Chemistry* **273**: 19532-19541.

Wheeler VC, Lebel L-A, Vrbanac V, Teed A, te Riele H and MacDonald ME. (2003) Mismatch repair gene *Msh2* modifies the timing of early disease in HdhQ111 striatum. *Human Molecular Genetics* **12**: 273-281.

Wong L, Ashizawa T, Monckton DG, Caskey C and Richards C. (1995) Somatic heterogeneity of the CTG repeat in myotonic dystrophy is age and size dependent. *American Journal of Human Genetics* **56**: 114.

Wu X, Xu Y, Chai W and Her C. (2011) Causal link between microsatellite instability and *hMRE11* dysfunction in human cancers. *Molecular Cancer Research* **9**: 1443-1448.

Zeichner SL, Palumbo P, Feng Y, Xiao X, Gee D, Sleasman J, Goodenow M, Biggar R and Dimitrov D. (1999) Rapid telomere shortening in children. *Blood* **93**: 2824-2830.

Zhao N, Zhu F, Yuan F, Haick AK, Fukushige S, Gu L and Her C. (2008) The interplay between *hMLH1* and *hMRE11*: role in MMR and the effect of *hMLH1* mutations. *Biochemical and biophysical research communications* **370**: 338-343.

Zubakov D, Liu F, Van Zelm M, Vermeulen J, Oostra B, Van Duijn C, Driessen G, Van Dongen J, Kayser M and Langerak A. (2010) Estimating human age from T-cell DNA rearrangements. *Current Biology* **20**: R970-R971.

Appendix I-III: Please see attached DVD

Appendix IV: Detailed analysis of SNPs

rs13067318	Base numbers in read											
Sample ID	**A**	**C**	**G**	**T**	**Total read count**	**Expected read count**	**Chi-square C**	**Chi-square T**	**Σx2**	**P value**	**% C/T**	**Genotype**
DMGV119C	0	66	0	0	66	33					100/0	C
DMGV18B	0	50	0	45	95	47.5	0.13	0.13	0.26	0.61	53/47	C/T
DMGV18C	0	35	0	32	67	33.5	0.07	0.07	0.13	0.71	52/48	C/T
DMGV99C	0	0	0	49	49	24.5					0/100	T
DMGV129C	0	0	0	69	69	34.5					0/100	T
DMGV133C	0	0	0	53	53	26.5					0/100	T
DMGV134C	0	0	0	86	86	43					0/100	T
DMGV22B	0	33	0	44	77	38.5	0.79	0.79	1.57	0.21	43/57	C/T
DMGV118C	0	0	0	65	65	32.5	32.50	32.50			0/100	T
DMGV13C	0	41	0	25	66	33	1.94	1.94	3.88	0.05	62/38	C/T
DMGV13B	0	32	0	27	59	29.5	0.21	0.21	0.42	0.52	54/46	C/T
DMGV106C	1	0	0	71	72	35.5					0/100	T
DMGV92C	0	78	0	0	78	39	39.00	39.00			100/0	C
DMGV91C	0	76	0	0	76	38	38.00	38.00			100/0	C
DMGV48C	0	32	0	22	54	27	0.93	0.93	1.85	0.17	59/41	C/T
DMGV29B	0	0	0	82	82	41					0/100	T
DMGV29C	0	0	1	89	90	44.5					0/100	T
DMGV47C	0	63	0	0	63	31.5	31.50	31.50			100/0	C
DMGV70B	0	42	0	40	82	41	0.02	0.02	0.05	0.83	51/49	C/T
DMGV70C	0	64	2	50	116	57	0.86	0.86	1.72	0.19	57/43	C/T
DMGV108C	0	76	0	0	76	38					100/0	C
DMGV57B	0	0	0	97	97	48.5					0/100	T
DMGV57C	0	0	0	92	92	46					0/100	T
DMGV86C	0	27	0	39	66	33	1.09	1.09	2.18	0.14	41/59	C/T
DMGV51C	0	49	0	25	74	37	3.89	3.89	7.78	0.01	66/34	C/T
DMGV51B	0	77	0	56	133	66.5	1.66	1.66	3.32	0.07	58/42	C/T
Total	1	841	3	1158	77							

Allele frequency for the rs13067318 SNP

Alleles	Number of Alleles observed	Allele Frequency observed	1,000 Genome GBR frequency
T	21	0.55	0.48
C	17	0.45	0.52

Hardy-Weinberg equilibrium calculation for the rs13067318 SNP

Genotype	Number of individuals observed	Genotype frequency	1,000 Genome GBR genotype frequency	Number of individual expected	Chi-square	
TT	7	0.31	0.23	4	1.57	
CT	7	0.49	0.50	9	0.65	
CC	5	0.20	0.27	5	0.00	
Total	19			Σx^2	2.22	P value = 0.33

The detailed analysis for the rs3855951 SNP

rs3855951	Base numbers in read											
Sample ID	A	C	G	T	Total read count	Expected read count	Chi-square C	Chi-square T	Σx2	P value	% T/C	Genotype
DMGV119C	0	1	0	72	73	36.5	34.53	34.53	69.05	9.58E-17	99/1	T
DMGV18B	0	1	2	100	103	50.5	48.52	48.52	97.04	6.80E-23	99/1	T
DMGV18C	0	0	0	132	132	66					100/0	T
DMGV99C	1	41	0	43	85	42	0.02	0.02	0.05	0.83	52/48	T/C
DMGV129C	0	1	0	114	115	57.5	55.52	55.52	111.03	5.81E-26	99/1	T
DMGV133C	0	0	1	76	77	38					100/0	T
DMGV134C	0	0	0	139	139	69.5					100/0	T
DMGV22B	0	1	1	96	98	48.5	46.52	46.52	93.04	5.12E-22	98/2	T
DMGV118C	0	1	2	97	100	49	47.02	47.02	94.04	3.09E-22	97/3	T
DMGV13C	0	0	0	89	89	44.5					100/0	T
DMGV13B	0	0	1	123	124	61.5					100/0	T
DMGV106C	0	1	1	131	133	66	64.02	64.02	128.03	1.11E-29	98/1	T
DMGV92C	1	0	1	72	74	36					100/0	T
DMGV91C	0	1	2	96	99	48.5	46.52	46.52	93.04	5.12E-22	99/1	T
DMGV48C	0	0	1	99	100	49.5					100/0	T
DMGV29B	1	2	0	94	97	48	44.08	44.08	88.17	6.02E-21	98/2	T
DMGV29C	1	3	1	120	125	61.5	55.65	55.65	111.29	5.10E-26	97/3	T
DMGV47C	0	0	0	106	106	53					100/0	T
DMGV70B	0	0	0	110	110	55					100/0	T
DMGV70C	0	0	0	132	132	66					100/0	T
DMGV108C	0	47	0	70	117	58.5	2.26	2.26	4.52	0.03	60/40	T/C
DMGV57B	0	0	0	106	106	53	53.00	53.00			100/0	T
DMGV57C	0	1	0	82	83	41.5	39.52	39.52	79.05	6.06E-19	99/1	T
DMGV86C	0	0	0	109	109	54.5	54.50	54.50			100/0	T
DMGV51C	2	0	0	124	126	62					100/0	T
DMGV51B	0	0	0	134	134	67					100/0	T
Total	6	101	13	2666	107							

Allele frequency for the rs3855951 SNP

Alleles	Number of Alleles observed	Allele frequency observed	1,000 Genome GBR frequency
T	36	0.95	0.96
C	2	0.05	0.04

Hardy-Weinberg equilibrium calculation for the rs3855951 SNP

Genotype	Number of individuals observed	Genotype frequency	1,000 Genome GBR genotype frequency	Number of individual expected	Chi-square	
TT	17	0.90	0.92	18	0.01	
TC	2	0.10	0.08	1	0.20	
CC	0	0.00	0.00	0	0.03	
Total	19			Σx^2	0.24	P value = 0.89

The detailed analysis for the rs10044327 SNP

rs10044327	Base numbers in read											
Sample ID	A	C	G	T	Total read count	Expected read count	Chi-square C	Chi-square T	Σx2	P value	% C/G	Genotype
DMGV119C	0	2	81	2	85	41.5	37.60	37.60	75.19	4.27E-18	5/95	G
DMGV18B	0	38	54	0	92	46	1.39	1.39	2.78	0.095	41/59	C/G
DMGV18C	0	34	53	2	89	43.5	2.07	2.07	4.15	0.04	39/61	C/G
DMGV99C	0	3	89	1	93	46	40.20	40.20	80.39	3.07E-19	3/97	G
DMGV129C	0	39	39	0	78	39	0.00	0.00	0.00	1.00	50/50	C/G
DMGV133C	0	43	30	0	73	36.5	1.16	1.16	2.32	0.13	59/41	C/G
DMGV134C	0	1	100	1	102	50.5	48.52	48.52	97.04	6.80E-23	2/98	G
DMGV22B	0	0	68	0	68	34					0/100	G
DMGV118C	1	67	0	0	68	33.5					100/0	C
DMGV13C	0	41	41	0	82	41					50/50	C/G
DMGV13B	0	48	46	0	94	47	0.02	0.02	0.04	0.84	51/49	C/G
DMGV106C	1	64	45	0	110	54.5	1.66	1.66	3.31	0.07	58/42	C/G
DMGV92C	0	50	29	0	79	39.5	2.79	2.79	5.58	0.02	63/37	C/G
DMGV91C	0	40	36	1	77	38	0.11	0.11	0.21	0.65	52/47	C/G
DMGV48C	0	75	2	1	78	38.5	34.60	34.60	69.21	0.00	96/3	C
DMGV29B	1	52	45	2	100	48.5	0.25	0.25	0.51	0.48	52/45	C/G
DMGV29C	1	46	40	0	87	43	0.21	0.21	0.42	0.52	53/46	C/G
DMGV47C	1	43	49	0	93	46	0.20	0.20	0.39	0.53	46/54	C/G
DMGV70B	1	100	0	0	101	50	50.00	50.00			100/0	C
DMGV70C	0	100	1	0	101	50.5	48.52	48.52	97.04	6.80E-23	99/1	C
DMGV108C	1	65	1	0	67	33	31.03	31.03	62.06	3.33E-15	97/3	C
DMGV57B	0	1	99	1	101	50	48.02	48.02	96.04	1.13E-22	99/1	G
DMGV57C	2	1	69	0	72	35	33.03	33.03	66.06	4.38E-16	96/4	G
DMGV86C	0	49	41	0	90	45	0.36	0.36	0.71	0.40	54/46	C/G
DMGV51C	0	0	69	1	70	34.5	34.50	34.50			0/100	G
DMGV51B	0	0	78	1	79	39	39.00	39.00			0/100	G
Total	9	1002	1205	13	86							

Allele frequency for the rs10044327 SNP

Alleles	Number of Alleles observed	Allele Frequency observed	1,000 Genome GBR frequency
C	19	0.50	0.43
G	19	0.50	0.57

Hardy-Weinberg equilibrium calculation for the rs10044327 SNP

Genotype	Number of individuals observed	Genotype frequency	1,000 Genome GBR genotype frequency	Number of individual expected	Chi-square	
CC	5	0.25	0.18	6	0.17	
CG	9	0.50	0.49	9	0.00	
GG	5	0.25	0.32	4	0.25	
Total	19			Σx^2	0.42	P value = 0.81

The detailed analysis for the rs267193 SNP

rs267193	Base numbers in read											
Sample ID	A	C	G	T	Total read count	Expected read count	Chi-square C	Chi-square T	Σx2	P value	% A/C	Genotype
DMGV119C	14	22	1	0	37	18	0.89	0.89	1.78	0.18	38/62	A/C
DMGV18B	0	47	1	0	48	23.5					0/100	C
DMGV18C	0	65	0	0	65	32.5					0/100	C
DMGV99C	15	24	0	0	39	19.5	1.04	1.04	2.08	0.15	38/62	A/C
DMGV129C	37	0	0	0	37	18.5					100/0	A
DMGV133C	10	12	0	0	22	11	0.09	0.09	0.18	0.67	45/55	A/C
DMGV134C	64	0	0	0	64	32					100/0	A
DMGV22B	19	25	0	0	44	22	0.41	0.41	0.82	0.37	43/57	A/C
DMGV118C	26	1	1	1	29	13.5	11.57	11.57	23.15	1.50E-06	93/6	A
DMGV13C	19	20	0	0	39	19.5	0.01	0.01	0.03	0.87	49/51	A/C
DMGV13B	14	24	0	1	39	19	1.32	1.32	2.63	0.10	38/62	A/C
DMGV106C	25	16	0	1	42	20.5	0.99	0.99	1.98	0.16	62/38	A/C
DMGV92C	17	15	0	0	32	16	0.06	0.06	0.13	0.72	53/47	A/C
DMGV91C	44	1	0	0	45	22.5	20.54	20.54	41.09	1.45E-10	98/2	A
DMGV48C	39	0	0	0	39	19.5					100/0	A
DMGV29B	18	12	0	0	30	15	0.60	0.60	1.20	0.27	60/40	A/C
DMGV29C	32	26	1	0	59	29	0.31	0.31	0.62	0.43	55/45	A/C
DMGV47C	13	30	0	0	43	21.5	3.36	3.36	6.72	0.01	30/70	C
DMGV70B	53	0	0	0	53	26.5					100/0	A
DMGV70C	53	0	0	0	53	26.5					100/0	A
DMGV108C	16	21	1	0	38	18.5	0.34	0.34	0.68	0.41	42/58	A/C
DMGV57B	53	0	0	0	53	26.5					100/0	A
DMGV57C	57	0	1	0	58	28.5					100/0	A
DMGV86C	33	0	0	0	33	16.5					100/0	A
DMGV51C	15	27	0	0	42	21	1.71	1.71	3.43	0.06	36/64	A/C
DMGV51B	10	21	0	0	31	15.5	1.95	1.95	3.90	0.05	32/68	A/C
Total	696	409	6	3	43							

Allele frequency for the rs267193 SNP

Alleles	Number of Alleles observed	Allele frequency observed	1,000 Genome GBR frequency
A	24	0.63	0.64
C	14	0.37	0.36

Hardy-Weinberg equilibrium calculation for the rs267193 SNP

Genotype	Number of individuals observed	Genotype frequency	1,000 Genome GBR genotype frequency	Number of individual expected	Chi-square	
AA	7	0.40	0.41	8	0.08	
AC	10	0.47	0.46	9	0.18	
CC	2	0.14	0.13	2	0.09	
Total	19			Σx^2	0.35	P value = 0.

The detailed analysis for the rs798755 SNP

rs798755	Base numbers in read											
Sample ID	A	C	G	T	Total read count	Expected read count	Chi-square C	Chi-square T	Σx2	P value	% A/G	Genotype
DMGV119C	0	0	23	0	23	11.5					0/100	G
DMGV18B	0	0	32	0	32	16					0/100	G
DMGV18C	0	0	30	0	30	15					0/100	G
DMGV99C	33	0	0	0	33	16.5					100/0	A
DMGV129C	4	0	4	0	8	4					50/50	A/G
DMGV133C	0	0	19	0	19	9.5					0/100	G
DMGV134C	0	0	35	0	35	17.5					0/100	G
DMGV22B	14	1	12	1	28	13	0.08	0.08	0.15	0.69	54/46	A/G
DMGV118C	0	0	26	1	27	13					0/100	G
DMGV13C	27	0	21	0	48	24	0.38	0.38	0.75	0.39	56/44	A/G
DMGV13B	32	1	8	0	41	20	7.20	7.20	14.40	0.00	80/20	A
DMGV106C	1	0	16	0	17	8.5	6.62	6.62	13.24	0.00	6/94	G
DMGV92C	0	0	18	0	18	9					0/100	G
DMGV91C	1	1	21	0	23	11	9.09	9.09	18.18	2.75E-04	4/96	G
DMGV48C	0	0	32	0	32	16					0/100	G
DMGV29B	21	0	12	0	33	16.5	1.23	1.23	2.45	0.12	64/36	A/G
DMGV29C	17	0	12	0	29	14.5	0.43	0.43	0.86	0.35	59/41	A/G
DMGV47C	0	0	34	0	34	17					0/100	G
DMGV70B	14	0	28	0	42	21	2.33	2.33	4.67	0.03	33/67	G
DMGV70C	15	0	16	0	31	15.5	0.02	0.02	0.03	0.86	48/52	A/G
DMGV108C	0	0	21	0	21	10.5					0/100	G
DMGV57B	0	2	29	0	31	14.5					0/100	G
DMGV57C	0	0	15	0	15	7.5					0/100	G
DMGV86C	0	0	28	0	28	14					0/100	G
DMGV51C	0	0	28	0	28	14					0/100	G
DMGV51B	0	0	11	0	11	5.5					0/100	G
Total	179	5	531	2	28							

Allele frequency for the rs798755 SNP

Alleles	Number of Alleles observed	Allele frequency Frequency observed	1,000 Genome GBR frequency
A	7	0.18	0.26
G	31	0.82	0.74

Hardy-Weinberg equilibrium calculation for the rs798755 SNP

Genotype	Number of individuals observed	Genotype frequency	1,000 Genome GBR genotype frequency	Number of individual expected	Chi-square	
AA	2	0.03	0.07	1	0.40	
AG	3	0.30	0.38	7	2.54	
GG	14	0.67	0.55	10	1.24	
Total	19			Σx^2	4.18	P value = 0.12

The detailed analysis for the rs13258459 SNP

rs13258459	Base numbers in read											
Sample ID	A	C	G	T	Total read count	Expected read count	Chi-square G	Chi-square T	Σx2	P value	% G/T	Genotype
DMGV119C	0	0	1	41	42	21	19.05	19.05	38.10	6.74E-10	3/97	T
DMGV18B	0	0	2	109	111	55.5	51.57	51.57	103.14	3.12E-24	2/98	T
DMGV18C	0	0	1	113	114	57	55.02	55.02	110.04	9.63E-26	1/99	T
DMGV99C	0	0	33	29	62	31	0.13	0.13	0.26	0.61	53/46	G/T
DMGV129C	0	0	25	39	64	32	1.53	1.53	3.06	0.08	39/61	G/T
DMGV133C	0	1	0	61	62	30.5					0/100	T
DMGV134C	0	0	0	116	116	58					0/100	T
DMGV22B	1	0	30	40	71	35	0.71	0.71	1.43	0.23	43/57	G/T
DMGV118C	0	0	32	33	65	32.5	0.01	0.01	0.02	0.90	49/51	G/T
DMGV13C	0	0	0	65	65	32.5					0/100	T
DMGV13B	0	0	0	73	73	36.5					0/100	T
DMGV106C	0	1	33	34	68	33.5	0.01	0.01	0.01	0.90	49/51	G/T
DMGV92C	0	0	34	51	85	42.5	1.70	1.70	3.40	0.07	40/60	G/T
DMGV91C	0	0	0	57	57	28.5					0/100	T
DMGV48C	0	0	30	43	73	36.5	1.16	1.16	2.32	0.13	41/59	G/T
DMGV29B	0	0	1	72	73	36.5	34.53	34.53	69.05	9.58E-17	2/98	T
DMGV29C	0	0	0	75	75	37.5					0/100	T
DMGV47C	0	0	36	47	83	41.5	0.73	0.73	1.46	0.23	43/57	G/T
DMGV70B	0	0	55	40	95	47.5	1.18	1.18	2.37	0.12	58/42	G/T
DMGV70C	0	0	37	34	71	35.5	0.06	0.06	0.13	0.72	52/48	G/T
DMGV108C	0	0	56	0	56	28					100/0	G
DMGV57B	0	0	72	0	72	36					100/0	G
DMGV57C	0	0	72	0	72	36					100/0	G
DMGV86C	0	0	58	0	58	29					100/0	G
DMGV51C	1	0	1	106	108	53.5	51.52	51.52	103.04	3.29E-24	1/99	T
DMGV51B	1	0	0	74	75	37					0/100	T
Total	3	2	609	1352	76							

Allele frequency for the rs13258459 SNP

Alleles	Number of Alleles observed	Allele frequency observed	1,000 Genome GBR frequency
T	23	0.61	0.58
G	15	0.39	0.42

Hardy-Weinberg equilibrium calculation for the rs13258459 SNP

Genotype	Number of individuals observed	Genotype frequency	1,000 Genome GBR genotype frequency	Number of individual expected	Chi-square	
TT	7	0.37	0.34	6	0.06	
GT	9	0.48	0.49	9	0.01	
GG	3	0.16	0.18	3	0.04	
Total	19			Σx^2	0.11	P value = 0.95

The detailed analysis for the rs1889073 SNP

rs1889073	Base numbers in read											
Sample ID	A	C	G	T	Total read count	Expected read count	Chi-square C	Chi-square G	Σx2	P value	% C/G	Genotype
DMGV119C	0	0	49	0	49	24.5					0/100	G
DMGV18B	0	46	21	0	67	33.5	4.66	4.66	9.33	2.26E-03	69/31	C/G
DMGV18C	1	80	39	1	121	59.5	7.06	7.06	14.13	1.71E-04	67/23	C/G
DMGV99C	0	78	1	0	79	39.5	37.53	37.53	75.05	4.59E-18	99/1	C
DMGV129C	0	48	41	0	89	44.5	0.28	0.28	0.55	0.46	54/46	C/G
DMGV133C	0	52	0	0	52	26					100/0	C
DMGV134C	0	51	33	0	84	42	1.93	1.93	3.86	0.05	61/39	C/G
DMGV22B	0	39	33	0	72	36	0.25	0.25	0.50	0.48	54/46	C/G
DMGV118C	0	61	0	0	61	30.5					100/0	C
DMGV13C	0	40	37	0	77	38.5	0.06	0.06	0.12	0.73	52/48	C/G
DMGV13B	0	46	60	0	106	53	0.92	0.92	1.85	0.17	43/57	C/G
DMGV106C	0	43	34	0	77	38.5	0.53	0.53	1.05	0.31	56/34	C/G
DMGV92C	0	0	97	0	97	48.5	48.50				0/100	G
DMGV91C	2	0	72	2	76	36	36.00				0/100	G
DMGV48C	0	35	56	0	91	45.5	2.42	2.42	4.85	0.03	39/61	C/G
DMGV29B	0	30	41	0	71	35.5	0.85	0.85	1.70	0.19	42/58	C/G
DMGV29C	0	53	38	0	91	45.5	1.24	1.24	2.47	0.12	58/42	C/G
DMGV47C	0	45	32	0	77	38.5	1.10	1.10	2.19	0.14	57/43	C/G
DMGV70B	0	48	36	0	84	42	0.86	0.86	1.71	0.19	57/43	C/G
DMGV70C	0	58	42	0	100	50	1.28	1.28	2.56	0.11	58/42	C/G
DMGV108C	0	109	1	2	112	55	53.02	53.02	106.04	7.24E-25	99/1	C
DMGV57B	0	49	49	1	99	49	0.00	0.00	0.00	1.00	50/50	C/G
DMGV57C	0	51	32	0	83	41.5	2.17	2.17	4.35	0.04	62/38	C/G
DMGV86C	0	112	0	0	112	56					100/0	C
DMGV51C	1	0	89	0	90	44.5					0/100	G
DMGV51B	0	1	92	0	93	46.5	44.52	44.52	89.04	3.86E-21	1/99	G
Total	4	1175	1025	6	85							

Allele frequency for the rs1889073 SNP

Alleles	Number of Alleles observed	Allele frequency observed	1,000 Genome GBR frequency
G	18	0.47	0.42
C	20	0.53	0.58

Hardy-Weinberg equilibrium calculation for the rs1889073 SNP

Genotype	Number of individuals observed	Genotype frequency	1,000 Genome GBR genotype frequency	Number of individual expected	Chi-square	
GG	4	0.22	0.18	3	0.13	
GC	10	0.50	0.49	9	0.06	
CC	5	0.28	0.34	6	0.30	
Total	19			Σx^2	0.49	P value = 0.78

The detailed analysis for the rs7090530 SNP

rs7090530	Base numbers in read											
Sample ID	A	C	G	T	Total read count	Expected read count	Chi-square A	Chi-square C	Σx2	P value	% A/C	Genotype
DMGV119C	4	12	0	0	16	8	2.00	2.00	4.00	0.045	25/75	A/C
DMGV18B	26	0	0	0	26	13					100/0	A
DMGV18C	30	0	0	0	30	15					100/0	A
DMGV99C	12	0	0	0	12	6					100/0	A
DMGV129C	10	12	0	0	22	11	0.09	0.09	0.18	0.67	46/54	A/C
DMGV133C	8	17	0	0	25	12.5	1.62	1.62	3.24	0.07	32/62	A/C
DMGV134C	18	24	0	1	43	21	0.43	0.43	0.86	0.35	43/57	A/C
DMGV22B	10	18	0	0	28	14	1.14	1.14	2.29	0.13	36/54	A/C
DMGV118C	21	0	0	0	21	10.5	10.50				100/0	A
DMGV13C	28	0	0	0	28	14	14.00				100/0	A
DMGV13B	30	0	0	0	30	15	15.00				100/0	A
DMGV106C	19	10	0	0	29	14.5	1.40	1.40	2.79	0.09	66/34	A/C
DMGV92C	0	21	0	0	21	10.5					0/100	C
DMGV91C	1	49	1	0	51	25	23.04	23.04	46.08	1.14E-11	2/98	C
DMGV48C	11	18	0	0	29	14.5	0.84	0.84	1.69	0.19	38/62	A/C
DMGV29B	30	0	0	0	30	15					100/0	A
DMGV29C	30	0	0	0	30	15					100/0	A
DMGV47C	10	9	0	0	19	9.5	0.03	0.03	0.05	0.82	53/47	A/C
DMGV70B	37	0	0	1	38	18.5					100/0	A
DMGV70C	33	0	0	0	33	16.5					100/0	A
DMGV108C	17	32	1	0	50	24.5	2.30	2.30	4.59	0.03	35/65	A/C
DMGV57B	52	0	0	0	52	26					100/0	A
DMGV57C	25	0	0	0	25	12.5					100/0	A
DMGV86C	10	10	0	0	20	10					50/50	A/C
DMGV51C	1	50	2	0	53	25.5	23.54	23.54	47.08	6.82E-12	1/99	C
DMGV51B	0	52	0	0	52	26					0/100	C
Total	473	334	4	2	31							

Allele frequency for the rs7090530 SNP

Alleles	Number of Alleles observed	Allele frequency observed	1,000 Genome GBR frequency
A	23	0.61	0.57
C	15	0.39	0.43

Hardy-Weinberg equilibrium calculation for the rs7090530 SNP

Genotype	Number of individuals observed	Genotype frequency	1,000 Genome GBR genotype frequency	Number of individual expected	Chi-square	
AA	7	0.37	0.32	6	0.11	
AC	9	0.48	0.49	9	0.01	
CC	3	0.16	0.18	4	0.07	
Total	19			Σx^2	0.19	P value = 0.91

The detailed analysis for the rs1079452 SNP

rs1079452	Base numbers in read											
Sample ID	A	C	G	T	Total read count	Expected read count	Chi-square C	Chi-square T	Σx2	P value	% C/T	Genotype
DMGV119C	1	36	0	38	75	37	0.03	0.03	0.05	0.82	48/52	C/T
DMGV18B	0	106	0	1	107	53.5	51.52	51.52	103.04	3.29E-24	99/1	C
DMGV18C	0	161	0	0	161	80.5					100/0	C
DMGV99C	1	88	0	1	90	44.5	42.52	42.52	85.04	2.92E-20	99/1	C
DMGV129C	1	56	0	57	114	56.5					50/50	C/T
DMGV133C	0	42	1	39	82	40.5	0.06	0.06	0.12	0.73	52/48	C/T
DMGV134C	1	55	3	60	119	57.5	0.11	0.11	0.22	0.64	48/52	C/T
DMGV22B	0	56	0	44	100	50	0.72	0.72	1.44	0.23	56/44	C/T
DMGV118C	0	33	0	32	65	32.5	0.01	0.01	0.02	0.90	51/49	C/T
DMGV13C	0	44	1	75	120	59.5	4.04	4.04	8.08	0.0045	37/63	C/T
DMGV13B	1	45	0	70	116	57.5	2.72	2.72	5.43	0.019	39/61	C/T
DMGV106C	0	58	0	58	116	58	0.00	0.00	0.00		50/50	C/T
DMGV92C	0	116	0	0	116	58					100/0	C
DMGV91C	0	106	0	0	106	53					100/0	C
DMGV48C	1	55	0	68	124	61.5	0.69	0.69	1.37	0.24	44/56	C/T
DMGV29B	1	121	0	0	122	60.5					100/0	C
DMGV29C	1	110	0	0	111	55					100/0	C
DMGV47C	0	37	1	63	101	50	3.38	3.38	6.76	0.01	37/63	C/T
DMGV70B	0	61	0	65	126	63	0.06	0.06	0.13	0.72	48/52	C/T
DMGV70C	0	75	1	57	133	66	1.23	1.23	2.45	0.12	57/43	C/T
DMGV108C	1	150	0	2	153	76	72.05	72.05	144.11	3.37E-33	99/1	C
DMGV57B	0	110	0	0	110	55					100/0	C
DMGV57C	1	97	0	0	98	48.5					100/0	C
DMGV86C	0	58	1	47	106	52.5	0.58	0.58	1.15	0.28	55/45	C/T
DMGV51C	0	52	0	47	99	49.5	0.13	0.13	0.25	0.62	53/47	C/T
DMGV51B	0	95	1	66	162	80.5	2.61	2.61	5.22	0.02	59/41	C/T
Total	10	2023	9	890	113							

Allele frequency for rs1079452 SNP

Alleles	Number of Alleles observed	Allele frequency observed	1,000 Genome GBR frequency
C	26	0.68	0.62
T	12	0.32	0.38

Hardy-Weinberg equilibrium calculation for the rs1079452 SNP

Genotype	Number of individuals observed	Genotype frequency	1,000 Genome GBR genotype frequency	Number of individual expected	Chi-square	
CC	7	0.47	0.38	7	0.01	
CT	12	0.43	0.47	9	1.04	
TT	0	0.10	0.14	3	2.74	
Total	19			Σx^2	3.79	P value = 0.15

The detailed analysis for the rs7131893 SNP

rs7131893	Base numbers in read											
Sample ID	A	C	G	T	Total read count	Expected read count	Chi-square A	Chi-square T	Σx2	P value	% A/T	Genotype
DMGV119C	0	0	0	74	74	37					0/100	T
DMGV18B	58	0	0	45	103	51.5	0.82	0.82	1.64	0.20	56/44	A/T
DMGV18C	75	0	1	50	126	62.5	2.50	2.50	5.00	0.03	60/40	A/T
DMGV99C	0	1	1	84	86	42					0/100	T
DMGV129C	58	1	0	44	103	51	0.96	0.96	1.92	0.17	57/43	A/T
DMGV133C	58	1	0	29	88	43.5	4.83	4.83	9.67	0.00	67/33	A/T
DMGV134C	82	0	0	56	138	69	2.45	2.45	4.90	0.03	59/41	A/T
DMGV22B	101	0	0	1	102	51	49.02	49.02	98.04	4.10E-23	99/1	A
DMGV118C	0	0	0	87	87	43.5					0/100	T
DMGV13C	88	2	1	40	131	64	9.00	9.00	18.00	0.00	69/31	A/T
DMGV13B	80	1	0	53	134	66.5	2.74	2.74	5.48	0.02	60/40	A/T
DMGV106C	0	0	0	89	89	44.5					0/100	T
DMGV92C	0	0	2	88	90	44					0/100	T
DMGV91C	1	0	2	85	88	43	41.02	41.02	82.05	1.33E-19	1/99	T
DMGV48C	65	0	1	52	118	58.5	0.72	0.72	1.44	0.23	56/44	A/T
DMGV29B	102	0	0	1	103	51.5	49.52	49.52	99.04	2.48E-23	99/1	A
DMGV29C	106	0	0	1	107	53.5	51.52	51.52	103.04	3.29E-24	99/1	A
DMGV47C	70	0	0	65	135	67.5	0.09	0.09	0.19	0.67	52/48	A/T
DMGV70B	0	0	0	126	126	63					0/100	T
DMGV70C	0	0	0	98	98	49					0/100	T
DMGV108C	63	0	0	38	101	50.5	3.09	3.09	6.19	0.01	62/38	A/T
DMGV57B	59	0	0	39	98	49	2.04	2.04	4.08	0.04	60/40	A/T
DMGV57C	45	0	0	50	95	47.5	0.13	0.13	0.26	0.61	47/53	A/T
DMGV86C	0	0	0	116	116	58					0/100	T
DMGV51C	84	0	0	41	125	62.5	7.40	7.40	14.79	0.00	67/33	A/T
DMGV51B	35	0	0	37	72	36	0.03	0.03	0.06	0.81	49/51	A/T
Total	1230	6	8	1489	105							

Allele frequency for the rs7131893 SNP

Alleles	Number of Alleles observed	Allele frequency observed	1,000 Genome GBR frequency
T	25	0.66	0.66
A	13	0.34	0.34

Hardy-Weinberg equilibrium calculation for the rs7131893 SNP

Genotype	Number of individuals observed	Genotype frequency	1,000 Genome GBR genotype frequency	Number of individual expected	Chi-square	
TT	8	0.44	0.44	8	0.01	
AT	9	0.45	0.45	9	0.03	
AA	2	0.12	0.12	2	0.02	
Total	19			Σx^2	0.06	P value = 0.97