Study of the hypervariable regions biology essay

Science, Biology

Aim: To determine the variability of the HV1 and HV2 regions in the mtDNA of the Maltese population. Objectives: To determine the sequences of the mtDNA HV1 and HV2 regions of 100 Maltese individuals. To predict the haplogroup of each mtDNA sample, and hence determine which human migrational event most influenced the Maltese mtDNA population. To determine how many different haplotypes are present in 100 samples. To determine which are the most common SNPs and base transversions in Maltese HV1 and HV2 regions. Introduction: MtDNA structure and functionMtDNA was originally the genetic code of α-proteobacteria, before becoming endosymbionts as stated in the endosymbiotic theory. During the course of evolution, many of the genes necessary for the mitochondria to function autonomously were either lost due to deleterious mutations or transposed into the host genome in the form of nuclear mitochondrial DNA (numts), many of which are inactive. As a result, modern mtDNA contains none of the non essential genes of the genetic material of the original endosymbiont (Kurland, Anderson, 2000). MtDNA is located within the matrix of mitochondria. Human mtDNA is about 5µm long with a molecular mass of 107 daltons (Giles et al, 1980). It includes 16, 569 bp. This is a very small number when compared to nDNA, which is over 3 billion bp long. However, it is essential for the normal functioning of mitochondria. MtDNA is circular, covalently closed and double-stranded (Taylor & Turnbull, 2007). One of the strands is Guanine-rich, while the other is Cytosein-rich. The Guanine-rich strand is heavy (H), while the Cytosein-rich strand is light (L), since Cytosein has one less ring then Guanine. The H strand is found on the outside of the mtDNA (Clayton, Hughes & Chase, 2000). 28 genes are located on the H strand, while nine more are located on the L strand. Thus, it has a total of 37 genes, which is a small minority when compared to the entire human genome of 20, 000 genes. 13 mtDNA genes code for oxidative phosphorylation enzymes. These play a role in the formation of ATP from Oxygen and sugars, or for proteins involved in electron transport. Another 22 genes code for tRna. The last two genes code for a small and large rRna subunit respectively. These subunits arrange amino acids to proteins (Ingman & Gyllensten, 1982). The Displacement loop (D-loop) is triple stranded, and contains most of the noncoding segment, called the control region, which is 1122 bp long. The sequences found in this region are important for regulating transcription of the H and L strands, and also acts as a promoter for the replication of the H strand (Clayton 1982). Since no genes are present, any mutation that occurs in this region is relatively non-lethal. As a result, there is much more variability in this region from one individual to another then the coding region. There are two regions of high variability in the control region. These are called the hypervariable regions 1 and 2 (HV1 and HV2) (Seyedhassani et al. 2010). The CCCCCCCTCCCCC area in the HV2 region is important for the formation of a persistent DNA-RNA hybrid, which is essential for H strand replication (Lee & Clayton 1998). The origin of replication of the H strand is located between HV1 and HV2. HV1 spans from the bp 16024 to 16400, while HV2 spans from 19 to 410. The mtDNA has a very efficient arrangement, with no introns, and overlapping reading frames (Taylor & Turnbull, 2007). There are from 100-10, 000 copies of mtDNA molecules per cell. Oocytes contain even more mtDNA molecules, with a total of about 100, 000 per oocyte. All metazoans have the same pattern of genes in their mtDNA but one or more genes may be absent, and sizes differ (Taylor & Turnbull, 2007). Replication and segregation of mtDNA and mitochondriaThe Light-Strand Promoter (LSP) is located in the D-loop and just prior to the origin of the H strand replication (OH). For mtDNA replication to begin, the LSP must first transcribe RNA primers on the OH. This would initiate replication. Transcription of the LSP also produces mRNA transcript of the ND6. Therefore, there must be a mechanism to switch from one function to another. There are three conserved sequence blocks (CSB) found downstream to the LSP. CSBII is important to increase the stability of the RNA-DNA hybrid (R-loop). The LSP transcript may be cleaved by RNase MRP at specific sites, and the resulting transcription fragments may act as primers. However, RNase MRP is mainly found in the nucleolus. CSBII is also able to terminate LSP transcription. This may be an alternative mechanism by which the LSP switches functions. CSBII ends the transcription when the 3' end of the nascent H strand reach bp numbers 300-282. This region happens to be where the transition from the RNA primer to the new DNA strand takes place in the D-loop (Brown et al. 2005). Once the RNA primer forms in the origin of replication of the H strand, synthesis commences unidirectionally, but commonly halts after 700 bp downstream. These 700 bp form the 7S DNA, and produces the triple stranded D-loop. The enzyme mtDNP may be responsible for terminating the replication bidirectionally, since it has contrahelicase activity. mtDNP may bind to termination-associated sequences (TAS) located at the 3' end of the nascant H strand (Brown et al. 2005). From this point onwards, there are two possible models by which replication continues to occur. The eldest model is called the strandsymmetric replication. In this model, the transcription recommences from the 3’ end of the 7S DNA in the D-loop. This strand is called the leading H strand. When the leading strand reaches two thirds of the mtDNA, the origin of L-strand replication (OL) gets exposed. The exposure activates the OL, and replication of the lagging L-strand commences in the opposite direction. The lagging strand stops replicating much after the leading strand is ended. Thus, the two daughter mtDNA molecules are not completed simultaneously. Segregation takes place before the lagging strand completes (Brown et al. 2005). The more modern model is called synchronous replication, where mtDNA replicates similarly to nDNA, with the leading and lagging strands replicating at the same time, and synchronously. Replication forks will form y arcs as a result of this mode of replication. These are mostly present between OH and OL. Evidence exists for both models (Brown et al. 2005). There are several enzymes involved in mtDNA replication. One enzyme is mtDNA polymerase y (POLy), which adds nucleotides to the 3’ end of the template. There are two subunits of POLy, called POLy A and B. these are associated with eachother to form POLyAB2, a heterotrimer. POLyB increases the catalytic properties of POLyA. Another enzyme involved is the mitochondrial TWINKLE Helcase. This functions to unwind the mtDNA, which exposes the bases to POLy. Another important enzyme is the single-strand binding protein, which enables manipulation of single stranded DNA. It also increases the rate that TWINKLE unwinds DNA. All this enzymes work together and form the Replisome (Brown et al. 2005). In contrast to nDNA, mtDNA replication does not depend on cell cycles. This makes mtDNA replication more relaxed, and some mtDNA molecules may get replicated multiple times in a cell cycle, or not at all. However, the number of copies of mtDNA is under strict regulation. Their numbers depend on the cell type (Shoubridge 2000). MtDNA replication is part of mitochondrial duplication. Replication is followed by mtDNA segregation, which is then followed by organellar segregation to form daughter cells. All this process is stochastic in nature. There is no control over which copies are replicated. Thus, within a cycle, some mtDNA molecules are replicated more often then others. If there are different haplotypes, this leads to a change in allele representation in the daughter mitochondria. It is not known whether different mtDNA molecules are segregated in the mitochondrion prior to mitochondrial division, or whether they are inherited depending on which side of the parent mitochondrion where happened to be in. Mitochondrial replication is independent of the cell cycle, and turnover may take place in all stages of the cell cycle. During mitosis, mitochondrial segregation into daughter cells also seems stochastic (Birky 2001). Recombination in human mtDNAmtDNA is currently thought to be clonal. This means that it is uniparentally inherited, and does not undergo recombination. However, there are two sources of evidence which imply that recomnination occurs. The first source is the excess homoplasy observed in mtDNA lineages. These are either caused by recombination or multiple mutations. However, if mtDNA is nonclonal, only multiple mutiatios would be possible. Since homoplasies are so common, If all the homoplasies are generated by multiple mutations, there must be some sites in the mtDNA that are hypervariable. The second source of evidence is the observed linkage disequilibrium in mtDNA. If recombination occurs, linkage disequilibrium should decrease, since linked genes at different loci will separate. Linkage disequilibrium also decreases with increasing distance between different loci in the mtDNA. This is also true for nDNA (Eyre-Walker 2000). There are three pathways by which mtDNA may become nonclonal. The first pathway is when two different types of mtDNA molecules in a heteroplasmic cell recombine. The second possibility is by a mtDNA molecule recombining with a numt, while the third is by a maternal mtDNA recombining with a parternal mtDNA. For all these mechanisms to be possible, a mitochondrion must be able to include enzymes that catalyse homologous recombination, and they must be able to take up recombinant DNA from the cytoplasm (Eyre-Walker 2000). Heteroplasmy is the simplest route towards recombination. However, this mode of recombination has not been detected yet, either because it does not occur in vivo, or because it occurs at a rate too slow to be detectable, or because mitochondrial fusion is rare (Eyre-Walker 2000). There are approximately 1000 copies of numts in the nDNA. For recombination with a numt to be possible, the numt must first be transcribed and then reverse transcribed in the nucleus. It must then be transported to the cytoplasm and then taken up by a mitochondrion. The mitochondrion must therefore be able to take up DNA. Although DNA can exit a mitochondrion, no flow of DNA into a mitochondrion has been detected yet (Eyre-Walker 2000). The most likely route towards nonclonality is via paternal leakage, since paternal mtDNA molecules are known to enter the oocyte. However, these are destroyed quickly by ubiquitination. Paternal inheritance itself does not contribute to nonclonality. Recombination between paternal and maternal mtDNA must occur either my mitochondrial fusion or by uptake of paternal mtDNA. Mitochondrial fusion is rare in humans, but does occur in embryogenesis, and as previously discussed, uptake of uptake of external DNA is not evident. However, if the elimination system of paternal mtDNA is somehow broken down, paternal mtDNA may recombine with maternal mtDNA during embryogenesis, since mtDNA does include genes for recombinase enzymes. Also, the very act of breaking down paternal mitochondria releases paternal mtDNA into the cytoplasm, where it might be taken up by a maternal mitochondrion (Eyre-Walker 2000). A single case was observed in the only known human with biparental mtDNA, where recombination was observed in about 0. 7% of all the mtDNA in the muscle tissue. In most cases, although there is no conclusive evidence of mtDNA recombination, even if it does occur, since mtDNA is maternally inherited, and provided that there is no heteroplasmic mtDNA molecules, the recombined mtDNA will be identical to the original, since the two parental mDNA molecules where also identical to each other (Pakendorf & Stoneking 2005). Even though there is very little evidence of recombination, the consequences of recombination make it important to determine whether it occurs or not. If it does occur, it would drastically change our understanding of human evolution and migration. These analyses all assume that mtDNA is clonal. However, if the mtDNA is nonclonal, the genealogy of a mtDNA molecule will not represent its matrilineal line, but the average of different genealogies. As a result, dates of coalescence of two lineages will be underestimated, since recombination would average the age of the different lineages found in the mtDNA. The more freely mtDNA recombines, the more averaged the ages of different lineages will become (Pakendorf & Stoneking 2005). Another reason why recombination would lead to an underestimation of the age of a lineage is because homoplasies would no longer come about solely due to the hypervariable regions. The hypervariable sites are taken into account when estimating the coalescence time, and this gives a more recent calculated age of a lineage. This is because they have a higher rate of mutation. Thus, the characteristic mutations of the lineage being analysed would seem to have taken less time to evolve. This calculation would be accurate if the hypervariable regions really are responsible for all observed homoplasies. However, if recombination is also responsible for the some of them, the inclusion of the hypervariable sites in the mutation rate would make the calculated coalescence time seem more recent then it actually is, since the actual mutation rate is slower then the calculated one (Eyre-Walker 2000). HeteroplasmyA single individual may have different types of mtDNA within its genome. This is known as heteroplasmy. The more common type of mtDNA is called the wild type, while the other type is called the heteroplasmic mtDNA. One study shows heteroplasmy to occur in 70% of a population. It most commonly occurs as length heteroplasmy in the poly(C) tracts of HV1 and HV2. 3. 81% of individuals also show point heteroplasmy (Santos et al 2008). The general homoplasmy of mtDNA implies that a bottleneck in the number of mitochondria takes place early on in oogenesis. Heteroplasmy is more common within aged cells. When cells are extremely deficient in mitochondria, a mutation of a single mitochondrion could outcompete the wild-type population of mitochondria in the cell. This is unexpected since the healthy mitochondria are being replaced by functionally impaired mutants. Aged cells also have more variety of mitochondrial mutations(Kowald & Kirkwood 2011). Since mtDNA has a very high mutation rate and is subject to genetic drift, heteroplasmy should be very common for a mtDNA population. In contrast, homoplasmy should be uncommon. However, the opposite situation is observed, and this is because several mechanisms exist that reduce the genetic diversity of mtDNA molecules within an individual. Such mechanisms include germline bottlenecks, gene transfer and selection against deleterious mutations. Despite the various mechanisms to promote homoplasmy, a small degree of heteroplasmy is still observed in humans, as well as other organisms. Heteroplasmy may not necessarily arise due to mutations, but may also come about by the coexistence of two separate lineages of mtDNA, as is the case in paternal leakage. Nevertheless, the most common form of heteroplasmy is in cases of point mutations, although the number of cases of paternal leakage is increasing (White, Wolff & Pierson 2008). The bottleneck that takes place during an embryo’s early development partly accounts for the great reduction in mtDNA diversity within an individual. It continually delays the accumulation of deleterious mutations and subsequent mutational meltdown via Muller’s ratchet. It is not known for certain whether the greatest bottleneck takes place during embryogenesis or oogenesis. During embryogenesis, the total number of mtDNA molecules remain constant. Most of the cells in the blastocyst will not form part of the embryo. Only a small group of cells called the inner cell mass (ICM) will contribute. This gives rise to a large bottleneck especially for rare haplotypes. During oogenesis, a small number of progenitor germ cells (PGCs) divide to form a large number of germ cells. Each PGC only contains a small number of mitochondria, and when they divide, the number of mitochondria dramatically increases. The large bottleneck during embryogenesis means that only a small number of mtDNA molecules will repopulate the embryo during oogenesis. This frequently leads to a return to homoplasmy if the individual is heteroplasmic. However, this could also lead to founder effects in the rare cases where the rare haplotype is not eliminated during embryogenesis (White, Wolff & Pierson 2008). Heteroplamsy can still be heritable. However, the persistence of a neutral heteroplasmic mtDNA depends on its population size in an individual. Heteroplasmy could either persist for many generations, or return to homoplasmy in a single generation. The heritability of a selectively neutral heteroplasmic mtDNA molecule can be influenced by the locus of its different mutations. If they are located next to an advantageous mutation, it is more likely to be inherited due to hitchhiking (White, Wolff & Pierson 2008). Heteroplasmy is selected against very strongly when associated with genetic diseases, often being fixed in one generation. However, they may persist in many generations, if they are below a certain threshold of deleteriousness. This is because the wild-type mtDNA that coexist with the deleterious mtDNA in the same individual can compensate for their functional deficiencies. Deleterious mutations can also persist due to a replicative advantage. For example, wild-type mtDNA molecules have a selective advantage over mtDNA with deleterious deletions, because they have better functioning. However, mtDNA with deleterious deletions also have a selective advantage over the wild type because they require less resources and less time to replicate (White, Wolff & Pierson 2008). There are currently two proposed mechanisms by which deleterious mutations are selected against. One theory is that a deleterious mutation makes the mitochondrion autophagic, while the alternative suggests that deleterious mutations makes the mitochondrion less efficient at replicating, leading to it being outcompeted by the functional wild type (White, Wolff & Pierson 2008). InheritanceThe theory of neutrality states that the great majority of intraspecific and interspecific molecular evolutionary change is caused by fixation due to genetic drift, rather then selection of fitter individuals. A modified version of the neutral theory is the nearly neutral theory, which states that slightly deleterious or slightly advantageous mutations could also become fixed by genetic drift. It also adds that the probability of fixation is inversely proportional to the population size (Ohta 2007). The control region of the mtDNA is commonly thought to be selectively neutral. However, this is debatable. It might be neutral due to its high evolutionary rate, the uniformity in substitution rates, and the relaxed translation of mitochondrial mRNAs. However, it more likely follows the mildly deleterious model. This is evidenced in the greater ratio of intraspecific replacement to silent polymorphisms then expected by the neutral model. The ratio of silent substitutions to replacement substitutions is used as a test used to determine the cause of molecular evolution. Silent or synonymous substitutions are substitutions in the DNA sequence that does not translate to an alteration of the amino acid sequence. This could be either because the substitution occurred in a noncoding region, or because of the degeneracy of the genetic code. On the other hand, replacement or nonsynonymous substitutions give rise to amino acid polymorphism (AAP), and thus are expressed phenotypically. According to the neutral theory, the frequency of silent and replacement mutations should both be caused by drifting, and their ratios should remain constant for a given locus. If the amount of replacement mutations are relatively increased, the ratio will alter, indicating a deviation from neutrality . this is the basis of the MacDonald and Kreitman test (Nielsen 1997). When all the mtDNA loci are considered together, there is an excess of AAP. Different lineages of mtDNA undergo different rates of heterogeneous substitution, and the disproportionate rate of replacement polymorphism continue to imply that selection influences mtDNA polymorphism. However, genetic drift may be the prevailing force. Given its small size, mtDNA does not provide many targets for directional selection (Nachman, Boyer & Aquadro 1994). In humans, and in the great majority of other species, mtDNA is mostly inherited from mother’s ovum. Sperm do not usually contribute any mitochondria when fertilizing the oocyte. A dominant or recessive mode of inheritance is unlikely when mtDNA is concerned because segregation is inevident. Evidence against segregation is that mtDNA inheritance shows a strong association between the phenotype of the progeny and the sex of the parent. By contrast, segregation does not show any such association (Giles et al, 1980). There are several reasons why the paternal mtDNA is generally not inherited. One reason is that most of the mtDNA present in sperm are present on the midpiece, which is found between the head and tail regions. In most cases, this is lost during fertilization since it does not enter the oocyte. Consequently, all progeny are hemizygous for maternal mtDNA (Ankel-Simons et al. 1996). A second reason arises due to the relative amounts of mtDNA in the oocyte and sperm respectively. The oocyte contains approximately 100, 000 copies of mtDNA while the sperm midpiece only contains about 100. Therefore, the minority of paternal mtDNA molecules that get incorporated in the egg undergo dilution. This low proportion of paternal mtDNA to maternal mtDNA makes the paternal mtDNA more susceptible to loss by drift. The effects of drift is further increased by the bottleneck during oogenesis (Birky 2001). A third reason is that paternal mtDNA within the fertilized oocyte get selectively destroyed by ubiquitination. During spermatogenesis, parental mitochondria of the secondary spermatocytes are first tagged with ubiquitin when still in the epididymis. The ubiquitin attaches to the mitochondrial membrane proteins. However, the ubiquitinated mitochondria are undetectable when outside the oocyte. This is because of the presence of sulfur cross bridges. Once some of the paternal mitochondria enter the oocyte, the sulfur cross bridges are broken down. Hence, the ubiquitinated mitochondria become detectable again, and are proteolytically digested by the oocyte’s machinery (Sutovsky et al. 1999). It is not known for certain why paternal mtDNA molecules are eliminated. One explanation is to prevent competition between different mtDNA haplotypes, possibly resulting in lethal genome conflict. An alternative reason is that it is an adaptation to prevent the inheritance of mitochondria that are potentially damaged by reactive oxygen species byproducts from oxidative phosphorylation (Bromham et al. 2003). Despite these reasons, some evidence indicates that paternal mtDNA inheritance is still possible. This occurs most commonly in interspecific hybrids of closely related species. One exception also exists in humans. A man with exercise intolerance was reported to have predominantly paternal muscular mtDNA. This raises concern about tracing back a population, since it may not always be accurate since biparental inheritance has a confounding effect. There are no more cases of paternal mtDNA inheritance in humans. It is very possible that this unique case is more likely a very rare example of a failure of the oocyte machinery to break down the paternal DNA, which is normally marked for destruction by ubiquitination. Due to the lack of further evidence of paternal inheritance in humans, maternal inheritance of mtDNA is still considered as the norm (Schwatz & Vissing 2002). Causes of mtDNA mutationsMtDNA is more prone to mutations then nDNA because mtDNA is not able to repair damage as efficiently as nDNA does. Also, mtDNA polymerase lacks fidelity, and no proof-reading is done (Birky 2001). One protective strategy it does make use of is to package itself by histone proteins. This will then aggregate into nucleoprotein complexes called nucleoids, with 2 to ten mtDNA molecules per nucleoid. However, this packaging is not as compact as those found in nDNA (White, Wolff & Pierson 2008). Furthermore, mtDNA is situated at the site of cellular respiration. This makes it even more vulnerable to mutation because the mitochondrial matrix is the site that produces reactive oxygen species such as •O2-, •OH and H2O2. These are by-products that form during oxygen reduction in the electron transfer system. These combined factors contribute to a higher mutation rate in the mtDNA (Lee & Wei 2007). MtDNA mutations could also arise from replication slippage due to mitochondrial short tandem repeats (STRs). Another cause of mtDNA mutations could be due to defects in nDNA. nDNA controls mtDNA replication by active intergenomic communication with the 5’ end of the D-loop. Thus a mutation in specific protein-coding regions of the nDNA could sabotage the normal functioning of the mtDNA in an inheritable manner (Zeviani et al. 1989). Genetic diseases associated with mtDNA mutationsMany genetic conditions are a result of somatic deleterious mutations on the mtDNA. Multiple mtDNA deletions could give rise to opthalmoplegia. Cancers such as breast, colon and stomach cancers may also result from mitochondrial mutations. During such cancers, the D-loop region of the mitochondria becomes a very common site of mutations, some of which are deletions. Deletions could alter the cell proliferation, and are associated with mitochondrial myopathies. On somatic cells, deletions could result in somatic cell disorders, while in germ cells, this could lead to early apoptosis and foetal loss (Solano & Playán 2001). Certain mtDNA mutations could also increase the rate of production of reactive oxygen species, which will further increase the amount of mutations. Mutations in the regulatory region of the D-loop could have an effect on the mtDNA copy number and its gene expression. This could in turn have a negative effect the oxidative phosphorylation processes and the production of ATP. Repeated pregnancy loss may also be associated with a high rate of point mutations in the regulatory region of the D-loop. There are over 200 mtDNA mutations that are thought to be linked with various human diseases. Some examples include the following: MutationDisease16126CBreast and endometrial cancers, glioblastoma multiform16189CType 2 diabetes, some metabolic syndromes16223Tinfantile sudden death syndrome, schizophrenia, age-related muscular degeneration16294Tinfantile sudden death syndrome, Parkinson disease, age-related muscular degeneration16311CProstatic cancer16519CDiabetes mellitus, gasterointestinal disordersTable 1: A table showing various SNPs in the HV1 region of human mtDNA, that are associated with various genetic diseases. Mutations associated with the 310 bp region are frequent mutations that are found in tumours. 152C is associated with respiratory morbidity in children. This mutation is a characteristic of H, U and K haplogroups (Seyedhassani et al. 2010). MtDNA mutations are also related to the aging process (Kowald & Kirkwood 2011). Mutation ratesThe human mitochondrial molecular clock is the rate at which mutations are accumulating in the mitochondria. The coding region of mtDNA has a mutation rate of 0. 017*10-6 substitutions per site per year (Ingman, Kaessmann, Paabo & Gyllensten 2000). The noncoding region has an even higher mutation rate. The rate of mutations in this region is more controversial, however an estimate from multiple phylogenetic comparisons is of 0. 075-0. 165*10-6 substitutions per site per year (Hasegawa, Di Rienzo, Kocher & Wilson 1993). Another estimate from direct observations from pedigree estimates is of 0. 0-1. 46*10-6 with an average of 0. 46*10-6. This discrepancy makes it hard to determine the true rate of mutation, and which is best suited for studies. An explanation of the large differences in estimation is that some sites in the control region are mutational hotspots, where mutation occurs approximately five times faster then average 37. One example of a hypermutable region is between 308-315 in the HV2 region. These regions are more vulnerable to damage (Hasegawa, Di Rienzo, Kocher & Wilson 1993). In pedigree estimates, the mutations occurring at hypermutable regions are more readily detected then other regions, while in phylogenetic estimates, mutations at the less mutable sites are detected as well (Heyer et al. 2001). Another study suggests that the difference does not arise because of one factor, but due to the combined effects of genetic drift, selection and low detection rates in phylogenetic analysis. Since such a large number of mutations detected in the pedigree analysis are recent, it is unlikely that they will become fixed, and the pedigree rate is better suited for studies concerning the recent population history, while the phylogenetic rate is used for deep history studies, since the mutations detected have had enough time to reach a significant frequency within the population (Macaulay et al. 1997). An alternative is to classify different sites with different mutation rates. One study showed that when this model is used to estimate the age of the mitochondrial eve, the value obtained was half that given if the rate was constant for every region. If each site has a different mutational rate, average estimates of the human mtDNA mutational rate do not reflect reality (Hasegawa, Di Rienzo, Kocher & Wilson 1993). Uses of mtDNAThe hypervariable regions 1 and 2 in the d-loop are highly variable from one individual to another. This makes them useful in forensic science. In contrast, the coding region not as useful because its sequences are much less variable in the human population. This is because a mutation in this region could easily be fatal, and hence would not be inherited. The sequence information is reported by comparing the analysed sequence to a standard called the revised Cambridge reference sequence (rCRS). The rCRS makes use of the light strand, therefore when analyzing mtDNA samples, the light strand should also be used (Bini et al. 2003). Somatic cells generally only have two copies of nDNA. In contrast, hundreds of mtDNA copies are found per cell 125. This feature, along with its extranuclear location, makes mtDNA more accessible for analysis. However, the multiple copies also complicates the its population genetics since there are multiple levels of mtDNA population genetics, be it in a single mitochondrion or a population of individuals {Pakendorf & Stoneking 2005}. Whether mtDNA is selectively neutral or not has not strongly influenced its uses in phylogenetic studies. However, mtDNA neutrality is important when estimating genetic distances and molecular clocks. Regardless of its neutrality, mtDNA remains a good indicator for processes in a population. It can be used to identify geographic clusters of related individuals or the matrilineal relationships within populations. It is also useful in discovering the genetic history of a population such as bottlenecks, or for hybrid zones. It is also useful in confirming a phylogenetic relationship between two taxa (Moritz et al. 1987). Since haplogroups of mtDNA share a common ancestry, they can be used to quantify the amount of interbreeding between two previously isolated populations {Richards et al. 2002}. Methods of studying human evolution and migrationThe lineage-based approach and the population-based approach are two methods that use mtDNA to study human evolution. The former is used to identify the origins of different lineages of mtDNA, called haplogroups, while the latter studies the history of populations, regions or migrations. This makes use of groups of different human populations and applying population genetic methods (Forster, Torroni, Renfrew & Röhl 2001}. One shortcoming of the lineage-based approach is that it only provides information on the history of the haplogroups themselves, and not of the populations they are found in. A haplogroup can be much older then the population it is found in, and the spread of a haplogroup need not be a separate migration. For example, a population contains multiple haplogroups. If the population had to migrate, the migration of its haplogroups would have occurred at the same time. The age of a haplogroup is the time since its characteristic mutations came about, and not when the population containing it migrated {Simoni et al. 2000}. When studying a population's genetic history, it is important to utilize statistics that evaluate population relationships. It is important to analyze both haplogroup affiliations and population affinities using a population genetics approach {Richards et al. 2002}. Despite the resolving power of the HV1 sequence is not high enough to reveal all the different haplogroups, its high mutation rate is enough for it to undergo population genetic analyses {Pakendorf & Stoneking 2005}. Sites analysed for the determination of phylogenetic relationshipsThe most commonly analysed site of the mtDNA is the HV1 region. This region is the only gives a low resolution. This means that if the HV1 regions of the two lineages match, they have a 50% chance of having a common ancestor within 52 generations. If a higher resolution is needed, the HV2 region can also be analysed. In a high resolution, if the HV1 and HV2 regions of the two lineages match, they have a 50% chance of sharing a common ancestor within 28 generations. However, it is becoming more common to analyse the whole mitochondrial genome. One reason is that phylogenetic trees based on entire genomes have higher resolution then those based on HV1 regions alone (Ingman & Gyllensten 1982). However, it is uncertain whether the additional cost and effort is worthwhile (Brandstatter, Parsons & Parson 2003). To date, whole genome analysis has not yet revealed a detail in human history that was not already revealed by HV1 sequence studies. Nevertheless, it is claimed that HV1 studies produce lower genetic distance between populations then the coding regions. A compromise is to combine HV1 analysis with partial coding region sequencing (Wilder et al. 2004). Different haplogroupsSince mtDNA lacks recombination and is uniparentally inherited, different mtDNA sequences can only arise from the accumulation of mutations. In the course of time, these molecular divergences between different mtDNA molecules give rise to different haplogroups. A haplogroup is a group of related genetic sequences, that all share common mutations at the same site. In order to classify a mtDNA molecule into a haplogroup, its sequence must be compared to the Cambridge reference sequence. Any mutations must be noted. The mutations are then compared to typical mutations of different haplogroups. The identification and naming of haplogroups is unorderly, resulting in para and polyphyletic groups being classified in the same group, or monophyletic groups being classified into different groups (Tanaka 2004). African human originNowadays, it is commonly accepted that modern humans originated from Africa, and spent most of their existence there. The migration routes of modern humans remain unanswered due to the lack of archeological evidence. This makes it important to track maternal lineages of human populations in Africa, using mtDNA. Fundamental routes of migration can be identified using the information obtained from sampling of recent mtDNA. Fundamental routes of migration can be identified using the information obtained from sampling of recent mtDNA. Individuals within a haplogroup are commonly located at particular geographic locations since these differences arise during and after colonization processes. For example, the macrohaplogroup L is only found in Africa {Salas et al. 2002}. mtDNA analysis has given evidence for the recent African origin hypothesis of modern humans. mtDNA analysis reveals that the most recent common ancestor of human mtDNA, popularly known as the mitochondrial eve, lived in Africa between 100, 000 to 200, 000 years before present (ybp). The mitochondrial eve is thought to have been a member of haplogroup L (Vigilant et al. 1991). MtDNA PhylogeographyAfricaThere are two branches of human mtDNA. These are the L0 and the L1’2’3’4’5’6. The former branch is common all over Africa and is ancestral to all other modern haplogroups around the world. The mitochondrial gene pool in Africa is much more diverse then the rest of the world, since humans had undergone about 150, 000 years of dispersal until 18, 000 ybp, when the first permanent settlements started to occur {Behar et al. 2008}. Haplogroup L0 was the first to spread about 180, 000 ybp, from East Africa to South Africa. The second major migration occured about 70, 000 ybp from East Africa to the West, and also towards the Arabian region and the rest of the world. Haplogroups L1’2’3’4’5’6 were all involved. The most recent widespread dispersal in Africa occurred 3 to 4, 000 ybp, and is known as the Bantu dispersal, which originated from Western Africa and spread Eastwards and Southwards. As a result, the present populations in East and South Africa have haplogroups characteristic of the original population, as well as haplogroups characteristic of Western Africa {Behar et al. 2008}. Despite some occasional mixing, different regions are populated with their own characteristic haplogroup. For example, North Africa is populated by individuals with M1 and U6 haplogroups. Both of these are not African, but are subclades of the M and N macrohaplogroups respectively. This arises due to back migrations from the Arabian region. Another example is the accumulation of the L1c haplogroup in Central Africa. L0d and L0k are only present in South Africa. L0a is common from Northeastern to Southeastern Africa. This implies a possible migration {Behar et al. 2008}. AsiaThe L3 haplogroup emigrated out of Africa to the Arabian region about 60 to 70, 000 ybp. M and N are two clades that diverged from the L3 haplogroup about 65, 000 ybp and colonized the rest of the world (Macaulay et al. 2005). Haplogroup M only emigrated Eastwards by travelling along the Indian coastal line. In contrast, N emigrated to Europe and Asia, gave rise to haplogroup R. Today, haplogroup N is found in Western and Eastern Eurasia. Subhaplogroups descended from N have their own distinct regions of prevalence. The present day haplogroup population in Asia is the result of a rapid migration from the middle East along the Southern Indian coast. This migration route is called the Southern Coastal Route (SCR), and is also evidenced in anthropology and archaeology (Macaulay et al. 2005). The haplogroups M, N and R all are approximately 60 to 63, 000 years old. This indicates that they all were involved in a colonization (Macaulay et al. 2005). All three haplogroups are present along the SCR. This includes the Near East, South Asia, South East Asia and Australia. This shows that the colonization event was rapid (about 10, 000 years), since all three of the founder haplogroups reached all destinations. If the migration was slower, subhaplogroups would have had time to evolve during the migration event, and they would have been found in the next area to be populated. However, only a fraction of the migrated group remained in the coastal area. About 40, 000 ybp, the rest of the group dispersed inland and developed their own haplogroups distict to that region (Metspalu et al. 2006). An alternative route taken during migration was the Northern Asian Route (NAR). This passed by the Himalayas and into Central Asia. However, this route is unsupported by genetic evidence, since no subhaplogroups from the Coastal area are found in this region {Derenko et al. 2007}. Also the climate at those time periods was to cold to allow such a migrational route to be possible (Tanaka et al. 2004). Central Asian haplogroups share common ancestry with Souther Asian haplogroups. Additionally, the Southern Asian haplogroups are more genetically diverse. This most likely occurred by Central Asia being influenced both from the East and from the West about 20, 000 years after the Southern coast was colonized {Yao et al. 2003}. AmericaMigrations from Eastern Asia eventually led to America via the Beringia land Bridge. America became populated about 25, 000 ybp with subhaplogroups that were previously developed during a delay period. Native American haplogroups include the A, B, C and D. A and B are subclades of N, while C and D are subclades of M. All four haplogroups are also found in Eastern Asia. Haplogroup X is another N subhaplogroup that colonized America, which migrated from Northern Europe. It is most common in the North America. 15 founder lineages in America have been identified to date. This make up the A1, B1, C1 and D1 pan-American haplogroups, along with other poorly dispersed lineages. The spread in America occurred south wards, probably along the Pacific coast. Recently, the genetic structure of America was influenced by European settlers or slave imports {Tamm et al. 2007}. AustraliaAustralia was populated through migrations from South Asia. It supports the view that there was only a single migration out of Africa. It is believed that the aborigine people are the first people to colonize Australia (Rasmussen et al. 2011). Common haplogroups found in Australia include haplogroup N and M. Australia also includes various subclades of N, such as

N12, N13 and N14, P and S. Humans with haplogroup M and N emigrated to Australia, by starting from the horn of Africa, and taking the coastal path. This occurred in a single migrational event about 50, 000 ybp {Hudjashov et al. 2007}. EuropeEurope has much less mtDNA diversity then Asia and Africa, and even the most isolated groups have a very short genetic distance. The European population structure hints about multiple bottlenecks that took place during the colonization process. European mtDNA haplogroups and their subclades are more evenly spread throughout the continent. Thus, they are not easily linked with any particular ethnicities. This is only the case on an intercontinental scale. However, they are sometimes useful as indicators of some medical conditions. The only exception is with the Saami from Northern Scandinavia. There are six regions of genetic boundaries in Europe. The Saami show the sharpest boundary, followed by the Near Easterns, Catalans, Belgians, Norwegians, and the Ladins. The only cline found in Europe is in the region around the Mediterranean. No cline is found in the North, provided that the Saami were excluded (Simoni et al. 2000). All European mtDNA haplogroups descended from N. N is currently only found in low frequencies in Eurasia. Typical European haplogroups include H, J, I, K, U, T, V, X and W (Lell et al. 2000). Subclades of these haplogroups are more associated to a particular area, but not with ethnic groups. One explanation is that women tended to marry men of a different ethnic group to maintain peace between different cultures. Ethnical associations only occur on a continental scale. The only exception is with the Saami. Ancient European mtDNA shows that early Europeans belonged to HV, H, V or U5a. Nowadays, these are the most common

European haplogroups. More then half of the European matrilineal lineages are descended from these Palaeolithic or Mesolithic Europeans (simony et al. 2000). Europe was colonized from the Near East in four migrational events. Two of these events occurred before the last ice age. The first migration occurred about 45, 000 ybp, when members of haplogroup U5, a descendent of N, migrated from the Middle East to Caucasus and eventually to the Iberian peninsula. About 7% of modern Europeans are descended from these people. The second migration took place about 26, 000 ybp, also from the Middle East, and brought haplogroups HV, U1, U2 and U4 (25%). The LGM took place about 20, 000 ybp and was responsible for the depopulation of Northern Europe. However, members of haplogroups HV, H, V and U5 survived in isolated refugia such as the Franco-Cantabrian refuge in the Iberian peninsula and another refuge near what is now Italy. These haplogroups in the refugia helped recolonize Northern Europe after the LGM{Lell et al. 2000}. The migration did not only occur northwards, but also southwards, and as a result, there is a relatively recent maternal link of a mere 9, 000 years between the Saami from Northern Scandinavia and Berber people from Northern Africa. They share many haplogroups such as H1, H3, V and U subclades.(Achilli et al. 2005) Another migration occurred about 14, 500 ybp, when members of haplogroups K, T, W and X emigrated from the Near East (36%). The last migration occurred after the LGM, about 9, 000 ybp, and brought haplogroups J and T and others to the continent (23%) {Lell et al. 2000}. European HaplogroupsHaplogroup H is the most common European haplogroup. It has a very wide distribution range and has very high frequencies in those ranges. In general, about 50% of Europeans belong in haplogroup H. It has origins in the Middle East about 25 ybp (Richards et al. 2002). H1 and H3 are the most frequent H subclades in South-West Europe. H1 is most common in the Ibarian peninsula and the surrounding areas. H3 is also most common in the Ibarian peninsula, but has much lower frequencies then H1. H1 and H3 frequencies both peak at the Basques in Spain, at 27. 8% and 13. 9% respectively (Achilli et al. 2004). Haplogroup V is the sister group of haplogroup H. They are both subclades of haplogroup HV. It is a derivative of haplogroup pre*V (Torroni et al. 2001). Unlike haplogroup H however, haplogroup V originated in Europe in the Franco-Cantabarian refuge soon after the LGM. In contrast, haplogroup pre*V seems to have existed before the LGM (Torroni et al. 2001). The distribution of haplogroup V mirrors that of haplogroup H1 and H3, but has a very high frequency of about 40% among the Saami people, but is less frequent in Central, and Eastern Europe, and is virtually absent in South Eastern Europe. Along with U5a and H, haplogroup V was the predominant haplogroup in South Western European hunter-gatherers in the last Ice Age. The similar trends of H1, H3 and V implies that they were part of the same migration when hunter-gatherers repopulated Central and Northern Europe about 15, 000 years ago, after the LGM {Torroni et al. 1998}. Haplogroup U is the oldest European haplogroup, with an estimated age of 50, 000 years. It is the only haplogroup that is common with Africans {Torroni et al. 1996}. It is estimated that 7% of Europeans are members of haplogroup U. There are 8 main subclades of U, named U1-8. U1 and U2 are mainly found in the Near East and Mediterranean {Macualay et al. 1999}. U5 is a very old subclade at 50, 000 years, and is only specific to Europe. It was the only cluster that arrived in

Europe in the first migration that occurred in the Upper Paleolithic. It makes up about 7% of Western European U subclades. U5a, dating to about 35, 000 years ago, is mainly found in Southern Europe. U5b is more common in central and Western Europe. Its U5b1 subclade is specific to the Saami {Richards et al. 1998}. Like U5, U4 expanded in Europe prior to the LGM (Corte-Real et al. 1996). U6 is the only subclade that is not found in Europe, but is specific to Berbers in Africa (Rando et al. 1998). This implies a back migration from the Near East. U6 is the sister clade of U5. Their similar ages hints that they diverged from a common ancestor (Di Rienzo & Wilson 1991). U8 is a large subclade of U. One of its subclades is haplogroup K, which is the sister group of subclade U8b. It covers the ranges of U8a and U8b, and is also found in India (Metspalu et al. 2004). In Europe, it is found in particularly high frequencies Western France (15. 3-17. 5%), but is also common in Norway and Bulgaria (13. 3%). However, the average European frequency is about 5. 6% (Debut et al. 2004, Simoni et al 2000). It originated approximately 16, 000 ybp in the Middle or Near East (Metspalu et al. 2004). Despite its recent origins, it has the highest numbers of subclades, the largest being K1a. 32% of Ashkenazi Jews belong to either K1a1b1a, K1a9, and K2a2a, all of which are K subclades. This contrasts the condition found in non Jewish Europeans and Middle Eastern Jews (6%). This implies that a bottleneck took place about 100 generations ago, followed by rapid population growth (Behar et al. 2004). Haplogroup J has the highest frequencies in the Near East (12%), particularly in the Araian peninsula (Richards et al. 2000). However, its frequency declines in Europe to about 11%, and is evenly spread. Also, its diversity in the Near East is higher

(Richards et al. 2000). This implies that haplogroup J also colonized Europe from the Near East. Its spread is also associated with the spread of agriculture (Ammerman & Cavalli-Sforza 1984). Haplogroup J is divided into J1 and J2. The former is further divided into J1a, J1b and J1c subclades. J1a is equally spread in Europe and the Near East, while J1b and J1c are more diverse in the Near East. The J2 subclade is relatively scarce when compared to J1. J2a and J2b are two subclades of J2. J2a is spread homogenously in Europe. J2b is divided in two more subclades, being J2b1 and J2b2. J2b1 is only found in Europe, while the latter is found in both Europe and the Near East. A bottleneck could have occurred during the peopling of Scandinavia, Albenia and Caucasus, since much less diversity is observed (Serk 2004). Haplogroup T is found throughout Europe, having a percentage frequency of about 5. 5%. It peaks at 10. 89% in France and Italy, but is absent in the Saami (Helgason et al. 2001). Haplogroup T is the sister clade of haplogroup J. They are both subclades of the haplogroup JT. Similarly to J, it also originated from the Middle East about 46, 500 ybp, but nowadays, it is more common in Europe (Richards et al. 1998). There are five T subclades, named T1-5, but the most common are T1 and T2. T1 by far the more common one, and is found especially in Bulgaria and Turkey. It is also common in European Russia and some Scottish islands. T2 is relatively scarce when compared to T1. It is mainly found in Iceland, but also in European Russia. It is much scarcer in other parts of Europe (Helgason et al. 2001). Haplogroup W is found in low frequencies in most Europe, with highest occurrences in Ukraine, European Russia, the Baltic and Finland. Haplogroup I has a similar distribution, but also has a higher presence around the Caspian Sea. It most likely origin in in the Proto-Indo-European cultures. It is absent in Western Europe, which is furthest from the Caspian Sea, but is more common in Eastern Europe. Haplogroup X is found all over Eurasia, but rarely exceeds 5% of the total population. It is a direct subclade of haplogroup N. Its European haplogroups are X2a, X2c, X2d and X2e, and are of Indo-European origin. It is found in North Africa, West and Central Asia, the Americas and Europe, but is totally absent in Siberia and East Asia (Reidla et al. 2003). On average, there are eight nucleotide differences between individuals amongst unrelated Caucasians and 15 nucleotide differences between individuals amongst unrelated Africans (Budowle et al. 1999, Melton et al. 2001). MtDNA haplogroup geographical variation also implies that there is selection on some haplogroups as humans populate areas with different climates. For example, some claim that the haplogroups A, C and D, which are found in Siberia and Native Americans, were selected to separate the production of heat and ATP from each other during oxidative phosphorylation. This would produce more heat, which is advantageous in a colder environment (Ruiz-Pesini et al. 2004). However, statistical and biochemical analyses are needed to determine whether there indeed is a change in the oxidative phosphorylation process in these haplogroups. However A, C and D also occur in tropical environments like the Americas, where they have been for at least 10, 000 years (Schurr et al. 2004). On the other hand, another study does not acknowledge any evidence for selection due to climate change. There is evidence of negative selection on 12 of the 13 genes coding for proteins. Most agree that human mtDNA is subject to purifying selection. Since the whole genome is inherited, and since recombination does not occur, the control region may also experience selection (Cavalli-Sforza &, Feldman et al. 2003). Problems associated with using mtDNA for studying historical population geneticsSome issues may jeopardize this method of study of historical population genetics. Firstly, it is now known that nuclear insertions of mtDNA (numts) are more common then previously expected (Bensasson, Feldman & Petrov 2003). There may be between 250 up to more then 600 insertions, with various lengths, the longest being nearly the whole mitochondrial genome (Tourmen et al. 2002). Since numts are part of the nuclear DNA, they mutate at much slower rates then mtDNA, and serve as molecular fossils (Zischler, Geisert, von Haeseler & Paabo 1995). However, they are hard to detect and are sometimes mistaken as mtDNA. The latter, along with their slower mutation rate, give rise to inaccurate phylogenetic conclusions (Olson et al. 2002). Another problem is an increase in the amount of publications and submittions of incorrect mtDNA sequences to databases (Rohl et al. 2001). Despite these errors, the conclusions taken are rarely affected by them. Nevertheless, the amount of errors should be kept to a minimum, and statistical methods should be employed frequently to detect sequencing artefacts. Statistical analysis alone is not sufficient, and could never replace good laboratory practice. Another issue is the debatable neutrality of mtDNA. In humans, the rate of nonsynonymous mutations is greater then that of synonymous mutations on several sites in the mtDNA. This is not true between humans and chimps (Nachman et al. 1996). This could be explained by most mutations being slightly deleterious, such that they would give rise to polymorphism without getting fixed. This would not contribute to interspecific differences (Hasegawa, Cao, Yang 1998). The surplus polymorphisms within a species indicate the that the intraspecific divergence is being underestimated since some mutations that occurred since our divergence from chimps have not been detected Gerber, Loggins, Kumar, Dowling 2001). One drawback is that mtDNA is a single locus and only reflects the population’s maternal history. The history of that locus may be a different one then that of the entire population due to the effects of drifting and of possible selection on that locus. This makes it imperative to compare mtDNA analysis with Y-chromosome and autosomal analysis. In contrast with mtDNA, the Y-chromosome is only paternally inherited (Bamshad 2003). Aim of this studyAlthough there is a substantial amount of published literature regarding the phylogeography of different European haplogroups, no literature was found about the different haplogroups found in Malta. The aim of this study is to assess the variability of the Maltese population of HV1 and HV2 regions by taking into account: What haplogroups are present in Malta and in what ratiosHow many different haplotypes can be found in 100 individualsThe amount of SNPs and what trnsversions take place