1 Introduction

Genetic research can be used to shed light on various aspects of the human species. By analyzing DNA variation between and within modern populations it is possible to make inferences about the genetic history and interaction of their ancestors, evolutionary processes, past demography and, if phenotypic information is also available, the genetic variants underlying these traits.

This chapter provides an introduction to the ideas and concepts discussed in this thesis. The first part describes the different types of variation observed in the human genome, ranging from single base changes to changes involving many kilobases. Genetic diversity has been shaped both by various demographic processes—such as population size, population structure and migrations—and by the forces of natural selection, as the human species became adapted to new environments and challenges. One of the most important tasks in population genetics is to distinguish between these demographic and selective signals. This is discussed in the second part where I describe what effects the different processes are expected to have on our genome. When these have been established it is possible to start looking for evidence of natural selection. In part three I will introduce tests used to identify candidate loci for selection and give examples of selection signatures that one should be looking for; such as a reduced variability, increased levels of population differentiation, increased linkage disequilibrium and skewed allele frequency spectra. The fourth part then gives a brief introduction to the evolution of modern humans, tracing their journey from their 'cradle' in Africa into the rest of the world, where they had to adapt to new conditions. For the special consideration of this thesis, the fifth part introduces the idea of gene loss as a process of evolutionary change and gives examples of genes whose loss has been advantageous for humans. Lastly, the sixth and final part describes the aims of this thesis.

1.1 Variation in the Human Genome

The human genome is made up of around 6 billion nucleotides stored on 23 chromosome pairs, one set inherited from each parent. Between two randomly-chosen human DNA sequences there will be several different types of variation occurring on different scales, ranging from single base changes to alterations of the copy number of larger segments. These will include single nucleotide polymorphisms (SNPs), insertion/deletion polymorphisms (indels), retroposon insertions, variations in the number of copies of a tandem repeat, copy number variants (CNVs), inversions and variants that may cut across these categories.

The Human Genome Project (The International Human Genome Mapping Consortium 2001; The International Human Genome Sequencing Consortium 2004) emphasised that the human genome was about 99.9% identical in all people (Sachidanandam et al. 2001). But more recent efforts such as the Haplotype Map of the Human Genome (Frazer et al. 2007; The International HapMap Consortium 2005), the CNV project (Redon et al. 2006; Stranger et al. 2007a) and the recently published sequences of two diploid genomes (Levy et al. 2007; Wheeler et al. 2008) have revealed a more complex picture. It is now clear that human genetic variation was underestimated and is much greater than the 0.1% difference found in earlier genome sequencing projects. In fact, when you take CNVs into account genetic variation is estimated to be at least 0.5% (99.5% similarity) or five times higher than the previous estimate (Levy et al. 2007).

1.1.1 SNP Variation

SNPs are the simplest and most common type of variation in the human genome and involve the exchange of one base for another. SNPs have been estimated to constitute roughly 75% of the total number of variants observed in the human genome (Levy et al. 2007).

Only about 1.5% of the genome encodes proteins, but this small proportion is of disproportionate importance for biology in general and this project in particular. The genetic code is read in triplets but is redundant as many amino acids are encoded by more than one codon. This redundancy is a consequence of the difference in number between the 64 possible triplets and 20 amino acids, and might also work as a defence against the deleterious effects of base substitutions occurring within an open reading frame (ORF). Synonymous-SNPs are base substitutions that do not alter an amino acid and are therefore often assumed to be selectively neutral. On the other hand, nonsynonymous-SNPs are base substitutions that lead to a change of amino acid and could potentially alter the function of the protein. There are two types of nonsynonymous mutations; missense mutations occur when an amino acid is changed into another amino acid and nonsense mutations occur when the substitution changes an amino acid codon into a termination codon (UAA, UAG or UGA after transcription to RNA).

Traditionally, these single base substitutions are said to be polymorphic when alleles are found at a frequency between 1% and 99% in the human population. The number of SNPs in the human genome has been estimated at more than 10 million. Thereof, about 7 million are designated as "common" SNPs with a minor allele frequency (MAF) of at least 5% across the entire human population (Crawford et al. 2005; Kruglyak and Nickerson 2001).

Therefore, any two unrelated humans are likely to have millions of such genetic differences between them. The average proportion of nucleotide differences (i.e. average nucleotide diversity, π) between two randomly chosen human chromosomes has been estimated at around 7×10^{-4}, meaning that on average you expect to see one SNP for every 1,430 base pairs (bp) (Altshuler et al. 2000; Sachidanandam et al. 2001; Schneider et al. 2003; Zhao et al. 2000). This difference is small compared to other species. For example, our closest living relative the chimpanzee (*Pan troglodytes*), occupies a much smaller geographic range and has a smaller population size, yet its nucleotide diversity is about 1.5 times higher than in humans (Fischer et al. 2004; Yu et al. 2003).

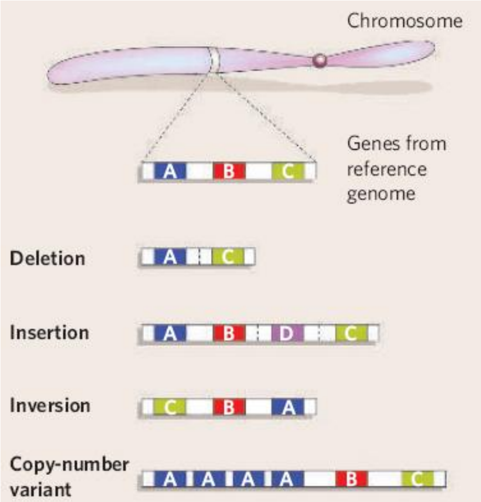

Figure 1 Different types of variation. A section of the reference sequence is given at the top, followed by examples of the way different types of variation could change that sequence. This figure is adapted from (Check 2005).

1.1.2 Other Forms of Variation

While SNPs are the most common form of variation in the genome, other types of variation (Figure 1) are worth noting as their importance in human evolution and susceptibility to disease is becoming ever more apparent. While these have been estimated to constitute about 22% of the total variation observed in the human genome, together they will affect a larger number of bases than SNPs (Levy et al. 2007).

It is not within the scope of this thesis to discuss all the different types of variation, but a few will be described briefly. Small insertion/deletion polymorphisms (together referred to as indels) are mutations which involve the insertion or deletion of a DNA sequence, either on the single base level or on a larger segment of DNA but conventionally less than one kilobase (kb). If the indel occurs within an ORF and is a multiple of three bases, it will lead to an insertion or deletion of one or more amino acids. A frameshift mutation occurs when the indel is not a multiple of three. It will cause all the codons occurring after the deletion or insertion to be read incorrectly during translation and thereby changes the reading frame. In this case the translation will keep going until a termination codon is reached, which will either lead to a prematurely terminated protein or an extended version of the protein (Jobling et al. 2003; Strachan and Read 2004). Inversions are segments of DNA that are reversed in orientation with respect to the reference sequence. They can affect almost any length of DNA, but are among the most difficult to study with the techniques available and thus the least-well characterised.

Variable number of tandem repeats (VNTRs) occur when a nucleotide sequence is organized as a tandem repeat and can be found at variable lengths between individuals. There are two main categories of VNTRs, microsatellites and minisatellites. The former refer to repeats of units less than roughly five base pairs in length while the latter involve longer blocks. While VNTRs are abundant in normal individuals, some have been associated with a number of genetic disorders in humans, collectively called nucleotide repeat expansion diseases (reviewed in Usdin 2008).

Retrotransposons are mobile repetitive DNA elements that have the ability to make an RNA copy of themselves which is then reverse-transcribed and inserted into a new location in the genome. The most famous of these, the *Alu* and LINE1 element insertion polymorphisms, have been used extensively to answer questions about human evolution (Jobling et al. 2003).

A CNV is a segment of DNA that is one kb or larger and is present at a variable copy number. It can be in the form of an insertion, deletion or duplication and will therefore involve gains or losses of one to several hundreds of kb of genomic DNA. Nothing is implied about their frequency but those occurring at 1% or more (as is traditional with SNPs) have been referred to as copy number polymorphisms (Feuk et al. 2006). CNVs can be neutral or involved in developmental disorders and susceptibility to disease (Inoue and Lupski 2002). It has been estimated that 12% of the human genome is subject to CNV (Redon et al. 2006) and that approximately 0.4% of the genomes of unrelated people will typically differ with respect to copy number (Redon et al. 2006), but these estimates are uncertain because the techniques used are not able to detect small (<50kb) CNVs or measure the sizes of those detected accurately.

1.1.3 The Good, the Bad and the Neutral — Consequences of Variation

Most of the genetic variation observed between individuals and populations is assumed to be neutral (Bamshad and Wooding 2003; Kimura 1983), having no obvious effect on the phenotype. However, our genome contains some variants that are being selected, either for or against, and these are of particular interest to those of us trying to decipher the forces behind human evolution.

The consequence of a mutation will first and foremost depend on its location within the genome. For example, mutations located outside a gene can affect its expression by altering promoters or enhancers, while a mutation within an intron can affect splicing or the regulation of an adjacent gene. For those mutations occurring within the ORF, the consequences can range from no effect to the complete loss of the protein product. These consequences will depend on the type of mutation and the position within the gene. Although one might generally expect that "the larger the mutation, the bigger the effect", this is not always the case, as even a single base substitution within a gene can cause a genetic disease, while changes in large segments might not have any detectable effect. But deleting large bits of DNA can result in the loss of important genes and having extra copies of a gene can cause unwanted overproduction of a protein.

The widespread existence of CNVs in the genomes of apparently healthy individuals (Iafrate et al. 2004; Sebat et al. 2004) was initially a big surprise to many researchers, as such large changes had previously been mainly associated with diseases. For example, a duplication of a 1.5 mega base (Mb) region from chromosome 17 has been associated with Charcot-Marie-Tooth disease type 1A (*CMTIA*) (King et al. 1998; Lupski et al. 1991) while a deletion of the region will lead to hereditary neuropathy with liability to pressure palsies (*HNPP*) (Chance et al. 1993). Many more such examples are likely to be discovered, as the investigation of the contribution of CNVs to complex traits and common human disorders has really only just started. This field has, up until now, mostly relied on the more easily typable SNPs, where associations have been reported with type 2 diabetes (Helgason et al. 2007; Sandhu et al. 2007; Sladek et al. 2007), breast cancer (Beeghly-Fadiel et al. 2008; Easton et al. 2007; Stacey et al. 2007) and coronary heart disease (Helgadottir et al. 2007; Ozaki et al. 2002), to name a few examples. Indeed, the Wellcome Trust Case Control Consortium (WTCCC) is a collaboration aimed at analysing hundreds of thousands of SNPs in thousands of DNA samples from patients suffering from different diseases to identify common genetic variation for each condition. Additionally, Icelandic women who carry a 900 kb inversion on chromosome 17 have been shown to have more children than those who don't. This inversion is found in 20% Europeans and, because of its selective advantage in child-bearing abilities, has spread through the population (Stefansson et al. 2005).

I have shown that most genetic variation is assumed to be neutral and that some variation is bad in its association with human diseases. Good variation— variation that has become advantageous for its carriers—is perhaps not as easily established, but some examples exist and these will be discussed in sections 1.3 and 1.5.

1.2 Processes Shaping Diversity

Modern human genetic diversity has been shaped by internal forces, such as recombination and mutation, as well as extrinsic events, like migration, gene flow, genetic drift and selection.

The neutral theory (Kimura 1983) holds that polymorphisms are generally neutral rather than affecting fitness, and deviations from this model have been taken as possible evidence for positive or other selection. Natural selection is, however, only one possible explanation out of many for a rejection of a simple neutral model. Demographic processes such as population bottlenecks, founder effects, migration and admixture can also influence sequence variation in human populations. However, while demographic processes affect the entire genome, natural selection leaves its signature at specific sites in the genome. Therefore, in order to make any judgment about a variant, it is essential to understand the processes shaping its diversity.

1.2.1 Recombination and Linkage Disequilibrium

The patterns of genetic variation observed in a sample of unrelated individuals are the product of many mutation and recombination events that have occurred over many generations. Recombination refers to the crossover (i.e. breaking up and exchange) of DNA segments between members of a chromosomal pair and occurs usually during meiosis. In this sense, recombination can be seen as a reciprocal process. Non-reciprocal transfer of genetic information (gene conversion) also occurs, but is much less studied by population geneticists. While only a few recombination events occur within a single meiosis (roughly one per chromosome arm), the ancestral history of the human population spans many thousand meioses, so any sizable region of the human genome is likely to have undergone several recombination events (reviewed in Hellenthal and Stephens 2006).

There is evidence for substantial variation in recombination rates across the genome, both at gross and fine scales (Crawford et al. 2004; McVean et al. 2004). Recombinations are often frequent near telomeres and rare near centromeres; at a finer scale recombination events seem to be concentrated into small regions and consequently 25,000 hotspots (i.e. small (~1 kb) regions with highly elevated rates of recombination separated by stretches of several kb with little recombination) have been identified in the human genome (Myers et al. 2005).

A haplotype is a combination of alleles at multiple loci that are inherited together on the same chromosomal region. Related to this, linkage disequilibrium (LD) is the extent of non-random association of alleles at neighbouring sites along chromosomes due to their tendency to be coinherited because of reduced recombination between them. Recombination will tend to reduce LD in the population. As a result, patterns of LD in the human genome are characterized by the amount of haplotype diversity in so-called LD blocks which are interspersed by apparent 'hot spots' of recombination. Thus, the expected amount of LD between markers depends on the recombination rate between them (Pritchard and Przeworski 2001) and the history of the population. For example, LD is generally lower in African populations than non-African populations (Jakobsson et al. 2008).

1.2.2 The neutral theory

Mutations can be roughly assigned to three categories: advantageous mutations that increase the individual's evolutionary fitness, deleterious mutations which decrease the individual's evolutionary fitness and are therefore eliminated, and neutral mutations that do not have any effect on evolutionary fitness.

While, as we have seen, some variation may undoubtedly have functional consequences, it has been widely accepted for many decades that most variation is neutral with respect to evolutionary fitness (Bamshad and Wooding 2003). The neutral theory of molecular evolution, as proposed by Kimura (1968; 1983), has served as a null hypothesis for researchers searching for selection (see further discussion in section 1.3). The neutral theory assumes that polymorphisms are either eliminated or become fixed in a population as a consequence of the stochastic effects of random genetic drift rather than natural selection. With the incorporation of additional simplifying assumptions, such as constant population size, a randomly mating population, with no migration and non-overlapping generations, the standard neutral, or 'Fisher-Wright', model can make quantitative predictions about many genetic properties, such as the level of variation expected at a locus in a population (Jobling et al. 2003). Therefore, despite relying on assumptions that clearly do not hold in human populations, the neutral model can serve as a null hypothesis from which departures of the data can be detected (Przeworski et al. 2000). Deviations from the neutral model can then be investigated as possible cases of selection (see section 1.3).

It should however be noted that natural selection is only one possible explanation out of many for a rejection of the neutral model. Demographic processes such as population bottlenecks, founder effects, migration and admixture can also influence sequence variation in human populations (Przeworski et al. 2000). Some demographic processes can mimic the signals of selection and could easily be mistaken for actual selection. A frequently quoted example is that both population expansion and positive selection lead to an excess of rare variants, reflected in negative values of the test statistic Tajima's D (described in section 1.3.1.2). Indeed, one of the greatest challenges in population genetics is to be able to distinguish between selection and demography in order to correctly infer the forces acting on a specific region. To this end, knowledge and understanding of the population history of humans is essential as the power of statistical tests to reject neutrality can be increased by appropriate modelling of the demographic parameters.

1.2.3 Demographic Processes

1.2.3.1 Population Structure

The most comprehensive summary of the origin and dispersal of human populations is *The history and geography of human genes* (Cavalli-Sforza et al. 1994). This extensive review, based on serogenetic loci, demonstrated unequivocally that the human gene pool is geographically structured—in other words, that people and their alleles are not randomly distributed over the surface of the Earth. There are many factors, both cultural and geographical, that have shaped the genetic relationships of human populations and these will have an effect on the genetic patterns observed in modern humans.

Furthermore, recent studies of large-scale variation data have shown that while humans are genetically similar, it is still possible to use the small differences to distinguish between the major geographical regions (Jakobsson et al. 2008; Rosenberg et al. 2002) and, on a finer scale, it is even possible to assign individuals to their likely sub-population of origin (Lao et al. 2008; Novembre et al. 2008).

1.2.3.2 Population Size and Bottleneck Events

Constant population size is one major assumption of the standard neutral theory. However, it is evident that the human population has not maintained a constant size, but has rather changed dramatically over the past 100,000 years. Most genetic studies (Takahata et al. 1995; Tenesa et al. 2007; Zhao et al. 2000) reflect this, estimating an 'effective population size' (the size of a Wright-Fisher population experiencing the same amount of drift) of around 10,000, which contrasts with the current census size of more than 6 billion. As large-scale data are becoming increasingly available, recent studies have been able to perform simulations (Hudson 2002; Schaffner et al. 2005) with some underlying demographic parameters defined to identify the best-fit model to explain the data.

With a sudden reduction in size, the whole genetic composition of a population can be changed. Such an event is referred to as a bottleneck. A bottleneck can be caused by the death of a large number of the population members by natural disasters, famine and/or disease or by the outward migration of people that do not return to reproduce. The result is that the genetic variation decreases and the smaller population becomes more prone to the effects of genetic drift, discussed in the next section.

1.2.3.3 Genetic Drift

Genetic drift is a concept introduced by Sewall Wright (1931) and refers to the random change in allele frequencies from one generation to the next. Some alleles will become common and others rare, and as time passes the end result will be that one allele becomes fixed (frequency = 1) at the expense of the other, which is eliminated (frequency = 0). With chance at play, genetic drift is distinct from natural selection where the alleles would increase/decrease in frequency in response to selection. In the case of genetic drift, the fate of the allele will depend on several factors, such as the size of the population and the initial allele frequency. On average, alleles drift to fixation or elimination faster in smaller populations than in larger populations (see Figure 2). This is due to a statistical effect of sampling error during the random sampling of the alleles from the overall population.

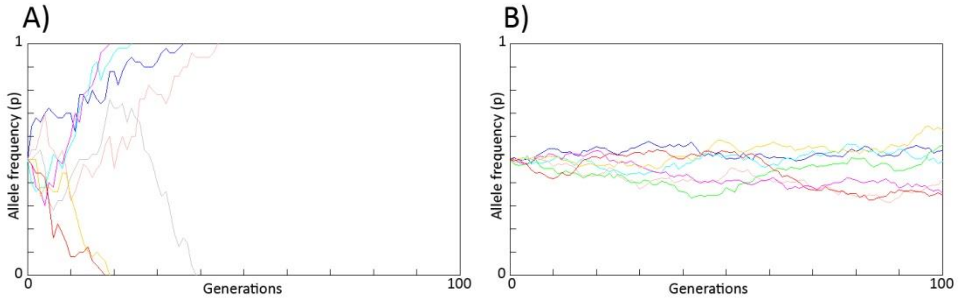

Figure 2 Simulations of genetic drift for an allele starting at a frequency of 0.5 over 100 generations. A) Population size = 25, the allele gets rapidly lost or fixed **B)** Population size = 1000, the allele frequency changes are more subtle. The simulation was created with an online simulator available from http://darwin.eeb.uconn.edu/simulations/drift.html.

1.2.3.4 Migration Events

The migration history of populations has influenced human genetic diversity and is therefore important to consider. In the case of colonization, a small group of people from a larger ancestral population may have moved into previously unoccupied land, causing the genetic diversity in the newly founded population to be reduced as it represents only a fraction of that of the parental population. This process is referred to as a founder effect, and in cases where the new population is extremely small it will continue to be sensitive to additional processes such as genetic drift after establishment.

By contrast, migration is the movement of people between occupied areas, causing alleles to be exchanged from one population to the other. If migrants successfully contribute their genetic material to the next generation in the new population then we talk about gene flow. Gene flow can lead to increased diversity within a population when new variants are introduced and can also lead to decreased diversity within two (or more) populations if migrations between them are reciprocal (Jobling et al. 2003).

1.2.4 Processes of Natural Selection

Natural selection is the process by which favourable alleles become more common in successive generations of a population while unfavourable alleles become less common, due to differential reproductive success of the genotypes. The term was defined by Charles Darwin in *The Origin of Species* (Darwin 1859) and was later elaborated by Fisher (Jobling et al. 2003).

Since the origin of our species (discussed in section 1.4), humans have had to adapt to new environments, nutritional sources, parasites and diseases, and as an adaptive response these are likely to have triggered selective forces. There are three main types of natural selection to consider: positive selection, balancing selection and negative selection (Figure 3). The nature of these will now be discussed briefly while the methods for identifying selective signals are described in chapter 1.3.

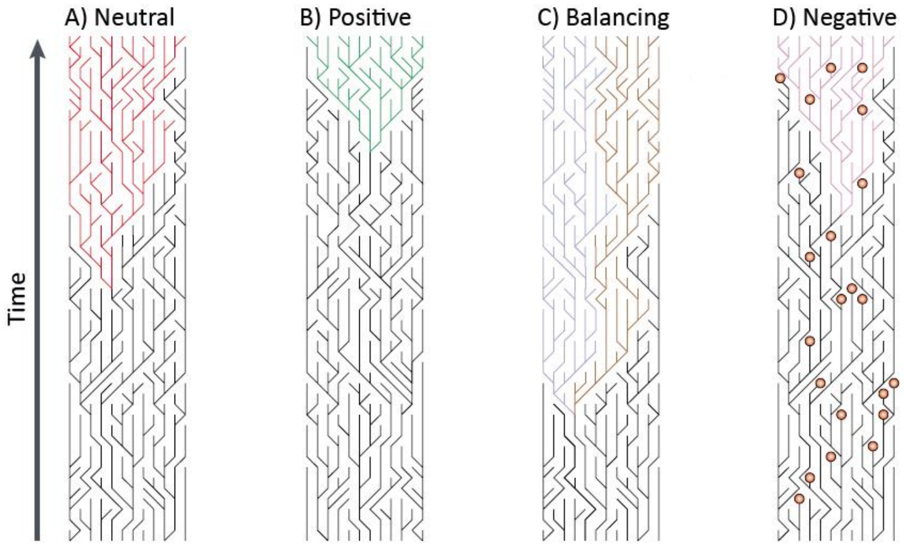

Figure 3 Effects of natural selection on gene genealogies and allele frequencies. A) The genealogy of a neutral allele (red) as it drifts to fixation. **B)** The genealogy of a positively selected allele (green) that is driven to fixation more quickly than is expected from neutrality. **C)** The genealogy of two alleles (blue and brown) under balancing selection, which are driven neither to fixation (100% frequency) nor to extinction (0% frequency). **D)** The genealogy of an allele (purple) that drifts to fixation with the elimination of a deleterious mutation (represented with circles). This figure is adapted from (Bamshad and Wooding 2003).

1.2.4.1 Positive Selection

Mutations that increase the evolutionary fitness of the carrier are likely to undergo positive selection. 'Hitchhiking' refers to the situation when neutral alleles closely linked to an advantageous allele are carried along with it in a selective sweep and reach a high frequency (Braverman et al. 1995; Fay and Wu 2000; Smith and Haigh 1974). A typical molecular signature of a newly-completed selective sweep is a reduction in genetic diversity in the region surrounding the beneficial allele. The amount of variation remaining will increase with the recombination distance from the selected allele. As new mutations accumulate after a complete sweep, there will initially be an excess of rare alleles in the swept region compared with unlinked neutral regions. This is described in more detail in section 1.3.1.

1.2.4.2 Balancing Selection

In some circumstances, selection for a beneficial allele will not lead to its fixation and alleles are thus maintained at intermediate frequencies at a locus. This is called balancing selection and it can arise because of frequency-dependent selection or heterozygous advantage. Heterozygote advantage is when the heterozygote state is more beneficial than either homozygote, as in the case of sickle cell (Hb^S) and normal (Hb^A) alleles observed at the β-hemoglobin locus in humans. Individuals homozygous for the Hb^S allele have a reduced fitness as they are inflicted with the sickle-cell disease in which red blood cells are grossly misshapen and this often results in a reduced lifespan. Heterozygotes will not suffer from the disease but have slightly irregularly shaped blood cells which protect against infection of the malaria parasite (Allison 1954; Cavalli-Sforza and Bodmer 1971). The frequency of the "disadvantageous" Hb^S allele is found at highest frequencies and at its greatest fitness in populations where malaria is endemic (Kwiatkowski 2005).

Balancing selection can also arise from frequency-dependent selection whereby the fitness of the genotype depends on its frequency in which case rare alleles may have a selective advantage and can be maintained over a long evolutionary time. A classical example of balancing selection is the amount of polymorphism observed at the human leukocyte antigen (HLA) loci wherein some human alleles are much more closely related to some chimpanzee (ancestral) alleles than they are to other human (derived) alleles. HLA encodes cell-surface antigen-presenting proteins that are used to recognize foreign invaders by cells of the immune system. The ancestral alleles have been maintained in the human population because either having rare alleles or having two different alleles has provided a selective advantage (Black and Hedrick 1997; Solberg et al. 2008).

By maintaining the frequency of two or more alleles at intermediate frequencies balancing selection will increase genetic variation within a population. Thus, a deviation from Hardy-Weinberg Equilibrium (HWE) would be one indication of balancing selection. Additionally, balancing selection could be proposed when observing an allele frequency distribution that is more even across populations than neutral expectations. It can be difficult to detect balancing selection and some studies have proposed that either balancing selection is a rare evolutionary phenomenon or it cannot be detected effectively by the methods currently used (Asthana et al. 2005; Bubb et al. 2006).

1.2.4.3 Negative Selection

Mutations that reduce the evolutionary fitness of the carrier are subject to negative selection (also called purifying selection). This may be the most pervasive form of selection in the human genome, and the easiest to detect. Indeed, much of the natural selection acting on genomes may be negative selection acting to remove new deleterious mutations (Kryukov et al. 2007). This type of selection will lead to a reduced genetic diversity at linked sites, as is observed in positive selection. Rates of elimination of slightly deleterious mutations are increased by negative selection, and rates of fixation of advantageous mutations are reduced (Charlesworth 1994). The strength of selection will depend on the magnitude of the selection, the mutation rate and the recombination rate (Charlesworth et al. 1993; Hudson and Kaplan 1995).

The specific molecular signatures of selection are discussed in the next section, but it is worth reminding ourselves of one general point here. While demographic processes affect the entire genome, natural selection leaves its signature at specific sites in the genome. Therefore, positively selected alleles can show distinct properties compared with the rest of the genome, such as rapid amino acid change, low diversity, high frequencies of rare and derived alleles, large differences between populations, and extended haplotypes. Let us now look at the tests used to detect these signatures.

1.3 HUNTING FOR SELECTION

I have previously mentioned that most genetic variation is assumed to be neutral. However, as more large-scale data become available, researchers are finding that selection in the human genome is not as rare as was thought previously (Akey et al. 2004; Bustamante et al. 2005; Lao et al. 2007; Sabeti et al. 2007; Vallender and Lahn 2004). Here I am interested in advantageous mutations that increase the evolutionary fitness of the individual.

However, detecting selection can be tricky as there is no single test for selection that applies to all circumstances (e.g. time and space) and all types of data (e.g. tests between species or within species). Even if I were to concentrate only on positive selection, there is no single test to detect it. For example, the ratio of non-synonymous to synonymous substitutions between species can be used to detect selective forces acting many millions of years ago (Nei and Gojobori 1986), whereas variation in allele frequencies (as calculated by the F_{ST} statistic) can suggest intra-species selection and the long-range haplotype test (Sabeti et al. 2002) can highlight even more recent events (acting within the past 10 thousand years or so).

While divergence data are commonly used to identify positive selection between species, this chapter will introduce the tests based on polymorphic data that are most commonly used to detect within-species selection—for the subject of this thesis, selection that has occurred in the human lineage after the split from the chimpanzee. To this end, patterns of nucleotide diversity, allele-frequency spectra, differentiation between populations, and haplotype structure can provide us with some information, where the expectation of low diversity, an excess of rare or derived alleles, large differences between populations and/or extended haplotypes might indicate positive selection (Ronald and Akey 2005; Sabeti et al. 2006).

1.3.1 Detecting Molecular Signatures of Selection

1.3.1.1 The Allele Frequency Spectrum

The allele frequency spectrum represents the distribution of the allele frequencies observed within a population and will identify selective sweeps occurring within the human species (less than 250 thousand years ago (KYA)). If a complete selective sweep has occurred, the swept region has very little variation and the amount of variation will depend on the recombination distance from the selected site. This will cause a skew in the frequency spectrum compared to what is expected under the standard neutral model. During a selective sweep, the hitchhiking effect drags variants to high or low frequency (Fay and Wu 2000). Therefore, in a nearly-complete sweep, there is an excess of high-frequency derived alleles. After the sweep, as new mutations accumulate, there will be an excess of rare variants. These may indicate a positively selected variant at a nearby site, but may also arise from negative selection or population expansion (Braverman et al. 1995; Przeworski 2002).

To summarize the signals expected from the allele frequency spectrum, positive selection will create a signature showing low overall diversity in the region but with an excess of rare alleles. We should remember, however, that demographic processes such as population expansion can also increase the frequency of rare alleles.

1.3.1.2 Neutrality Tests

The starting point of any selection test is to distinguish neutral variation from variation that has been subject to selection. The null hypothesis of neutrality tests assumes that all variants are neutral and deviations from the expected pattern are interpreted as possible selection. As has been noted, demographic changes can sometimes produce similar results, and some neutrality tests incorporate a demographic model in an attempt to allow for these effects (Schaffner et al. 2005).

One of the most important parameters in population genetics which underlies the neutrality tests is used as a measure of variation, theta (θ), defined as $4N_e\mu$ where N_e is the effective population size and μ is the rate of mutation per nucleotide per generation. The most commonly used neutrality tests is Tajima's *D* (Tajima 1989c) which summarizes the allele frequency spectrum. Tajima's *D* is the most robust test for identifying regions with an excess of common alleles or an excess of rare alleles. However, Tajima's *D* is also affected by population demography (Przeworski et al. 2000; Tajima 1989b). The test compares the average number of nucleotide differences between pairs of sequences to the total number of segregating sites (SNPs). If the difference between these two measures of variability is larger than expected under the neutral model, neutrality is rejected. Under the standard neutral model, the expectation of *D* is zero. A negative value of *D* reflects an excess of rare variants as might be expected after exponential growth (Slatkin and Hudson 1991) or a selective sweep. In contrast, a positive value of *D* reveals an excess of alleles at intermediate frequencies which may indicate population subdivision (Tajima 1989a) *or* balancing selection (Hudson and Kaplan 1988).

Another test based on the frequency spectrum is Fay and Wu's H test (Fay and Wu 2000). As an excess of rare alleles can indicate either positive or negative selection, this test, by focusing on identifying an excess of high frequency derived alleles, can help to distinguish between the two selective forces. Other commonly-used tests are Fu and Li's *D, D**, *F* and *F** (Fu and Li 1993). Fu and Li's tests compare the number of singletons with the number of polymorphic sites (giving *D, D**) or the nucleotide diversity (giving *F, F**); * indicates an unrooted tree and negative values an excess of singleton mutations.

1.3.1.3 Levels of Population Differentiation

Previous analyses of global allele frequency distributions indicate that the human population is not simply divided into a few clearly distinct groups ('races'). Roughly 84% of genetic diversity is represented by differences among individuals within populations, whereas differences among continents account for only around 10% (Barbujani et al. 1997). Even though these genetic differences between populations are small, statistical methods based on large variation datasets can be used to distinguish populations and assign individuals to their population of origin with high reliability (Jakobsson et al. 2008; Novembre et al. 2008; Rosenberg et al. 2002).

Allele frequency variation between populations provides an estimate of population differentiation and is largely determined by random genetic drift (Jobling et al. 2003). However, if a variant is under positive selection in a geographically isolated population, the allele frequencies around the selected variant change rapidly and this will lead to high levels of population differentiation in both the variant and the surrounding region. Therefore, as human population differentiation is not expected to be high, increased levels of diversity between populations could indicate positive selection (Nielsen 2005; The International HapMap Consortium 2005; Weir et al. 2005)

Adaptation, while not the only explanation for increased differentiation between populations, can be the effect of geographically localised selection, (i.e. local adaptation). Indeed, there are many accepted examples of selection in human populations at genes associated with locally adapted traits such as resistance to malaria (Hamblin and Di Rienzo 2000; Tishkoff et al. 2001), lactase persistance (Bersaglieri et al. 2004; Hollox et al. 2001; Tishkoff et al. 2007) and skin pigmentation (Lao et al. 2007; Norton et al. 2007).

1.3.1.4 Measures of Population Differentiation

Measures of population genetic differentiation are useful for detecting natural selection because they are highly sensitive to a large spectrum of adaptive events, varying in both strength and duration (Barreiro et al. 2008).

F-statistics, such as F_{ST} (Weir and Cockerham 1984), have traditionally been used to estimated population differentiation and they are good at identifying certain selective events (Sabeti et al. 2006). Genetic differentiation between two populations will increase when those populations become isolated and gene flow between them is limited. In humans, continental population differentiation started after they left Africa around 50 to 75 KYA (Barreiro et al. 2008). The *F*-statistic can therefore potentially reveal the effects of natural selection over the past 75 thousand years.

Under the assumption of neutrality, F_{ST} is determined by demographic history which will affect all loci similarly. The value of F_{ST} is between 0 and 1, where 0 implies no differentiation between populations and 1 implies complete differentiation. If natural selection is acting on a locus, the F_{ST} value will decrease in the case of negative or balancing selection and increase when positively selected. Furthermore, positive selection might often be expected to be specific for one population or region. The genome-wide average F_{ST} has been estimated to be around 0.12 (Akey et al. 2002; Barreiro et al. 2008; The International HapMap Consortium 2005; Weir et al. 2005), so that any values significantly higher can be considered to indicate possible candidates for positive selection. On average, however, this means that 12% of genetic differences are ascribable to differences among subpopulations and 88% of the total genetic variation exists within the subpopulations themselves.

The highest classical F_{ST} value observed in humans (F_{ST} = 0.78) (Cavalli-Sforza et al. 1994) is that of the Fy*O allele in the Duffy blood group. This mutation is widely accepted to be under positive selection with the Fy*O derived allele almost fixed in most sub-Saharan African populations but very rare outside Africa (Hamblin and Di Rienzo 2000; Hamblin et al. 2002). Significantly, this allele has been shown to be associated with resistance to malaria infection by *Plasmodium vivax* (Livingstone 1984).

1.3.1.5 Linkage Disequilibrium and Haplotype Structure

An incomplete sweep (when the adaptive mutation has not yet been fixed in the population) leaves a distinct pattern in the haplotype structure (Sabeti et al. 2002). Thus, recent selected sweeps are expected to increase the amount of LD around a selected variant producing long-range haplotypes (Przeworski 2002; Sabeti et al. 2002) while old or recurrent selective sweeps will not lead to high levels of LD (Przeworski 2002). Furthermore, neutrality or old balancing selection will tend to reduce LD and generate short-range haplotypes. However, long-range haplotypes caused by a selective sweep will likely be short-lived as recombination will rapidly break up allelic associations after the sweep, and high-frequency alleles will drift to fixation (Przeworski 2002; Sabeti et al. 2006), so will be associated only with recent selection events. However, some of the most dramatic changes to the human environment have occurred within the past 10 thousand years, so recent selective events are of great interest.

1.3.1.6 Long-Range Haplotype Tests

With increased knowledge of LD and recombination rates in the human genome, so-called long-range haplotype (LRH) tests have been developed to detect unusual patterns indicating selection. Among the most commonly used are the Relative Extended Haplotype Homozygosity (REHH) test (Sabeti et al. 2002) and the Integrated Haplotype Score (iHS) (Voight et al. 2006). The LRH test relies on the relationship between an allele's frequency and the amount of LD surrounding it. As the test is based on SNPs, which are still segregating in a population, it will detect recent selective sweeps, generally occurring within the past 10 thousand years. When a new allele comes into a population, the amount of LD surrounding it will be long (long-range LD). If the allele turns out to be neutral, it will on average take a long time to reach a high frequency so that LD will decay as it is broken up by recombination, leading to a pattern of short-range LD. If, however, the allele is advantageous it can increase in frequency faster than it takes for recombination to break up the LD surrounding it (see Figure 4)

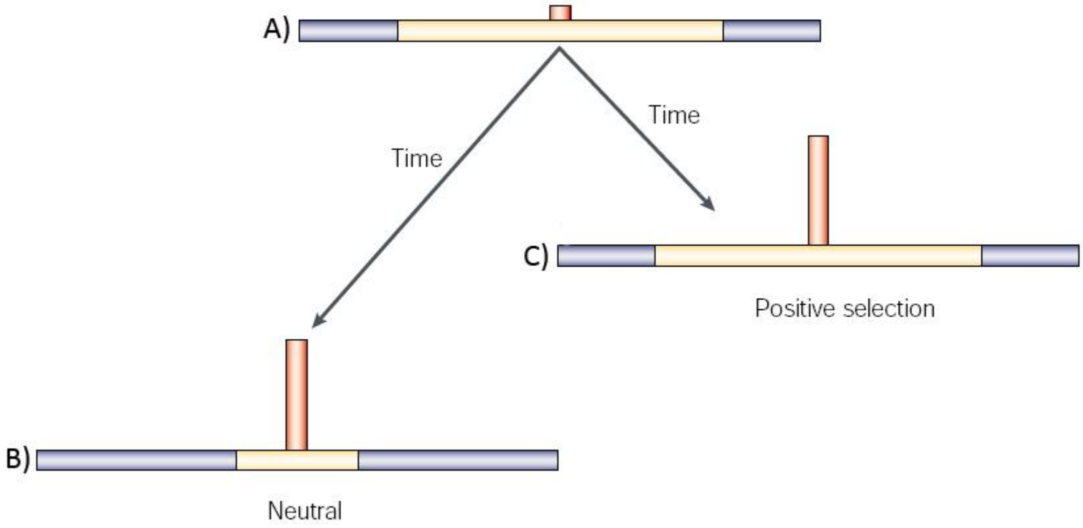

Figure 4 Detecting recent positive selection using linkage disequilibrium analysis. A) A new allele (red) starts out at a relatively low frequency on a background haplotype (blue) that is characterized by long-range LD (yellow) between the allele and the linked markers. Time passes and if **B)** the allele is neutral, its frequency may increase as a result of genetic drift, but if so recombination breaks up the LD surrounding it and short-range LD is produced; however, if **C)** the allele turns out to be advantageous it might increase in frequency much faster than it will take for recombination to break up the LD between the allele and the linked markers, and a pattern of long-range LD is observed. This figure is taken from (Bamshad and Wooding 2003).

Therefore, an allele at a high frequency with unusually long-range LD can be taken as a candidate for positive selection. Indeed, several studies have identified long-range haplotypes in genes previously suggested to be under positive selection (Sabeti et al. 2007; The International HapMap Consortium 2005).

1.4 Recent Human Evolution

The environment that we live in now is radically different from the environment that ancestral human populations were adapted to. The human species has travelled far since its origin in Africa some 200 KYA, and throughout this journey, episodes of bottlenecks, founder effects, migration, gene flow, mutation, genetic drift and selection have taken place, and ultimately shaped the genomes of modern human populations.

In particular, changes in the past 10 thousand years, mainly because of the domestication of plants and animals, have been the most dramatic, affecting the environment and lifestyle of nearly all humans. These changes are bound to have led to the evolution of new adaptive traits.

1.4.1 The Origin and Dispersal of Modern Humans

For the past few decades the origin of anatomically modern humans (AMH) has been a subject of hot debate. Today, there seems to be general agreement among scientists that AMH arose in Africa and spread from there throughout the world. However, the agreement usually stops there as the timing, routes, possibilities for admixture and expansion events are still under consideration (Yngvadottir and Carvalho-Silva 2008).

The early debate centred around two main but opposing views, the 'Recent African Origin' model and the 'Multiregional Evolution' model. The debate has been resolved to most scientists' satisfaction and while some (Eswaran et al. 2005; Fagundes et al. 2007; Plagnol and Wall 2006) have shown that the Multiregional model cannot be completely ruled out, the introduction and results discussed here will be based on the more widely accepted Recent African Origin theory. Briefly, the story goes something like this: AMH arose in East Africa approximately 200 KYA (Jobling et al. 2003; Liu et al. 2006; Ray et al. 2005) and their effective population size was around 10,000 at that time (Harpending et al. 1998; Schaffner et al. 2005; Takahata et al. 1995; Tenesa et al. 2007). In support of this view, recent re-analysis of fossil evidence, the skull Omo 1 from Ethiopia, has provided an age of around 195 thousand years, which makes it the earliest known AMH yet found (McDougall et al. 2005).

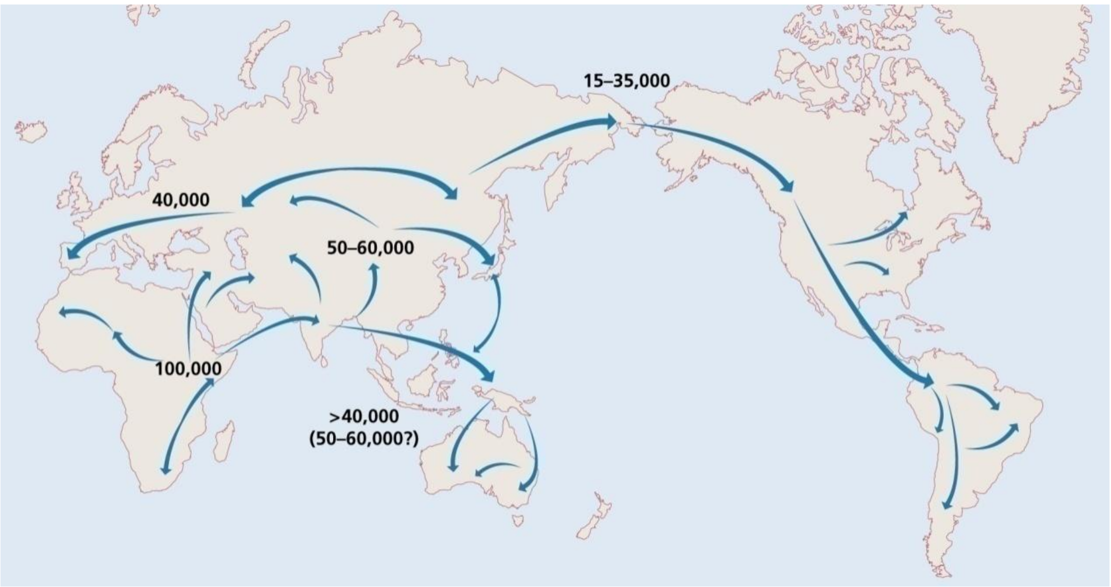

Figure 5 Possible scenario for the timings and dispersal routes of AMH's journey out of Africa. A range of expansions within Africa occurred ~100 thousand years ago, which was then followed by subsequent expansions into the rest of the world. This figure is taken from (Cavalli-Sforza and Feldman 2003).

Around 100 KYA there was a warm interglacial period allowing a range of population expansions within Africa extending to the Levant, followed by contraction when the climate deteriorated after 80-90 KYA (Lahr and Foley 1994; Lahr and Foley 1998; Mellars 2006). Then, within Africa, further key steps in the evolution of modern humans occurred with the evolution of modern behaviour (Henshilwood et al. 2002). Subsequent dispersals of anatomically and behaviourally modern humans into Asia, and Oceania occurred around 40-60 KYA, with Europe colonized after 40 KYA and the final colonization of the Americas 15-20 KYA (Figure 5) (Cavalli-Sforza and Feldman 2003; Liu et al. 2006). Thus, modern populations inherited their genes almost entirely from these humans that were both anatomically and behaviourally modern (Jobling et al. 2003), although there is still debate about whether interbreeding with archaic humans occurred and contributed a small amount of genetic material to the modern gene pool.

The population that left Africa must have experienced a bottleneck, i.e. a reduction in population size followed by a recovery, resulting in a relatively small ancestral population from which all modern humans outside Africa originate.

Studies of the allele frequency spectrum have suggested that the African-American population (taken to represent Africans) shows a history of moderate but uninterrupted expansion and larger effective population size while Asian and European populations have a bottleneck shaped history as they experienced a reduction of effective population size in the past followed by a recovery (Falush et al. 2003; Marth et al. 2004). In further support of these views, a greater genetic diversity (Cann et al. 1987; Harpending and Rogers 2000; Przeworski et al. 2000; Zhao et al. 2000) and decreased LD (Jakobsson et al. 2008) in African compared to non-African populations have been revealed, which emphasises Africa as the place of origin for AMH. Figure 6 illustrates this by showing how genetic diversity decreases as we move further away from Africa. This is consistent with a bottleneck occurring at the time of migration out of Africa and thereby reducing genetic diversity of non-African populations.

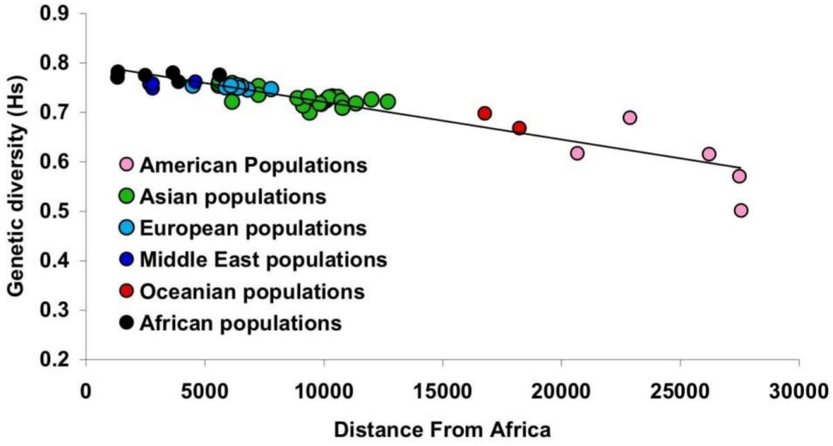

Figure 6 Relationship between genetic diversity and geographic distance from Africa for 51 distinct present-day populations. There is a decline in the genetic diversity of human populations with increasing distance from our assumed place of origin in Africa. This figure is taken from (Prugnolle et al. 2005).

1.4.2 Out of the Cradle—and The Neolithic Revolution

As humans found themselves in new environments outside Africa, they encountered many differences including colder temperatures and novel animals and plants. When the climate warmed and stabilised at the beginning of the Holocene ~10,000 years ago, there was a shift away from a hunting and gathering lifestyle towards subsistence based on agriculture and the domestication of animals for many humans. This change is marked in the archaeological record by the beginning of the Neolithic period (starting ~8,000-10,000 years ago in several independent centres) and as a result the human population experienced dramatic changes in population size, population density and cultural conditions. As a consequence humans had to adapt to new environments, diets and diseases.

Increased population densities implying close proximity to other people, and also close contact with domestic animals facilitated both the origin and spread of infectious diseases in human populations (Wolfe et al. 2007). In response to infectious diseases, the human genome has adapted in various ways, notably by favouring variation in genes involved in the immune system as well as several expressed in red blood cells, the sites of malaria parasite (*Plasmodium sp.*) replication, such as the Duffy antigen, the alpha and beta globins, and G6PD.. Malaria is a leading cause of death in the world today and thus it is likely that selective forces have acted in response. In fact, it has been suggested that the strongest force of selection in humans is played by the infectious disease malaria (Kwiatkowski 2005). The sickle cell variant in the haemoglobin gene was mentioned in section 1.2.4.2 as an example of heterozygote advantage and the Duffy Fy*O allele with extremely high levels of population differentiation in section 1.3.1.4. Both variants have been reported with the highest frequency of the protective allele in Africa, where malaria is endemic. In addition to this, haplotype analysis of two variants in the *G6PD* gene, "A-" and "Med", has provided an age estimate between 3,840-11,760 years ago and 1,600-6,640 years ago, respectively. The variants result in enzyme deficiency and have been implicated in resistance to malaria and these age estimates therefore suggest that malaria did not become hyperendemic until the origin of agriculture ~10 KYA when people started settling down (Tishkoff et al. 2001).

The domestication of livestock led to a change in diet, studied especially in relation to the practice of milk consumption by adults during and after the Neolithic revolution. The inability to digest the major sugar, lactose, in milk (lactase nonpersistance, LNP) in adulthood is normal to all mammals. LNP is therefore the ancestral state, whereas lactase persistance (LP) is observed in some human populations and may have become advantagous when milk from domesticated animals became available for adults to drink. In fact, there is a relationship between the frequency of LNP in a population and the population's history of dairy farming (reviewed in Swallow 2003). Therefore, it comes as no surprise that the highest levels of LP are found in northern European populations (>90% in Swedes and Danes), which are known to have practiced dairying for a long time, and in some pastoral African populations that rely on milk in their diet (~90% in the Tutsi and ~50% in the Fulani). The lowest values, on the other hand, are reported in populations of Asian ancestry (1% in the Chinese), who were not dairy farmers, and in agricultural populations within Africa (~5-20% in West Africa) (Bloom and Sherman 2005; Swallow 2003; Tishkoff et al. 2007).

A SNP ~14 kb upstream of the *LCT* gene (within *MCM6*), has been associated with LP in several European populations (Enattah et al. 2002) but this variant was not found at significant frequencies in pastoral African populations, some of which were LP (Mulcare et al. 2004). Indeed, evidence from the LRH test has revealed that the *LCT* gene has undergone recent positive selection in the populations of European ancestry but not in the populations of African (although see later findings below) or Asian ancestry examined (Bersaglieri et al. 2004; The International HapMap Consortium 2005) and was estimated to have arisen in the the past ~2,000–20,000 years (Bersaglieri et al. 2004). However, a recent study found that the LP allele was absent from ancient DNA samples dated to the Neolithic, and thus concluded that LP was in fact rare in early European farmers (Burger et al. 2007).

The mystery of the causative allele for LP in African pastoralists was partially solved with the identification of other SNPs, also in the same region of the *MCM6* gene, found to be associated with persistance specifically in some African populations (Ingram et al. 2007; Tishkoff et al. 2007). This variant was also revealed to be positively selected based on evidence from the LRH test, and the selective sweep was proposed to have started ~3,000-7,000 years ago (confidence interval 1,200-23,200 years ago) (Tishkoff et al. 2007). Thus, Africans and Europeans have a similar LP phenotype but the causative variant is different between the two populations. This was taken to be an example of convergent evolution occurring independently in the two populations that were exposed to dairy farming at different times.

1.5 Evolution by Gene Loss?

The theory that gene duplication is the major factor in shaping evolution was proposed many years ago by Susumu Ohno (1970) and is now widely accepted. The theory that gene loss can also have such an effect is, however, a relatively new one and was first proposed by Maynard Olson (1999). Common sense may lead us to consider gene loss as a bad thing and to associate adaptation with genes that are somehow "better". However, as the thrifty gene theory has proposed, some genes that were good in the past may have become a burden in modern life. In this section I will explore the possibility that gene loss may be good for one's evolutionary fitness.

1.5.1 Different Types of Gene Loss

Section 1.1 gave an introduction to several types of variation observed in the human genome and considered some possible consequences. In this section I will focus on the types of mutations that cause a gene to lose its function. One molecular mechanism for gene loss is the introduction of a premature termination codon (PTC). This can be caused by nonsense mutations and frame shifting indels (mentioned in sections 1.1.1 and 1.1.2) as well as by splice site mutations with the skipping of a single exon containing a number of nucleotides that cannot be divided by three (reviewed in Cartegni et al. 2002). These mutations must have severe consequences as they can alter the stability of transcripts and function of proteins and might therefore be expected to be rare. However, examination of alternative transcripts in humans revealed that one-third of mRNA isoforms contained PTCs (Lewis et al. 2003).

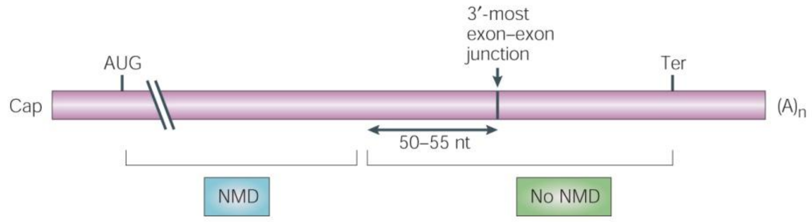

Figure 7 NMD prediction according to the "50-55 nucleotide" rule. If the PTC is located more than 50-55 nucleotides upstream of the 3'most exon-exon junction (region indicated in blue) NMD is triggered and the transcript is degraded. If the PTC is located in the last exon or less than 50-55 nucleotides away (region indicated in green) NMD is escaped and results in a truncated protein. Figure is taken from (Maquat 2004).

The PTC-causing mutations might be expected to result in a shorter protein, but truncated proteins are likely to be deleterious and are usually eliminated by a process called nonsense-mediated mRNA decay (NMD) (Hentze and Kulozik 1999; Maquat 2004). NMD is a quality control-based mRNA surveillance system that recognizes transcripts with PTCs at specific positions and degrades them (see Figure 7).

NMD thereby prevents the accumulation of truncated and potentially harmful proteins in addition to regulating gene expression. As a rule, in most mammalian cells NMD is triggered if the PTC is present more than 50-55 nucleotides upstream of the 3'-most exon-exon junction (Maquat 2004; Nagy and Maquat 1998). If the NMD pathway is triggered it will eliminate the production of the protein and the gene product is completely lost. However, if the PTC is located either in the last exon or less than 50-55 nucleotides upstream of the last exon-exon boundary, NMD can be escaped resulting in the production of a truncated protein (Maquat 2004; Mort et al. 2008). While the 50-55 nucleotide rule is often applied to mammalian cells, exceptions have been reported (see e.g. Inacio et al. 2004; Isken and Maquat 2007; Zhang and Maquat 1997).

PTCs can be disadvantageous and such mutations are common causes of genetic disease (Frischmeyer and Dietz 1999; Olson and Varki 2003). However, sometimes the mutation is neutral, and can increase in frequency as a result of drift, or advantageous, and can increase in frequency because of selection (see examples in section 1.5.4).

1.5.2 The Thrifty Gene Hypothesis

The thrifty gene hypothesis (Neel 1962) was introduced to explain the high prevalence of type II diabetes and obesity in modern human populations. According to the hypothesis, certain genetic variants evolved in the past to better enable the storage of fat and carbohydrates and were thus advantageous for our hunter-gatherer ancestors as they went through seasonal cycles of feast and famine. However, as modern food production, processing and storage has provided western populations with an abundance of food, these variants have become disadvantageous as they predispose their carriers to obesity and diabetes. In this respect we have genes that were good in the past but have become a burden today and we might be better off losing them. While the thrifty gene hypothesis has been used extensively in medical genetics over the past decades, recent studies have cast doubts on its validity and relevance to modern human populations, both on general theoretical grounds (Speakman 2006) and as a result of specific studies of diabetes susceptibility alleles (Helgason et al. 2007) as well as those associated with obesity (Ohashi et al. 2007).

However, the effect of changes in the environment and lifestyle of human populations are still emphasized by the large number of ancestral alleles known to increase risk to common diseases (Di Rienzo and Hudson 2005). These ancestral alleles were likely adapted to our ancient lifestyles, but have become disadvantageous after changes in the environment, while the derived alleles may have become advantageous or neutral. For example, the *ENPP1* gene has a mutation in which the derived allele provides protection against obesity and type II diabetes (Meyre et al. 2005) and is present in ~90% non-Africans (Barreiro et al. 2008).

1.5.3 Less is More—An Evolutionary Theory of Gene Loss

In 1999 Maynard Olson introduced his “less-is-more” hypothesis, where he proposes gene loss to be a plausible mechanism for adaptive evolutionary change (Olson 1999). As discussed above and further below, gene loss may sometimes be advantageous in itself. In addition, if a gene loses its function without being completely deleted, it can persist in the genome and might therefore be available for subsequent evolutionary forces to act upon. Furthermore, as I have discussed in section 1.2.4.2, heterozygous advantage can also keep disrupting alleles in a population that would otherwise be disadvantageous in a homozygote state (see also discussion in Dean et al. 2002). While the focus in this thesis is on gene loss events that are still segregating in humans, gene loss that has occurred in the human lineage after the split from the chimpanzee can potentially explain some of the differences observed between the two species. Examples of this are suggested to include delayed postnatal development as well as loss of muscle strength and hair in humans (Olson and Varki 2003). In response to this, three studies have recently explored gene loss events in an evolutionary context. One study focused on events occurring since the common ancestor of primates and rodents during the past ~75 million years (Zhu et al. 2007). The other two focused on more recent events, one on inactivation that has occurred in the human lineage after its separation from the chimpanzee 5-7 million years ago (MYA) (Wang et al. 2006), and the other on nonsense-SNPs which are still segregating in human populations (Savas et al. 2006) as will be done in this thesis.

Wang *et al* (2006) found that lost genes were mainly found to be involved in chemoreception and immune response, which suggests potential species-specific features in these aspects of the human physiology. Using publicly available data from dbSNP, Savas *et al* (2006) identified 28 nonsense-SNPs with the minor allele frequency (MAF) information reported. These were found to be more common (~79% had a MAF≥0.05 in one or more populations) than would be expected if they were simply deleterious. They furthermore identified a non-uniform distribution across the three human populations they analysed, as eight SNPs were reported to be prevalent in all three whereas six SNPs were found exclusively in one or two population(s). By looking at the position of each nonsense-SNP within the gene and resolving whether they triggered NMD or not, they concluded that the 28 nonsense-SNPs were likely to affect the gene function.

It seems that while gene loss may, in many cases, be detrimental for one's health, such inactivating mutations are nevertheless prevalent in the human genome. In fact as will be discussed in the next section, several reports have revealed the selective advantage of losing a particular gene.

1.5.4 You Lose, You Gain—Examples of Advantageous Gene Loss

On a deeper evolutionary scale, the human *MYH16* gene contains a frameshift deletion giving rise to a PTC inactivating the gene, whereas other primates have the active version which is expressed strongly in muscles of the cheeks (Stedman et al. 2004). Initially, this mutation was thought to have occurred about 2.4 MYA and the loss of *MYH16* was suggested to have influenced the anatomy of the head and to have removed a constraint which may have paved the way to the development of the modern human brain (Stedman et al. 2004). Another study has, however, raised doubts about this gene being positively selected and re-dated the mutation at about 5.3 MYA (Perry et al. 2005). The case remains unsolved.

Interestingly, many examples of advantageous gene loss seem to be related to immune response and such genes have previously been reported to be overrepresented in human-specific gene loss (Wang et al. 2006). I have already discussed the advantageous loss of the Duffy Fy*O allele in section 1.3.1.4 and the heterozygote advantage observed in having one copy of the sickle cell allele in section 1.2.4.2, because of their resistance to malaria. Malaria is endemic in many countries in Africa but other infectious diseases such as AIDS are becoming prevalent as well. *CCR5* is polymorphic in humans for a 32 base pair deletion which inactivates the gene which is itself a receptor for HIV. Consequently homozygotes for the deletion are protected against HIV infection and AIDS whereas heterozygotes receive some level of protection (Dean et al. 1996). The loss of this gene is clearly advantageous now but it still shows a pattern of variation consistent with neutral evolution in the past (Sabeti et al. 2005) and the reason for the relatively high frequency of the deletion in European and West Asian populations—neutral drift or past selection—remains unclear. An additional example relating to the immune system, and which will be discussed in greater detail in section 4.1.1 is that of the *CASP12*. This gene is polymorphic for an inactivating mutation in human populations, with carriers of the inactivated allele being more resistant against severe sepsis (Saleh et al. 2004; Xue et al. 2006).

On a non-immune related level, the *ACTN3* gene, dubbed "the gene for speed" has an interesting story to tell. *ACTN3* is an actin-binding protein mainly expressed in skeletal muscles. A nonsense-SNP was identified within the gene and the inactive homozygous form was found at a high frequency in the human population (MacArthur and North 2004). The complete loss of this gene does not result in a disease phenotype, an observation which may be explained by the compensation of a closely related homolog (*ACTN2*). However, the loss of *ACTN3* was also found to have a consequence of its own, as the homozygote state was found to be associated with athletic performance. Elite sprint athletes were found to have significantly higher frequencies of the normal (active) allele than control samples, suggesting that the active form has an evolutionary advantage in terms of increased sprint performance. However, the heterozygous state was found at high frequencies in female sprint and at lower frequencies in endurance athletes, suggesting the possibility of a sexual difference in the effect of the nonsense-SNP (Yang et al. 2003). Furthermore, it was shown that loss of alpha-actinin-3 expression in a knockout mouse model results in an increase in intrinsic endurance performance (Chan et al. 2008; MacArthur et al. 2007). As the nonsense-SNP had different effects on sprint and endurance performance in humans, it was at first proposed to be undergoing balancing selection in the human population (Yang et al. 2003) but was later suggested to be positively selected in populations of European and East Asian ancestry (MacArthur et al. 2007).

Perhaps more examples of advantageous gene loss in the human genome have yet to be revealed. In any case, this study will attempt to survey nonsense-SNP inactivation mutations on a genome-wide scale in order to describe their prevalence and distribution more completely.

1.6 Thesis Aim

In section 1.5.3 I mentioned three recent studies focused on the identification of gene loss events. Two of them (Wang et al. 2006; Zhu et al. 2007) were focused on older events involving human-specific loss, and one (Savas et al. 2006) was looking for nonsense-SNPs still segregating in the human species. While this study, performed solely *in silico,* made excellent use of available data, it was limited by the data as it relied on a specific MAF observed in a set of three populations. Their identification of 977 nonsense-SNPs in dbSNP was thus greatly reduced to 28 SNPs analysed in detail. I also started with a large set of nonsense-SNPs (n = 805) identifiable in dbSNP that were compatible with the genotyping platform used at the Wellcome Trust Sanger Institute (WTSI), and they were subsequently genotyped in 1,151 individuals from 56 geographically distinct worldwide populations. With a larger dataset, I was able to investigate the prevalence and selective forces acting on 169 nonsense-SNPs found to be variable in humans.

My curiosity about the evolutionary forces acting on nonsense-SNPs was first triggered by our initial study (Xue et al. 2006) of the *CASP12* gene which provided an excellent example of advantageous gene loss. Together with Maynard Olsons's "less-is-more" hypothesis, I decided to put his theory to a more systematic test by embarking on a genome-wide study of loss events. A number of nonsense-SNPs had been identified in dbSNP, and since such SNPs are perhaps the easiest form of gene loss to analyse on a large scale with the new genotyping assays, nonsense-SNPs became my target of choice. With this study I wanted to identify the general pattern of selection acting on the class of nonsense-SNPs as a whole, and determine whether the inactive form had always spread because of neutral drift, or more excitingly sometimes by positive selection. The ultimate aim was thus to identify outliers that could potentially reveal some additional interesting contributions of gene loss to the evolution of our species.