Evolution by Gene Loss?

A genome-wide survey of human SNPs that introduce premature termination codons

Bryndís Yngvadóttir

Queens' College
University of Cambridge
September 2008

This dissertation is submitted for the degree of Doctor of Philosophy

UNIVERSITY OF CAMBRIDGE

Declaration

This thesis describes my work undertaken in the laboratory of Dr Chris Tyler-Smith, at The Wellcome Trust Sanger Institute, in fulfilment of the requirements for the degree of Doctor of Philosophy, at Queens' College, University of Cambridge. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. The work described here has not been submitted for a degree, diploma, or any other qualification at any other university or institution. I confirm that this thesis does not exceed the page limit specified by the Biology Degree Committee.

Bryndís Yngvadóttir
Cambridge, September 2008

Abstract

Nonsense-SNPs introduce premature termination codons into genes, and can result in the absence of a gene product or a truncated and potentially harmful protein, so are often considered disadvantageous and associated with disease susceptibility. As such, the disrupted allele might be expected to be rare and, in healthy people, observed only in a heterozygous state. However, some, like those in the caspase-12 and actinin-3 genes, are found at high frequencies with many homozygotes and seem to have been advantageous in recent human evolution.

The goal of this project was to perform a genome-wide survey of nonsense SNPs in the human genome and evaluate the selective forces acting on them. Most available nonsense-SNPs (n=805) and a set of synonymous control SNPs (n=731) were genotyped in 1,151 individuals from 56 geographically distinct worldwide populations.

I identified 169 genes containing nonsense-SNPs that were polymorphic in the samples, of which 99 were found in a homozygous state, showing that both copies of these genes can be truncated in healthy subjects without any obvious consequences. This study illustrates how much the human gene content varies between individuals: on average by 24 genes (out of about 20,000) by nonsense-SNPs alone. Gene Ontology analysis revealed that there was significant overrepresentation of genes involved in olfactory reception and the nervous system.

As might be expected, these SNPs as a class were found to be slightly disadvantageous over evolutionary timescales, but a few nevertheless showed signs of being advantageous, indicated by unusually high levels of population differentiation or a departure from neutrality in tests based on resequencing the region surrounding the SNP in multiple individuals. In addition to caspase-12, a *SEMA4C* nonsense-SNP was confined to the Americas where it reached high frequency, while a *MAGEE2* nonsense-SNP was present at high frequency only in East Asia and showed evidence of positive selection. Several examples of beneficial

gene loss could thus be found, and have contributed in a small but significant way to human evolution.

Publications

Publications arising during the course of the work described in this thesis by the time of submission:

Yngvadottir B, Xue Y, Searle S, Hunt S, Delgado M, Morrison J, Whittaker P, Deloukas P, Tyler-Smith C (2009). A genomewide survey of the prevalence and evolutionary forces acting on human nonsense-SNPs. *American Journal of Human Genetics*, **84**(2):1-11.

Xue Y, Zhang X, Huang N, Daly A, Gillson CJ, **Yngvadottir B**, Nica AC, Woodwark C, Chen Y, Ayub Q, Mehdi SQ, Li P, Tyler-Smith, C (submitted). Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation.

Yngvadottir B, Carvalho-Silva DR. (2008) *Reconstructing Human History Using Autosomal, Y-Chromosomal and Mitochondrial Markers*. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0020819

Yngvadottir B. (2007) Insights into modern disease from our distant evolutionary past. *European Journal of Human Genetics,* **15**(5):603-6.

Xue Y, Daly A, **Yngvadottir B**, Liu M, Coop G, Kim Y, Sabeti P, Chen Y, Stalker J, Huckle E, Burton J, Leonard S, Rogers J, Tyler-Smith C. (2006) Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *American Journal of Human Genetics*, **78**(4):659-70.

Gutala R, Carvalho-Silva DR, Jin L, **Yngvadottir B**, Avadhanula V, Nanne K, Singh L, Chakraborty R, and Tyler-Smith C. 2006. A shared Y-chromosomal heritage between Muslims and Hindus in India. *Human genetics,* **120**(4):543-551.

Acknowledgements

Firstly, I would like to thank my supervisor, Chris Tyler-Smith, whose office door was always open and who believed in me and helped me in every way throughout my PhD. I feel truly privileged to work with him. Many thanks go to Yali Xue who (in addition to giving me moral support and all the time in the world) has taught me so many things in the lab and helped me to understand the data and analysis in every detail possible.

Thanks go to Matt Hurles, Alex Bateman and Bill Amos for advice and support throughout the PhD. Thanks go also to Dan Turner who managed to teach me the ropes in the lab without getting impatient with my ignorance of lab-based techniques. I thank the sample donors, and Howard Cann for making the HGDP-CEPH data freely available. Thanks go to the many people passing through our Team 19 of human evolution in the past four years for their enthusiasm and expertise. In particular, I would like to mention Denise Carvalho-Silva, Tatiana Zerjal and Cara Woodwark. Thanks to all the great people in the Sanger Genotyping Platform and Large Scale Sequencing Pipeline groups, without whom I would have no data. I would especially like to thank Panos Deloukas and members of his team, Sarah Hunter, Pam Whittaker, Marcos Delgado and Rhian Gwilliam for the genotyping and QC processing. Special thanks go to Barbara Stranger and Manolis Dermitzakis for allowing me to make use of their gene expression data and then helping me understand what it was all about. I would like to thank Pardis Sabeti and Pat Varilly for advice on the LRH-tests and for giving me the source code for Sweep. Thanks go also to all the HelpDesk people at Ensembl and HapMap that I have bugged with endless queries throughout the years. Thanks to Areum Han who gave me the full data from SNP2NMD and to Ni Huang, Yuan Chen and Jim Stalker for various useful scripts. Great thanks to Joan Green and Andrew King for excellent "Journal Picks" pointers and for the endless renewal of my library books. I would also like to thank Christina Hedberg-Delouka deeply for all her support and kind words in the past years. Big thanks go to the Wellcome Trust for an excellent PhD program and for its generous fellowship that not only put a roof over my head but also allowed me to go to all those great conferences.

A special thanks go to Agnar Helgason for giving me a head start in my transition from the studies of social anthropology (old but not forgotten) to the exciting world of genetics—things are changing so fast in this line of work. Lots of love goes to my Sanger girls, Raffaella, Eleni and Antigone for "keeping it real". These past years have been tough, but because of you they have also been a joy ride of fabulous dinner parties, awesome red wine and some great vibes playing in the background. I'll miss us. Thanks go also to my Ice girls, Ellen, Þurý and Hulda, who have always encouraged me with their hugs and kisses, be they natural or electronic. I would especially like to thank my parents for their love and never-ending support, for standing behind me in the good times and the bad, and for making me less home-sick by making Cambridge their second home away from home. Thanks to Guðrún and Þorsteinn for being the greatest siblings a girl could hope for. And finally, my greatest thanks go to my love Bjarki, whom I could not have done this PhD without. Your support and understanding has been essential throughout the past eleven years and surprisingly so has your cooking for the past few weeks! Hopefully I can do the same for you in a year's time. This thesis is dedicated to you.

Table of Contents

Abbreviations

AMH	anatomically modern humans
ASO	allele-specific oligo
bp	base pairs
BP	biological process
CEU	CEPH Utah residents with ancestry from northern and western Europe
CHB	Han Chinese in Beijing
CNV	copy number variant
DAF	derived allele frequency
DAVID	Database for Annotation, Visualization and Integrated Discovery
EHH	extended haplotype homozygosity
GC	Gene Call
GO	gene ontology
HGDP-CEPH	CEPH Human Genome Diversity Cell Line Panel
HGMD	Human Gene Mutation Database
HLA	human leukocyte antigen
HWE	Hardy-Weinberg Equilibrium
iHS	Integrated Haplotype Score
JPT	Japanese in Tokyo
kb	kilobases
KYA	thousand years ago
LD	linkage disequilibrium
LNP	lactase nonpersistance
LP	lactase persistance
LRH	long-range haplotype
LSO	locus-specific oligo
LWK	Luhya in Webuye, Kenya
MAF	minor allele frequency
Mb	mega base
MF	molecular function
MYA	million years ago
NCBI	National Center for Biotechnology Information
NMD	nonsense-mediated mRNA decay
OR	olfactory receptor
ORF	open reading frame
PCR	polymerase chain reaction
PHASE	Phylogenetics And Sequence Evolution
PTC	premature termination codon
REHH	relative extended haplotype homozygosity
SNP	single nucleotide polymorphism
STS	sequence tag site
UCSC	University of California Santa Cruz
VNTR	Variable number of tandem repeat
WGA	whole-genome-amplification
WTCCC	Wellcome Trust Case Control Consortium
WTSI	Wellcome Trust Sanger Institute
YRI	Yoruba in Ibadan, Nigeria

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).

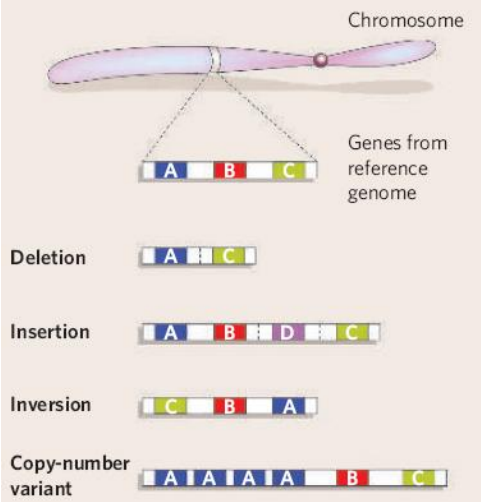


Figure 1 Different types of variation. A section of the reference sequence is given at the top, followed by examples of the way different types of variation could change that sequence. This figure is adapted from (Check 2005).

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.

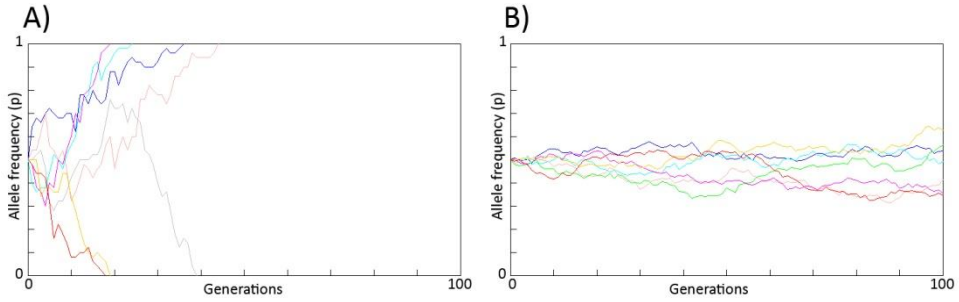


Figure 2 Simulations of genetic drift for an allele starting at a frequency of 0.5 over 100 generations. A) Population size = 25, the allele gets rapidly lost or fixed **B)** Population size = 1000, the allele frequency changes are more subtle. The simulation was created with an online simulator available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

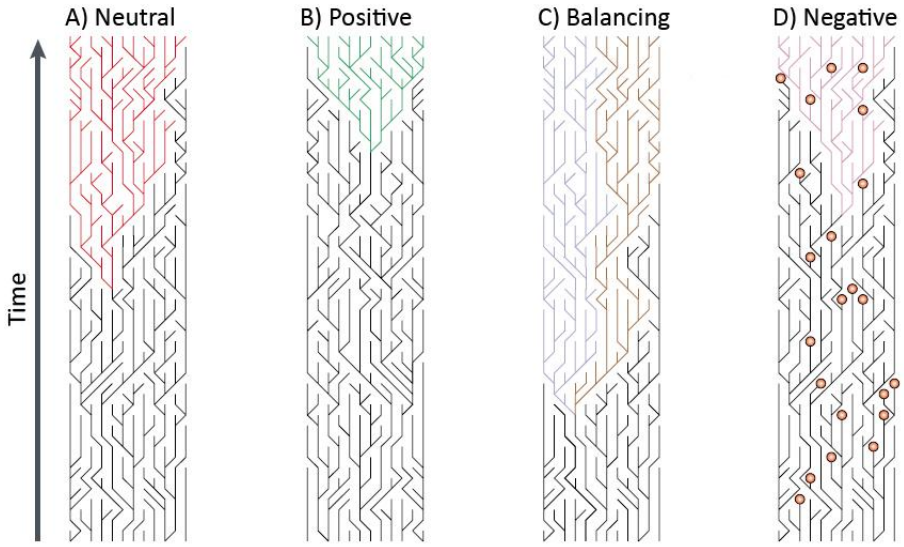


Figure 3 Effects of natural selection on gene genealogies and allele frequencies. A) The genealogy of a neutral allele (red) as it drifts to fixation. **B)** The genealogy of a positively selected allele (green) that is driven to fixation more quickly than is expected from neutrality. **C)** The genealogy 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)

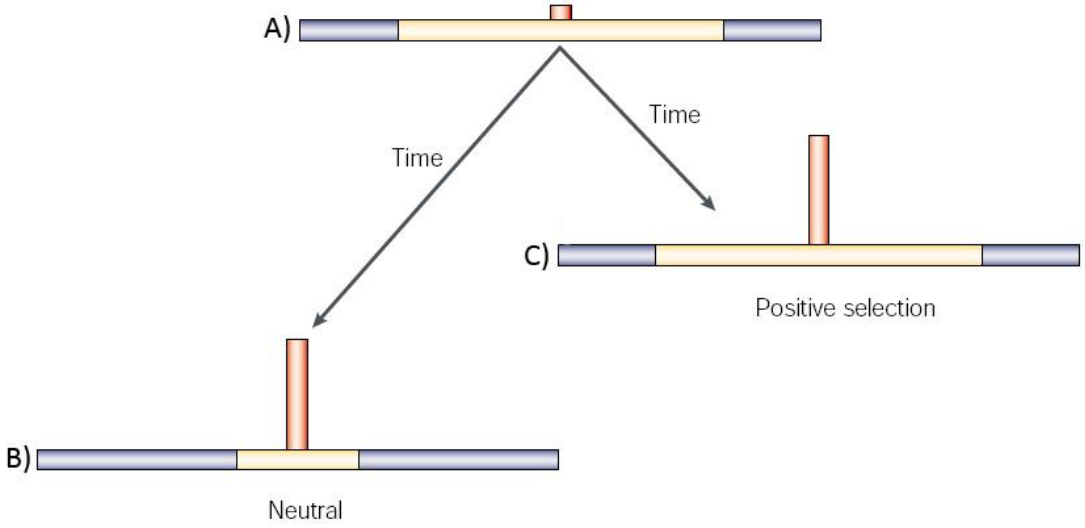


Figure 4 Detecting recent positive selection using linkage disequilibrium analysis. A) A new allele (red) starts out at a relatively low frequency on a background haplotype (blue) that is characterized by long-range LD (yellow) between the allele and the linked markers. Time passes and if **B)** the 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).

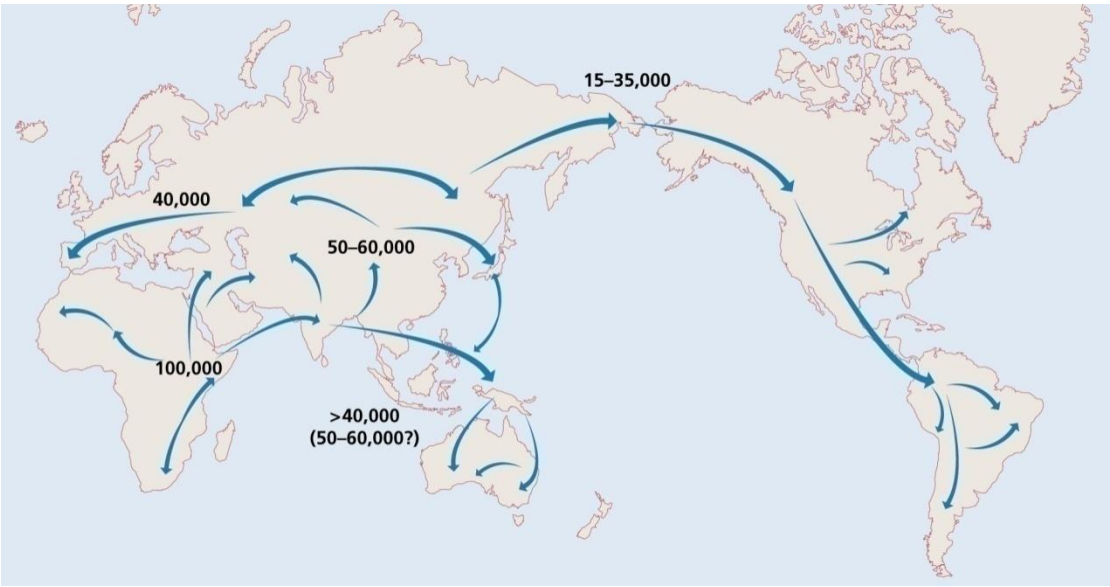


Figure 5 Possible scenario for the timings and dispersal routes of AMH's journey out of Africa. A range of expansions within Africa occurred ~100 thousand years ago, which was then followed by subsequent expansions into the rest of the world. This figure is taken from (Cavalli-Sforza and Feldman 2003).

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.

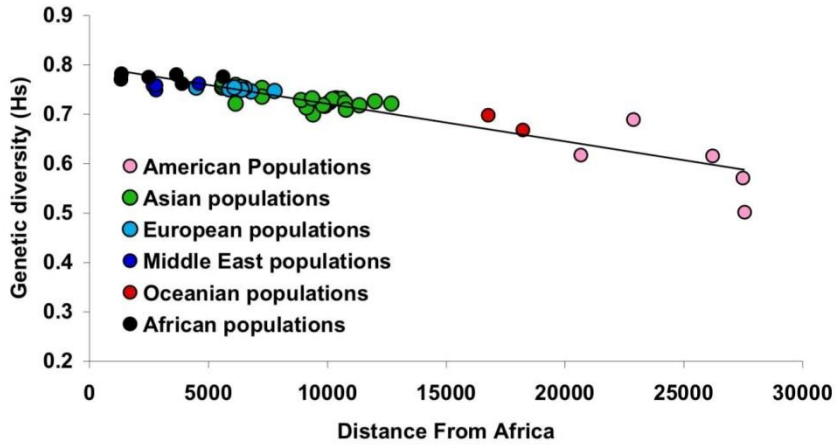


Figure 6 Relationship between genetic diversity and geographic distance from Africa for 51 distinct present-day populations. There is a decline in the genetic diversity of human populations with increasing distance from our assumed place of origin in Africa. This figure is taken from (Prugnolle et al.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

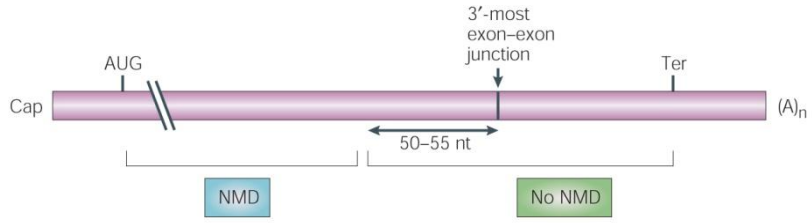


Figure 7 NMD prediction according to the "50-55 nucleotide" rule. If the PTC is located more than 50-55 nucleotides upstream of the 3'most exon-exon junction (region indicated in blue) NMD is triggered and the transcript is degraded. If the PTC is located in the last exon or less than 50-55 nucleotides 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.

2 MATERIALS AND METHODS

This chapter describes the materials used in this study and the methods of analyses that were applied to the data. The first part presents the source of the DNA samples with information on their geographical origin as well as noting the criteria applied to select the SNPs for the study. The second part describes the laboratory methods applied, primers designed and protocols that were followed. The third and final part lists the programs, databases and scripts used and describes the computational methods that were used in analysing the data – inferring the ancestral state of the alleles, predicting the protein truncation and NMD, looking at the gene ontology and applying summary statistics to search for selection.

2.1 THE DATA

2.1.1 The Samples

The samples genotyped were derived from 1,191 individuals from 56 geographically diverse populations. 1,064 samples were obtained from the Foundation Jean Dausset, the CEPH Human Genome Diversity Cell Line Panel (HGDP-CEPH) (Cann et al. 2002) and 127 unrelated individuals from the four HapMap populations – CEPH Utah residents with ancestry from northern and western Europe (CEU), Yoruba in Ibadan, Nigeria (YRI), Han Chinese in Beijing (CHB), Japanese in Tokyo (JPT) (The International HapMap Consortium 2005).

The samples used for the re-sequencing analysis were from three HapMap (23 YRI, 23 CHB, 22 CEU) and 23 individuals from one extended HapMap population , the Luhya in Webuye, Kenya (LWK). In addition, one chimpanzee (*Pan troglodytes*) sample was included as an outgroup.

All HapMap samples were purchased from the Coriell Institute for Medical Research (Camden, New Jersey, USA), the HGDP-CEPH collection (Cann et al. 2002) was kindly provided by Howard Cann (CEPH, Paris, France) and the chimpanzee

sample was purchased from the ECACC (Salisbury, Wiltshire, UK). The HGDP-CEPH samples were whole-genome amplified before use (see section 2.2.1). The HapMap samples were used as genomic DNA.

In the end, 1,151 of the original 1,191 samples were used in the final genotype analyses. A total of 40 samples were thus excluded. These included 16 samples from HGDP-CEPH which were excluded according to Rosenberg's suggestions for using standardized subsets of the original diversity panel (see Rosenberg 2006). I used the H1048 subset which contains no duplicated samples or individuals that are extremely atypical for their populations. According to this subset 18 individuals should be excluded, but two of these were not found in my dataset. The exclusion of duplicated samples followed the convention of discarding duplicates with higher identification numbers. I followed this rule except when the sample with the lower number yielded more genotype data. A further 24 samples were excluded because their genotyping failed completely. Of the 91 HapMap samples used in the re-sequencing analysis, 88 were successfully re-sequenced.

The coordinates for the HGDP-CEPH populations were obtained from the CEPH website at http://www.cephb.fr/en/hgdp/diversity.php/table.php and the locations were projected onto a map (see Figure 8A). As exact coordinates were not available for the HapMap samples, their location is not shown on this map. When displaying pie charts with allele frequencies (in chapters 3 and 4) I grouped some closely related populations together to avoid population size bias, resulting in a total of 37 populations instead of 56 (Figure 8B). Additionally, the coordinates of a few HGDP-CEPH populations (in Israel, France, Italy, and Brazil) were changed slightly so that the pie charts would not overlap and the allele frequency proportions could be easily viewed. The HapMap pie charts were inserted separately onto the map. The details of all population names are further displayed in Table 1 and a full list of all samples used is given in Appendix A (on accompanying CD).

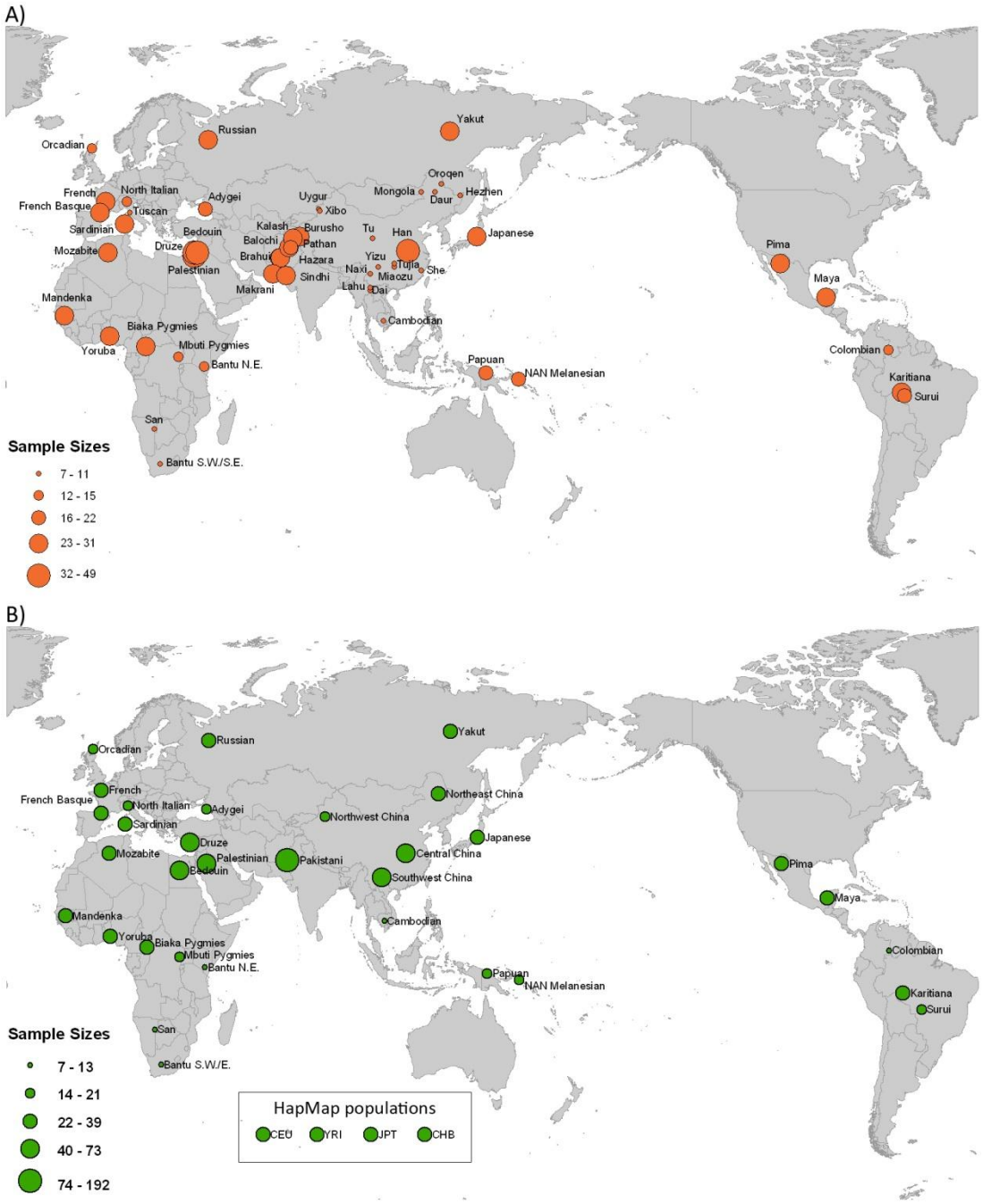


Figure 8 Population locations of genotyped samples. A) Geographical locations of the 52 HGDP-CEPH populations genotyped. The diameter of the orange circles is proportional to sample sizes. The HapMap populations are not shown. **B)** Geographical locations of the genotyped populations as they appear in allele frequency pie charts. Some related populations were clustered together to reduce population size bias, resulting in 37 populations displayed on the map. The diameter of the green circles is proportional to sample sizes. Coordinates were slightly shifted for populations too close to each other for the pie charts not to overlap. HapMap populations are inserted at the bottom of the map as they do not have geographical coordinates.

Population (N = 56)	Sample No.	Source	Geographic Origin	Population (N = 37)
Mozabite	30	HGDP-CEPH	Algeria (Mzab)	Mozabite
NAN Melanesian	19	HGDP-CEPH	Bougainville	NAN Melanesian
Karitiana	24	HGDP-CEPH	Brazil	Karitiana
Surui	21	HGDP-CEPH	Brazil	Surui
Cambodian	11	HGDP-CEPH	Cambodia	Cambodian
Biaka Pygmies	31	HGDP-CEPH	Central African Republic	Biaka Pygmy
Dai	10	HGDP-CEPH	China	Southwest Chinese
Daur	10	HGDP-CEPH	China	Northeast Chinese
Han	43	HGDP-CEPH	China	Central Chinese
Han Chinese in Beijing	32	HapMap	China	CHB
Hezhen	9	HGDP-CEPH	China	Northeast Chinese
Lahu	10	HGDP-CEPH	China	Southwest Chinese
Miaozu	10	HGDP-CEPH	China	Southwest Chinese
Mongola	10	HGDP-CEPH	China	Northeast Chinese
Naxi	10	HGDP-CEPH	China	Southwest Chinese
Oroqen	10	HGDP-CEPH	China	Northeast Chinese
She	10	HGDP-CEPH	China	Central Chinese
Tu	10	HGDP-CEPH	China	Central Chinese
Tujia	10	HGDP-CEPH	China	Central Chinese
Uygur	10	HGDP-CEPH	China	Northwest Chinese
Xibo	9	HGDP-CEPH	China	Northwest Chinese
Yizu	10	HGDP-CEPH	China	Southwest Chinese
Colombian	13	HGDP-CEPH	Colombia	Colombian
Mbuti Pygmies	15	HGDP-CEPH	Democratic Republic of Congo	Mbuti Pygmy
CEPH Utah residents with ancestry from northern and western Europe	32	HapMap	Europe	CEU
French	25	HGDP-CEPH	France	French

Table 1 Genotyped populations. Shown are the population labels as given by the source (HGDP-CEPH and HapMap) for the 56 populations, the number of samples genotyped in each population, as well as the geographical origin and a broader division of the populations (N=37). The table is sorted by geographical origin.

Population (N = 56)	Sample No.	Source	Geographic Origin	Population (N = 37)
French Basque	24	HGDP-CEPH	France	French Basque
Druze	43	HGDP-CEPH	Israel (Carmel)	Druze
Palestinian	49	HGDP-CEPH	Israel (Central)	Palestinian
Bedouin	47	HGDP-CEPH	Israel (Negev)	Bedouin
Sardinian	28	HGDP-CEPH	Italy	Sardinian
Tuscan	8	HGDP-CEPH	Italy	Italian (mainland)
North Italian	13	HGDP-CEPH	Italy (Bergamo)	Italian (mainland)
Japanese	29	HGDP-CEPH	Japan	Japanese
Japanese in Tokyo	31	HapMap	Japan	JPT
Bantu N.E.	12	HGDP-CEPH	Kenya	Bantu N.E.
Maya	24	HGDP-CEPH	Mexico	Maya
Pima	24	HGDP-CEPH	Mexico	Pima
San	7	HGDP-CEPH	Namidia	San
Papuan	17	HGDP-CEPH	New Guinea	Papuan
Yoruba	25	HGDP-CEPH	Nigeria	Yoruba
Yoruba in Ibadan	30	HapMap	Nigeria	YRI
Orcadian	15	HGDP-CEPH	Orkney Islands	Orcadian
Balochi	25	HGDP-CEPH	Pakistan	Pakistani
Brahui	25	HGDP-CEPH	Pakistan	Pakistani
Burusho	24	HGDP-CEPH	Pakistan	Pakistani
Hazara	24	HGDP-CEPH	Pakistan	Pakistani
Kalash	23	HGDP-CEPH	Pakistan	Pakistani
Makrani	25	HGDP-CEPH	Pakistan	Pakistani
Pathan	22	HGDP-CEPH	Pakistan	Pakistani
Sindhi	24	HGDP-CEPH	Pakistan	Pakistani
Russian	25	HGDP-CEPH	Russia	Russian
Adygei	17	HGDP-CEPH	Russia Caucasus	Adygei
Mandenka	24	HGDP-CEPH	Senegal	Mandenka
Yakut	25	HGDP-CEPH	Siberia	Yakut
Bantu S.W./E.	8	HGDP-CEPH	South Africa	Bantu S.W./E.

Table 1 continued

2.1.2 The SNPs

Nonsense-SNPs were identified from their annotation in dbSNP in early 2005 (build 121), resulting in a list of 1,230. In designing the project, I excluded nonsense-SNPs that were known to be incompatible with the typing method used, but ignored prior information about their frequency if it was available. Synonymous-SNPs were chosen to act as controls in this study; although not perfectly neutral they provide an approximation to neutral variants. They were selected to roughly match the sources (submitter) of the nonsense-SNPs in order to match SNPs that might have been called on the basis of poor sequencing or the use of particular populations.

Most SNP data has been obtained through various different discovery processes that often involve the discovery (ascertainment) of the SNPs in a larger sample (typically non-African) which is then followed by genotyping in a larger sample of different populations. This causes ascertainment bias in the data and often the ascertainment schemes have not been recorded systematically and thus it can be difficult to correct for this bias (discussed in Nielsen et al. 2004). However, since the nonsense-SNPs and synonymous-SNPs were chosen in the same way we expect them to be affected by the same ascertainment bias and the effect of such a bias should therefore be reduced at least when the two types of SNPs are compared.

In the end, assays were designed for 805 nonsense-SNPs and 732 synonymous-SNPs, a total of 1,536 SNPs which is the number required for one bundle of an Illumina BeadArray™. All SNPs were genotyped in the HGDP-CEPH and HapMap samples using a multiplexed genotyping assay, the GoldenGate™ assay (Fan et al. 2003). The genotyping is further described in section 2.2.3.

The genotyping results were subjected to sequential quality control filters by the Sanger Genotyping Platform Group (Team 67). Each plate contained three duplicates, and SNPs with more than 33% discrepancies between duplicates were excluded. The Gene Call (GC) score which gives the confidence of the genotype read (intensity) was then estimated. A very low value is not to be trusted. Genotypes

without call, individual genotypes with a GC score less than 0.25, assays with a median GC score lower than 0.3 and assays with less than 80% data were also discarded. 406 SNPs (181 nonsense and 225 synonymous) were excluded because they failed these quality control filters. A further 494 SNPs (387 nonsense and 107 synonymous) were excluded by me as they were monomorphic in the combined samples. The SNP had to show variation in at least one individual to be kept. Lastly, I excluded 183 SNPs (68 nonsense and 115 synonymous) that did not pass my manual reassessment of gene annotation incorporating information that became available after the assays were designed. For manual assessment I looked to see whether the nonsense-SNP genes overlapped with Vega pseudogenes (manually annotated and curated by the international vertebrate genome annotation (VEGA) project)(Ashurst et al. 2005) and excluded them if they were found to do so. I used the Tblastx tool to search for the ORF of the sequence surrounding SNPs that had "Stop lost" listed as a consequence and removed those where the ancestral state (chimpanzee) was found to be the PTC and the derived state (human) was found to be a read through of the protein. One SNP, in the *PCDH11XY*, was excluded as the variation observed was found be due to variation between the X and Y chromosomes and not because of polymorphism of the SNP. In addition I excluded those synonymous-SNPs that were found to be intronic. As derived allele information is essential to the analysis, I also excluded SNPs were the ancestral state could not be inferred (1 nonsense- and 8 synonymous SNPs). My final dataset consisted of 452 polymorphic SNPs, 169 nonsense SNPs in 167 genes and 283 synonymous SNPs, and this was used in subsequent analyses. Table 2 lists the number of SNPs kept after each of the above filtering steps.

SNP Status	Nonsense	Synonymous	Total
Original number of SNPs	805	731	1536
Successfully genotyped	624	506	1130
Polymorphic in our dataset	237	399	636
Passed manual assessment*	169	283	452

Table 2 The number of SNPs kept in the dataset after the various filtering stages. *See description in text.

2.2 Laboratory Methods and Protocols

2.2.1 Whole Genome Amplification

The samples from the HGDP-CEPH panel had low amounts of DNA and were thus subjected to whole-genome-amplification (WGA) on 11/10/2004 by Yali Xue at the WTSI using the GenomiPhi DNA Amplification Kit by GE Healthcare (formerly Amersham Bioscience) and the protocol was performed according to the manufacturer's guidelines. The resulting stock was then stored at -20°C as there is some indication that WGA DNA degradation in time is temperature dependent.

2.2.2 DNA Quantitation

The Illumina GoldenGate™ assay for genotyping required 22µl of ≥50 ng/µl DNA. I performed quantiation on the DNA samples with the Quant-iT™ PicoGreen® dsDNA Assay Kit from Molecular Probes (Invitrogen) and diluted the samples accordingly to ~50ng/µl. The assay was performed according to the manufacturer's guidelines, with the following modification. For the DNA standard curve the Lambda DNA standard was diluted to 5 µg/ml, instead of to 2 µg/ml.

The assay plates were read and fluorescence was measured using a Cytofluor 4000 Fluorescence Plate Reader (MTX Lab Systems, Inc.) with excitation light and filter settings set for excitation at 480 nm and emission at 520 nm. Using the DNA standards, the amount of DNA versus fluorescence intensity was plotted and a line was fitted to the points. This standard curve was then used to determine the amount of DNA from the fluorescence intensity for each sample.

2.2.3 Genotyping

Once the DNA samples had been diluted to a concentration of ≥50 ng/µl I submitted them to the Sanger Genotyping Platform Group (Team 67). 1,536 SNPs were genotyped in 1,191 samples (but see later sample and SNP exclusions in sections 2.1.1 and 2.1.2) with the GoldenGate™ assay protocol (Illumina) (Fan et al. 2003)

according to the manufacturer's instructions. The GoldenGate™ assay workflow is displayed and described in Figure 9.

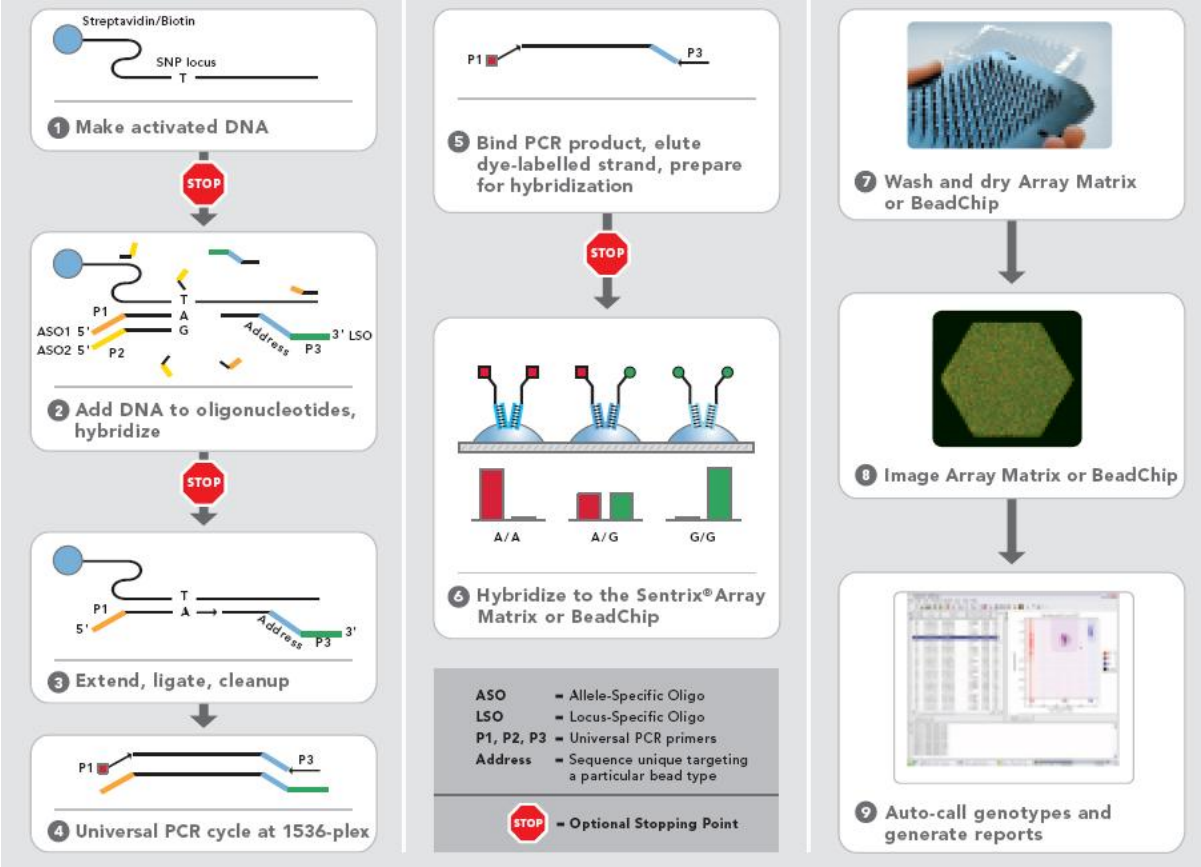


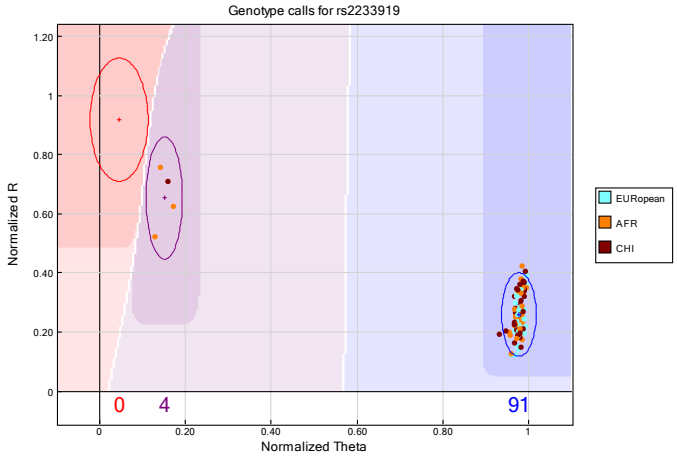
Figure 9 GoldenGate™ Assay Overview. Step 1: The DNA sample is activated for binding to paramagnetic particles. **Step 2:** Assay oligonucleotides (oligos), hybridization buffer, and paramagnetic particles are then combined with the activated DNA. Three oligos are designed for each SNP locus. Two oligos are specific to each allele of the SNP site (Allele-Specific Oligos, ASOs). A third oligo that hybridizes several bases downstream from the SNP site is the Locus-Specific Oligo (LSO). All three oligo sequences contain regions of genomic complementarity and universal PCR primer sites; the LSO also contains a unique address sequence that targets a particular bead type. The hybridization is followed by several wash steps. **Step 3:** Extension of the appropriate ASO and ligation of the extended product to the LSO joins information about the genotype present at the SNP site to address the sequence on the LSO. **Step 4:** These joined, full-length products thus provide a template for PCR using universal PCR primers P1, P2 and P3. **Step 5:** Universal PCR primers P1 and P2 are Cy3- and Cy5-labeled. **Step 6:** After downstream processing, the single-stranded, dye-labeled DNAs are hybridized to their compliment bead type through their unique address sequences. **Step 7:** Hybridization of the GoldenGate assay products onto the BeadChip allows for the separation of the assay products in solution, onto a solid surface for individual SNP genotype readout. **Step 8:** After hybridization, the BeadArray reader is used to analyze the florescent signal on the BeadChip. **Step 9:** GeneCall software is then used for automated genotype clustering and calling. Figure and assay description were obtained from http://www.illumina.com/

The Illumina primer sequences are given in Appendix B.1. (on accompanying CD). The genotypes were inferred from genotype clusters in GeneCall (from Illumina) and the quality control filters have already been described in section 2.1.2.

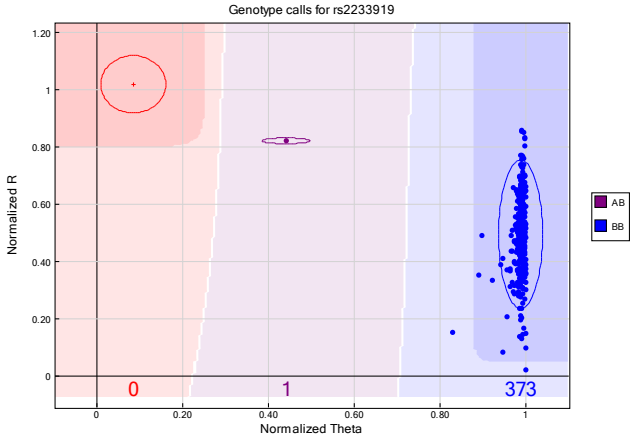
2.2.3.1 Problems with Genotype Clusters

At the WTSI it is customary for the genotype clustering and quality control for large scale surveys to be handled by Team 67, as was the case here. However, at a later stage in the analyses I detected a number of cases causing a deviation from Hardy-Weinberg Equilibrium (HWE) and when I looked back at the raw data I noticed some odd genotype calls. This problem was subsequently resolved and will now be explained with the example of one SNP (rs2233919). The alleles observed at this SNP are A/G, and our samples showed the following numbers of genotypes – 29 AA, 5 AG and 1105 GG – which deviates significantly from HWE (chi-square, $P<0.0001$). I then looked at the genotype clusters as they appear in GeneCall (Figure 10). At this point it should be noted that plates containing the samples were submitted in batches to Team 67 for genotyping, starting with plate 1 (containing only HapMap samples), then plates 2-5 (containing only HGDP-CEPH samples) were submitted and finally plates 6-13 (containing both HapMap and HGDP-CEPH samples). In Figure 10 each dot represents a sample and the genotype clusters are revealed with different colours, where the pink area designates AA homozygotes, the purple area the heterozygotes (AG) and the blue represents clusters of GG homozygotes. I will not go into the details of the clustering method performed, but note that the clusters observed in Figure 10A and B returned the expected genotypes in our dataset, e.g. in Figure 10A you see 4 dots in the purple area and these corresponded to the 4 (out of 5) heterozygotes observed for this SNP.

A)



B)



C)

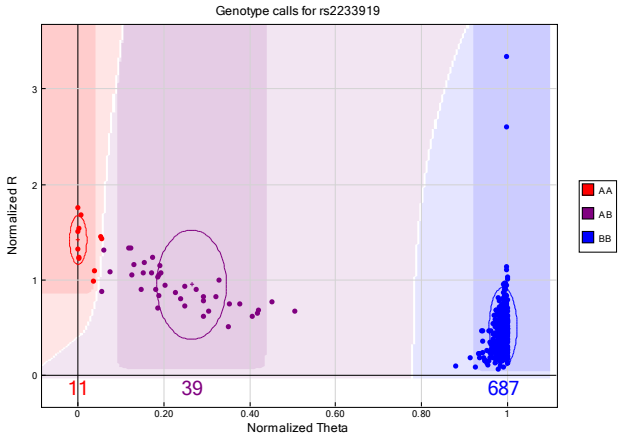


Figure 10 Genotype clusters for SNP rs2233919 as displayed in GeneCall. A) Plate 1 (containing only HapMap samples) **B)** Plates 2-5 (containing only HGDP-CEPH samples) **C)** Plates 6-13 (which contained samples from both HGDP-CEPH and HapMap). Each cluster has a plus sign to indicate the mean of the data.

However, the clustering in Figure 10C looked odd, as I only had one more heterozygote reported for this SNP although the purple cluster has a large number of dots filling the purple area. When I extracted the sample names for these dots, one was indeed the expected heterozygote, but the rest were reported as AA homozygotes which should then be represented in the pink area. After various discussions with several members of Team 67 we came to the conclusion that the problem came from analysing the clusters of both HapMap (genomic) and HGDP-CEPH (WGA) samples together and that genomic DNA and WGA DNA should not be analysed together because of different properties. As a consequence, the clustering and subsequent quality control was redone for the whole data set, by analysing the HapMap and HGDP-CEPH samples separately. Unfortunately, the SNP given as an example above was consequently excluded by the quality control filters and so I am unable to represent its new genotype calls here.

2.2.3.2 Additional Quality Control

When we got the new genotype results back I performed additional quality controls of my own to investigate the genotype calls further. I checked for deviation from HWE for each SNP in the individual populations and found none that deviated from HWE. I also decided to compare my genotyping results to the publicly available genotypes of the HapMap. I used the SNP IDs (rs numbers) of my 452 SNPs and extracted their genotypes for the four HapMap populations (CEU, YRI, CHB and JPT) using the HapMart tool from the HapMap website at http://hapmart.hapmap.org/BioMart/martview and then compared those genotype results to the genotypes of my typed HapMap samples. 77% of my SNPs were included in HapMap Phase II, and I only found inconsistencies for 0.692% of the genotype comparisons (i.e. ~seven inconsistencies per 1000 genotypes SNPs), which is similar to the reproducibility of the HapMap results and other comparisons with HapMap data carried out in our team, and therefore acceptable. Thus, I conclude

that the quality control filters were satisfactory and that the genotype calls are to be trusted.

In the end, 452 SNPs were successfully genotyped in 1,151 samples and the whole dataset is available as a tab delimited text file on the accompanying CD (Appendix D).

2.2.4 Resequencing

In addition to genotyping 1,536 SNPs, we decided to follow up on two nonsense-SNPs, rs1343879 in *MAGEE2* and rs16982743 in *SIGLEC12*, which were observed as outliers in the nonsense-SNP data set, by resequencing the genes. We also followed up on rs497116 in *CASP12*, but as the re-sequencing of *CASP12* was not performed for this project, the methods used are described elsewhere (Xue et al. 2006).

All primers were ordered from Sigma-Genosys and their sequence is given in Appendix B. The machine used for all PCRs was Alpha™ Unit Bloc Assembly for DNA Engine System, ALS1296, BIO-RAD.

2.2.4.1 Long-Range Polymerase Chain Reaction

The regions we chose to analyse were around 13 kb in length for each gene with the nonsense-SNP in the middle. Primers were designed for human and chimpanzee with Primer3 (Rozen and Skaletsky 2000) and a custom Perl script, pcr_overlap.pl (see description in section 2.3.3.1), and were selected to amplify two long polymerase chain reaction (PCR) fragments, ~6.5 kb, for each gene. The sequences of the long-range PCR primers for *MAGEE2* and *SIGLEC12* are given in Appendix B.2.

The Platinum® Taq DNA polymerase High Fidelity (Invitrogen) was used for all long PCRs. A long PCR reaction mastermix sufficient for the number of reactions to be carried out was prepared and the recipe for one reaction is given in Table 3.

Reagent	Volume (µl) x1
ddH_2O	8.96
10X High Fidelity PCR Buffer	1.50
$MgSO_4$ (50 mM)	0.60
dNTPs (25mM each)	0.12
Forward Primers (10 µM)	0.60
Reverse Primers (10 µM)	0.60
Platinum Taq High Fidelity (5 U)	0.12
Total volume added to plate	**12.50**
DNA template (50 ng/µl)	2.50
Total volume	**15.00**

Table 3 Recipe for amplification of long PCR products.

The following PCR cycle conditions were used for the long PCR reactions:

94°C for 2min

94°C for 30sec
68°C for 30sec (decrease 0.5C/cycle) } 15 cycles
68°C for 6min

94°C for 30sec
58°C for 30sec } 20cycles
68°C for 6min

68°C 7min
4°C forever

2.2.4.2 Nested PCR

In order to get good quality sequence traces, it is better to re-amplify segments of the long PCR product with overlaps rather than sequence the long PCR product directly. Therefore, a set of nested primers was designed using a perl script, pcroverlap.pl (see description in section 2.3.3.1). The primers were conditioned to amplify nested PCR products of 500x(1±15%) bp length overlapping by 240x(1±30%) bp. The sequences of the nested primers for *MAGEE2* and *SIGLEC12* are listed in Appendix B.3.

The Platinum® Taq DNA Polymerase (Invitrogen) was used for the nested PCRs. A nested PCR reaction mastermix sufficient for the number of reactions to be carried out was prepared and the recipe for one reaction is given in Table 4.

Reagent	Volume (µl) x1
ddH_2O	9.65
Platinum Buffer 10x	1.50
$MgCl_2$ (50 mM)	0.48
dNTPs (25mM each)	0.12
Forward primers (100uM)	0.10
F&R primers (100uM)	0.10
Platinum Taq (5 U)	0.05
Total volume added to plate	**12.00**
400x diluted long PCR products	3.00
Total volume	**15.0**

Table 4 Recipe for amplification of nested PCR products.

The following PCR cycle conditions were used for the nested PCR reactions:

94°C for 15min

94°C for 45sec
61°C for 45sec
72°C for 45sec
} 15 cycles

72°C for 7min
4°C forever

2.2.4.3 Electrophoresis

Products were analysed by electrophoresis on a 1.5% agarose gel containing ethidium bromide to check that a band of the expected size was present at an adequate concentration. ~20% of each plate was checked.

2.2.4.4 PCR-Product Purification

The PCR-products were purified before they were sent off for re-sequencing. A mastermix of Shrimp Alkaline Phosphatase (USB) and Exonuclease I (USB) sufficient for the number of reactions to be cleaned was prepared and the recipe for one reaction is given in Table 5.

Reagent	Volume (µl) x1
ddH_2O	1.380
ExoSAP buffer*	0.670
Exonuclase I (20U/ul)	0.033
Shrimp Alkaline Phosphatase (1U/ul)	0.670
Total volume added to plate	2.000
PCR product	8.000
Total volume	10.00

Table 5 Recipe for one reaction of mastermix required for PCR-product clean-up. *ExoSAP buffer: 1M Tris (PH8.0) 20ml, 1M $MgCl_2$, ddH_2O 70ml

The following PCR conditions were used for the clean-up of PCR products:

Step 1. Incubate at 37°C for 1 hour

Step 2. 80°C for 20 min

Step 3. 4°C forever.

Products were sequenced on both strands by the Sanger Large Scale Sequencing Pipeline using BigDye Sanger sequencing technology with an 3730 *xl* DNA Analyzer (Applied Biosystems).

2.3 Computational Methods

2.3.1 Programs and Databases

The complete data set was stored in a Microsoft Access database and was handled and queried using SQL query language implemented therein. Many online databases enabled us to browse, extract data and use various tools supplied. The most commonly used were NCBI, Ensembl, HapMap, UCSC Genome Browser (Kent et al. 2002), The Human Gene Mutation Database, SNP2NMD (Han et al. 2007) and DAVID (Dennis et al. 2003); the usage of some of these is described in other sections.

In order to visualise the geographical distribution of alleles, the geographical coordinates of the sampled individuals were imported into the ESRI ArcGIS 8.2 software (projected with the Gall Stereographic coordinate system with the central meridian set at 145) and pie charts were then produced from allele frequencies.

Basic statistical analyses were performed in Microsoft Excel, Minitab® (release 14) and in R. To test for the significance of the differences in the distribution of values observed for the nonsense-SNPs versus the synonymous-SNPs we applied the Kolmogorov-Smirnov test with an online calculator, http://www.physics.csbsju.edu/stats/KS-test.n.plot_form.html. F_{ST} was calculated using the R package HIERFSTAT (Goudet 2005) for autosomal SNPs and in Arlequin (Schneider et al. 2000) for X-chromosomal SNPs. Pairwise difference was calculated using Arlequin (Schneider et al. 2000). Calculations of summary statistics were performed in DnaSP (Rozas et al. 2003). The LRH-test (Sabeti et al. 2002) was performed for the whole SNP dataset (with extra controls) with a java version of Sweep™ and individual SNPs were visualised in Haplotter (Voight et al. 2006). Haplotypes were inferred using PHASE 2.1 (Stephens and Donnelly 2003; Stephens et al. 2001), and median-joining networks (Bandelt et al. 1999) were constructed with Network (http://www.fluxus-engineering.com/sharenet.htm). The use of these programs is further described in the appropriate sections in 2.3.8.

2.3.2 Detection of Variants

Potential variable positions in sequence traces were flagged by Mutation Surveyor® v. 2.0. (SoftGenetics, LLC., PA, USA) and checked manually. A Perl script, merge_sts.pl, was then used to check the SNP calling consistency between the overlapping sequence tag sites (STSs) as well as the four duplicates (see description in section 2.3.3.1). Unfortunately, at this stage it was apparent that we could not use the resequenced data from the *SIGLEC12* gene as the sequence traces were unreadable and full of complications. This gene was thus not analysed in the end.

2.3.3 Programming Scripts

Several custom computer scripts written in the Perl and Java programming languages were used. All input files were tab delimited. The scripts are found on the

CD accompanying this thesis (Appendix C), with a detailed description of the input files and command lines required.

2.3.3.1 Perl Scripts

pcroverlap.pl: This program takes large tracts of sequence data in FASTA file format, and produces PCR products in overlapping segments to span the entire region. It divides up the given sequence, based on the user's criteria for PCR product size (e.g. 500-700 bp) and overlap between adjacent segments (e.g. 200-400 bp), and passes these choices to the PCR primer-selecting program Primer3 (Rozen and Skaletsky 2000). Primer3 then chooses a set of nested primers based on specific selection criteria. The output file is a list of nested primers consisting of the primer sequence, melting temperature (Primer3 calculated), "quality" of nested primers (lower is better; Primer3 calculated), primer positions, primer lengths, PCR product length, and amount of overlap between adjacent fragments. The script was originally obtained from the SeattleSNPs website (http://droog.gs.washington.edu/PCR-Overlap.html) and was modified slightly by Yuan Chen & Cara Woodwark.

hgdp2sweep.pl: This program takes a genotype file as input and gives you as output the .snp and .many input files needed to run Sweep™. The PHASE program (Stephens and Donnelly 2003; Stephens et al. 2001) needs to be installed as this script will take the genotype input file, run PHASE to infer the haplotypes, and then use the phased data to create the Sweep input file. The input file should be space delimited and contain the following information: SNP id, chromosome, position and genotypes for all samples. This script was originally created by Yuan Chen and modified by myself.

create_fstat_input.pl: This program takes a tab delimited text file with the following information: SNP name, SNP number, sample name, population number, Genotype Code (i.e. 11 = homozygote for first allele, 22 = homozygote for second

allele and 12 = heterozygote) and converts it into the file input required by HIERFSTAT (Goudet 2005). This script was created by Jim Stalker.

merge_sts.pl: This script was used to check the SNP calling consistency between the overlapping STSs as well as the four duplicates. When the callings were consistent, the script joined the different segments together to reconstruct the whole resequenced region. It then created a table with the variable positions listed for each sample (a SNP table) This script was created by Ni Huang.

snptab2phase.pl: This script converts the SNP table produced by merge_sts.pl into the PHASE input file format population by population. Additionally, it requires a file with sample id for each population. This script was created by Ni Huang.

phase2fasta.pl: This script converts the PHASE output files from different populations into FASTA format and converts them into a format that can be read into the DNaSP program for the neutralisty tests. A file containing all the PHASE output file names is needed. This script was created by Cara Woodwark.

phase2network.pl: This script converts the PHASE output files from different populations into .rdf format and converts them into a format required for the Network program in order to create median-joining networks. A file with the all PHASE output file name list is needed. This script was created by Ni Huang.

2.3.3.2 Java Scripts

InputFileTransformer.java: This program will convert a crosstab table created in Access with homozygote and heterozygote codes (00, 11 and 01) into the format required in Arlequin (Schneider et al. 2000) to calculate the number of pairwise differences. This script was created by Bjarki Holm.

DelimitedFileTransformer.java: This program was designed to convert the HapMart output from HapMap so that it would correspond to the format of our genotyping results in order to make the comparison between the two easier. This script was created by Bjarki Holm.

SweepFileConversion.java: This program collects the HapMap phased data from a URL for a region of choice and outputs the .snp and .many files required by Sweep™ for each SNP and each HapMap population. This script was created by Bjarki Holm.

2.3.4 Inferring the Ancestral State

In order to calculate the derived allele frequency (DAF), we needed to know the ancestral state of each allele. The chimpanzee (*Pan troglodytes*) base was primarily used as the ancestral state, but when the chimp sequence was not available or differed from both the observed human alleles, we accepted sequence from other primates (*Macaca mulatta* or *Lagothrix lagotricha*). The derived allele was then defined as the other observed human allele.

We used the Table Browser on the UCSC Genome Browser website (http://genome.ucsc.edu/cgi-bin/hgTables) and retrieved the ancestral allele for ~98% (445 SNPs) from the "snp126OrthoPanTro2RheMac2" table. We then looked manually for the ancestral state of the missing 2% (8 SNPs). We obtained FASTA sequences surrounding the SNPs and used the NCBI Blastn algorithm to find the best hit with a primate reference sequence and thereby identified the ancestral allele for 6 of these at the appropriate position.

The derived allele frequency was obtained by direct allele counting and a Kolmogorov-Smirnov test was used to evaluate the difference between the distributions of nonsense- and synonymous-SNPs.

2.3.5 Predicted Truncations and Calculations of NMD

In order to visualize the predicted effect of these nonsense-SNPs on the gene product, we first estimated the proportion of protein truncation each SNP would cause. 112 genes bearing nonsense-SNPs were found to code for a single transcript. The remaining 57 nonsense-SNPs were found in genes undergoing alternative splicing and were reported in more than one transcript. For such SNPs we used the

transcript showing the largest truncation. The truncation was calculated as a percentage of the ancestral sequence ORF length (100-(SNP protein position/protein length*100)).

The nonsense SNP could lead to a truncated protein with an altered function but if it is located more than 50-55 nucleotides upstream of the 3'-most exon-exon junction the transcript will be eliminated by NMD (Maquat 2004). In order to assess whether our nonsense-SNPs were likely to trigger NMD we used the SNP2NMD database (Han et al. 2007) available from http://bioportal.kobic.re.kr/SNP2NMD. This database contains human nonsense-SNPs with an estimate of whether or not NMD is expected to be triggered according to the 50-55 nucleotide rule. 107 (~63%) of our nonsense-SNP were in SNP2NMD and we used the default setting of the "NMD distance" (distance between a SNP and the 3'-most exon-exon junction) to be >50 nucleotides for the NMD pathway to be triggered. As the transcripts used in SNP2NMD were obtained from different sources from our data, we applied the same rule as mentioned above and selected the transcript with the maximum truncation when having to choose from multiple transcripts. For the remaining 62 (~37%) SNPs missing from SNP2NMD we extracted information on the location of the nonsense- SNP with respect to exon-intron boundaries from Ensembl (release 37 and 43) and calculated the prediction for NMD manually.

2.3.6 Gene Expression

In collaboration with Barbara Stranger and Manolis Dermitzakis, of Team 16 (Population and Comparative Genomics) at the WTSI, we used their available expression data to test the association between nonsense-SNP genotypes and expression levels. Gene expression quantification and normalization had already been performed by Barbara Stranger *et al* (Stranger and Dermitzakis 2006; Stranger et al. 2007b)

Gene expression data were obtained for approximately 48,000 transcripts, including a subset of 14,456 probes (13,643 unique autosomal genes) that were

highly variable among lymphoblastoid cell lines of the 210 unrelated HapMap individuals (Stranger et al. 2007b). Hybridization intensity values were normalized on a log2 scale using a quantile normalization method (Kuhn et al. 2004) across all replicates of a single individual followed by a median normalization method across all 210 individuals. A subset of 14,456 probes (13,643 unique autosomal genes) that were highly variable within and between populations was selected from the 47,294 probes on the array, and were used for the analysis. A detailed description can be found in Stranger *et al* (2007b).

We first attempted to test our set of 169 nonsense-SNPs for association with expression of these variable genes, but found that only 57 of the SNPs mapped within the genes corresponding to the 14,456 probes, and of these, only 19 were polymorphic and genotyped in the HapMap (The International HapMap Consortium 2005). This gave us little power to draw any conclusions and we thus resorted to using all available nonsense-SNPs (dbSNP126) which gave us a starting dataset of 1,624 SNPs instead of our original 169. In the end, 588 of these had been typed in HapMap and 105 of those could be mapped within genes corresponding to the expression probes exhibiting variable gene expression.

We tested the nonsense-SNP genotype for association with expression levels of the gene by using an additive linear regression model (Stranger et al. 2005; Stranger et al. 2007a; Stranger et al. 2007b) applied to each population separately. Our association analysis employed: 1) nonsense-SNP genotypes for the unrelated individuals of each HapMap population (MAF<0.05) from the HapMap phase II map for each population (version 21, NCBI Build 35) and 2) normalized log2 quantitative gene expression measurements for the 210 unrelated individuals from the original four HapMap populations (60 CEU 45 CHB, 45 JPT, 60 YRI).

To assess the significance of association between nonsense-SNP genotypes and expression variation of the gene harbouring the nonsense-SNP, we performed 10,000 permutations of each expression phenotype relative to the genotypes (Stranger et al. 2007b). An association to gene expression was considered significant if the nominal

p-value from the linear regression test was lower than the 0.01 tail of the distribution of the minimal p-values (among all comparisons for a given gene) from each of the 10,000 permutations of the expression phenotypes. For genes containing more than one nonsense-SNP, the most stringent permuted p-value was retained.

2.3.7 Gene Ontology Term Enrichment Analysis

To find out if the set of genes containing nonsense-SNPs have an overrepresentation of a particular molecular function (MF) or biological process (BP), their relevant gene ontology (GO) (Ashburner et al. 2000) terms were identified. We performed the GO term enrichment analysis with the DAVID chart analysis tool in DAVID (Dennis et al. 2003) (http://david.abcc.ncifcrf.gov/summary.jsp, 26/05/08). All available GO terms were used and all human genes (implemented in DAVID) were defined as the background. Ensembl gene IDs were collected for each of the 169 nonsense-SNPs (167 genes) with the BioMart query system (http://www.ensembl.org/biomart/index.html, 26/05/2008) and these were used as input for the enrichment analysis. P-values were calculated by the EASE score which is a modified conservative adjustment of the one-tailed Fisher Exact test (Hosack et al. 2003) and is implemented in DAVID. Terms with values below 0.05 were considered to be enriched. While a multiple correction is often applied for these tests, the authors of DAVID attest that it will be too conservative on the cost of the biological importance (revealed in a personal communication through their website). Thus, while the Bonferroni correction is given with our results, it should not be taken too seriously. Of the total 167 genes analysed, 71 were not included in the output for BP and 88 for MF. For the 71 (BP) and 88 (MF) missing, 26 (BP) and 59 (MF) had GO terms associated with the genes but the terms did not pass the filter of the EASE score (enrichment analysis), while 45 (BP) and 29 (MF) did not have any GO annotation because the functional annotation of the human genome is incomplete.

2.3.8 Population Genetic Calculations

2.3.8.1 Population Differentiation Calculations (FST)

*F*ST was used as a measurement of population differentiation. *F*ST values were calculated by conventional F-statistic methods with the HIERFSTAT (Goudet 2005) package for R using the *varcomp* function to calculate the *F*ST (theta) from Weir and Cockerham (1984). This F-statistic uses the allele frequencies to quantify the proportion of the total variance among the human populations. *F*ST values were calculated for each SNP across the 37 populations (see division in Table 1). The Kolmogorov-Smirnov test was used to assess whether there was a significant difference between the distributions of nonsense- and synonymous-SNPs. For comparison with empirical data we downloaded the genotypes for the HapMap phase II SNPs and for a set of 650K publicly available SNPs genotyped in the HGDP-CEPH populations and calculated their *F*ST values to find out if our SNPs were significant outliers (i.e. lying above the 95th or 99th percentiles). The values calculated for the HGDP-CEPH were calculated from the 32 HGDP-CEPH populations as well as for the combination of those 32 populations into five major groups to match the K=5 division in Rosenberg (2002).

Traditionally, the range of *F*ST is between 0 and 1, where 0 would imply no differentiation between populations and 1 complete differentiation. However, it is possible for the unbiased estimate of *F*ST to give negative values. When this occurred, we assigned negative values of *F*ST to zero as suggested by Nei (1987).

2.3.8.2 Heterozygosity

Nei's measure of heterozygosity (Nei 1987), the probability that any two randomly chosen samples from a population are the same, was calculated for each SNP by:

$$H = \frac{n}{n-1}\left(1 - \sum_{i=1}^{k} p_i^2\right)$$

Where n is the number of alleles, k is the number of haplotypes and p_i is the frequency of the ith haplotype.

2.3.8.3 Pairwise Differences

To estimate how much human individuals differ with respect to the nonsense-SNPs we calculated the mean number of pairwise differences as implemented in Arlequin (Schneider et al. 2000).

2.3.8.4 Long-Range Haplotype Test

To gain a better insight into the possible action of natural selection, we applied the REHH test (Sabeti et al. 2002). This test has been implemented in the Sweep™ program which requires phased haplotype data as input and analyses haplotype structure in the genome by determining the frequency and long-range LD for each allele. The method uses LD to measure the association between a single allele at one locus with multiple loci at various distances (see 1.3.1.6). We identified our nonsense-SNPs as the so-called "core" haplotype (SNP) and then increasingly distant SNPs were added to quantify the decay of LD from the core. The assumption is that a positively selected SNP will be found at a high frequency on an unusually long haplotype.

We used the SweepFileConversion.java programme to collect the phased haplotypes from the HapMap populations (CEU, YRI, CHB+JPT) and to convert them into the format required to run Sweep (see section 2.3.3.2). We caution that the CHB+JPT phased haplotypes were later withdrawn from the HapMap as they were under scrutiny and were not available again to use in time for this thesis.

We used the Phase II data (Build 36) which contained 131 out of the 169 nonsense-SNPs. We chose to use Build 36 as it contained a higher number of our SNPs than did Build 35, 131 compared to 106. As the current version of Sweep will only accept coordinates from a Build as high as 35, we used Build 35 coordinates for the Build 36 SNPs when available, and collected the coordinates for the 25 SNPs present in Build 36 but not in 35 manually using the Ensembl Genome Browser archive (Ensembl release 42).

For each of the 131 nonsense-SNPs genotyped in HapMap we chose to use a 100 kb region on each side of the SNP to infer the haplotypes. In addition, we chose 30 ENCODE random regions, which are assumed to be neutral, to act as controls. The coordinates of these were obtained from the UCSC Genome Browser. Each ENCODE region was roughly ~500 kb in length. REHH was calculated with the default setting of a 0.04 marker breakdown from the core SNP.

To evaluate whether or not our nonsense-SNPs found at high frequencies with unusually extended haplotypes were significant, we plotted the SNP frequency against its REHH value for both the nonsense-SNPs and the ENCODE SNPs (used as empirical controls), calculated the 95th and 99th percentiles, and considered a nonsense-SNP significant if it was above those.

2.3.9 Neutrality Tests

Two genes (*MAGEE2* and *SIGLEC12*) were re-sequenced, but only the *MAGEE2* sequence was of good enough quality to be further analysed (see explanation in section 2.3.2). We used DnaSP (Rozas et al. 2003) to calculate traditional neutrality tests (discussed in section 1.3.1.2). These included Tajima's *D* (Tajima 1989c), Fu and Li's *D*, *D**, *F* and *F** (Fu and Li 1993), Fu's F_S (Fu 1997) and Fay and Wu's *H* (Fay and Wu 2000). Null distributions were obtained by the custom modified ms program (Hudson 2002) incorporating the best-fit demographic model (Schaffner et al. 2005).

2.3.10 Median-Joining Network

Haplotypes for the resequenced data were inferred using PHASE 2.1 (Stephens and Donnelly 2003; Stephens et al. 2001). Median-joining networks (Bandelt et al. 1999) were constructed from the inferred haplotypes with Network (http://www.fluxus-engineering.com/sharenet.htm).

3 NONSENSE-SNPS IN THE HUMAN GENOME

Nonsense-SNPs introduce premature termination codons into genes, and can result in the absence of a gene product or a truncated and potentially harmful protein, so are often considered disadvantageous and associated with disease susceptibility. As such, we might expect the disrupted allele to be rare and, in healthy people, observed only in a heterozygous state. However, the "less-is-more" hypothesis proposes that gene loss could be a common mechanism for adaptive evolutionary change (Olson 1999).

The main aim of this thesis was to test this idea of advantageous gene loss by genotyping a large set of nonsense-SNPs in a number of populations to reveal which evolutionary forces are acting on these SNPs. This chapter describes the results from this genome-wide survey, starting with the frequencies observed in the world-wide population samples, moving on to the consequences these SNPs could be having on the gene product and finally evaluating the selective forces in play.

3.1 RESULTS

3.1.1 The Nonsense in Our Genome

We identified 167 genes containing 169 nonsense-SNPs that were polymorphic in our dataset of world-wide human samples, of which only eight genes were found in the Human Gene Mutation Database (HGMD March 2008; www.hgmd.org) of mutations associated with human inherited disease (Stenson et al. 2003). Two genes, *CDKL1* and *FMO2*, were found with two nonsense-SNPs each (*CDKL1* with rs11570829 and rs7148089; *FMO2* with rs2020866 and rs6661174) and might therefore be suspected to be pseudogenes. *CDKL1* is a cyclin-dependent kinase-like 1 (CDC2-related kinase), and according to the Entrez Gene database (http://www.ncbi.nlm.nih.gov/sites/entrez) several alternatively spliced variants have been identified (some of which differ by truncation of the 5' end and others by

a truncation of the 3' end) but their full-length nature has not been determined and it is unclear whether or not the gene is functional. A nonsense mutation (rs6661174 in our data) has previously been reported in the *FMO2* gene, a flavin-containing monooxygenase, resulting in a truncated and catalytically inactive polypeptide (Dolphin et al. 1998) which is the derived state and is nearly fixed in the human population (Veeramah et al. 2008). As these two genes were not found to overlap with the Vega set of pseudogenes (see section 2.1.2 in Materials and Methods) and the nonsense-SNPs are polymorphic in our samples we do not consider them to be pseudogenes. Therefore, these four SNPs in two genes have been kept and are included in the results presented here.

Hereafter we will refer to the disrupting allele as the "stop allele" and the non-disrupting allele as the "normal allele". Genotyping revealed that on average the individuals in our samples have ~14 stop/stop homozygous SNPs and ~18 stop/normal heterozygous SNPs in their genome, a total of ~46 stop alleles in their diploid genome or ~23 per haploid genome. As our data is based on previously ascertained SNPs this is likely to be a minimum estimate of the actual value. The issues of ascertainment bias will be discussed in more detail in section 3.2.1. Furthermore, these individuals were found to differ on average by 24 genes per diploid genome because of nonsense-SNPs.

While the nonsense-SNPs analysed here are only a fraction of all nonsense-SNPs reported in the human genome (see Figure 11), we observed that their distribution around the genome appears random and they can thus be thought to represent nonsense-SNPs as a class.

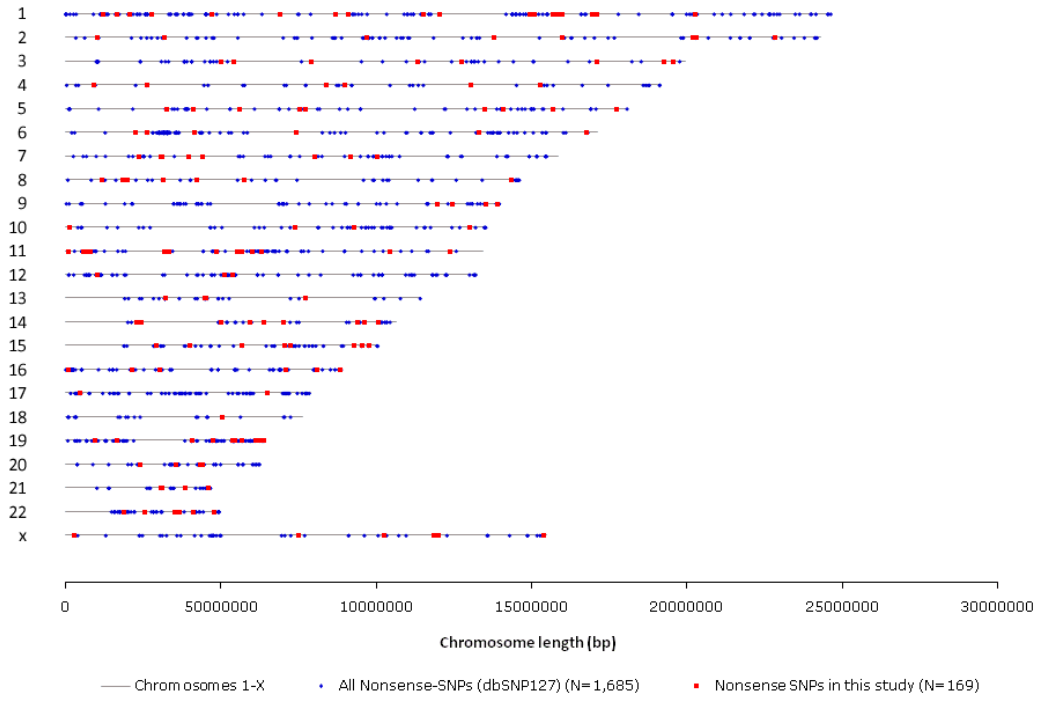


Figure 11 Genome-wide distribution of nonsense-SNPs on chromosomes one to X in the human genome. The nonsense-SNPs used in this study are displayed in red and all nonsense-SNPs reported in the human genome (dbSNP127) are shown for comparison in blue.

3.1.1.1 The Derived Allele Frequency Spectrum

A derived allele might reach a high frequency in a population if it becomes more advantageous than the ancestral allele and a high derived allele frequency can thus be a sign of positive selection and/or genetic hitchhiking. In contrast, a low derived allele frequency (DAF) could indicate negative selection whereby the derived allele is less advantageous or perhaps even more harmful than the ancestral allele. However, as was noted previously, allele frequencies also increase and decrease by pure chance because of genetic drift. Despite this, it is useful to look at the DAF spectrum to get the first clues of the processes affecting the SNPs as a class.

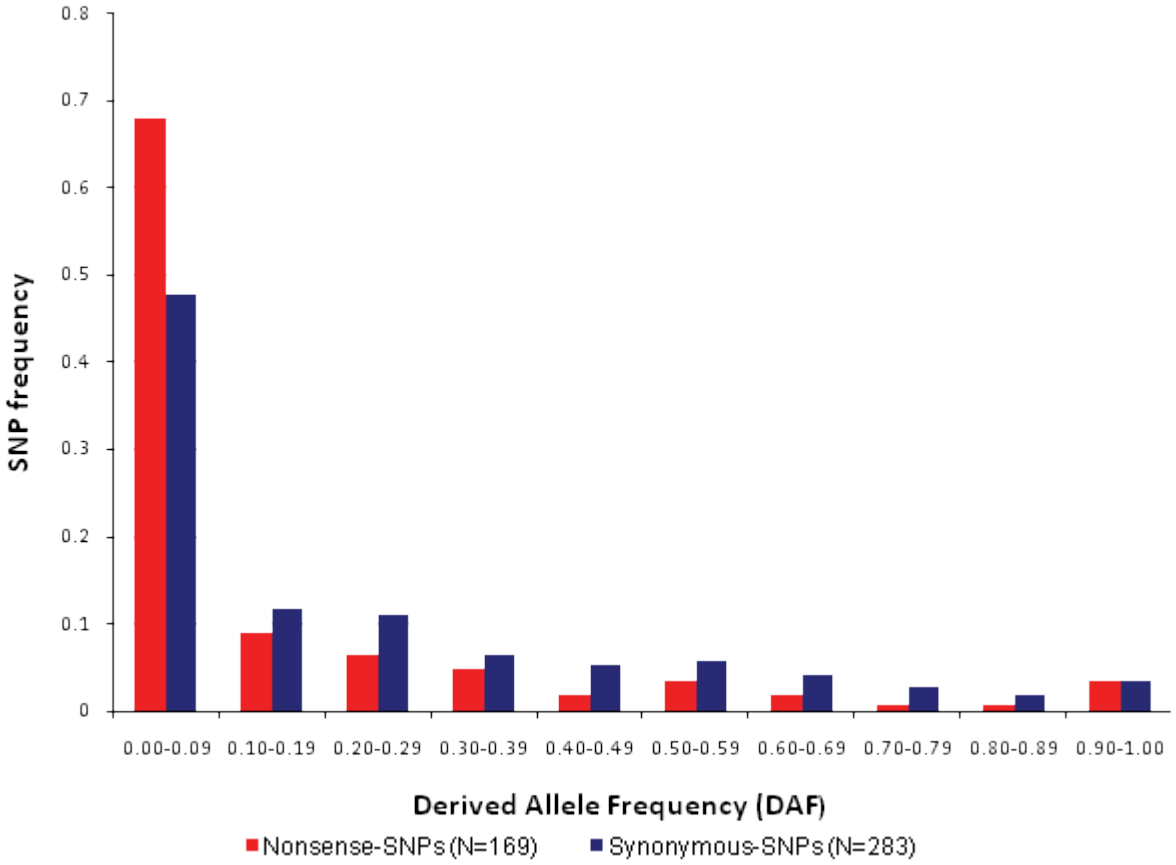


Figure 12 Comparison of the DAF spectrum of nonsense- and synonymous-SNPs in the sample of worldwide populations. The derived allele frequency was calculated for each SNP and sorted into ten bins. The DAF of nonsense-SNPs (red) was significantly lower than the DAF of the synonymous-SNPs (blue) (Kolmogorov-Smirnov, P<<0.001). The shape of the derived allele spectrum is consistent with previous reports (Williamson et al. 2005) and, while the tail of high frequency derived alleles has been explained by ancestral misspecification, it might also include potentially interesting positively selected genes.

It has previously been shown that there is a clear difference in the frequency spectrum between nonsynonymous and synonymous human SNPs where nonsynonymous-SNPs showed a relative excess of SNPs with rare derived alleles (Williamson et al. 2005). We thus compared the DAF of the nonsense-SNPs with those of synonymous-SNPs in the same samples. The derived stop allele of nonsense-SNPs was indeed found to be generally rarer than the derived allele of synonymous-SNPs (Kolmogorov-Smirnov, P<<0.001). This suggests that, as expected, negative selection is acting on stop alleles as a class to remove variants that are harmful over an evolutionary timescale.

Figure 12 shows the DAF spectrum in the combined populations for nonsense- and synonymous-SNPs divided into ten frequency bins. The DAF was also viewed by separating the populations into five categories (according to K=5 in Rosenberg et al. 2002) and the distributions were found to be similar to the distribution observed in Figure 12 (See Appendix E). The highest number of SNPs (68% nonsense and 48% synonymous) fall in the lowest DAF bin (0.00-0.09) which is not surprising as the derived allele is the new one to arrive in the population. Looking at the highest DAF bin (0.90-1.00) there are equal proportions of nonsense- and synonymous- SNPs (~3%). This excess of very high-frequency variants has been observed previously in the normalized site-frequency spectrum and has been explained by ancestral misspecification (Williamson et al. 2005). However, we double-checked our SNPs by comparing them to the chimpanzee sequence and found that ancestral allele was correctly inferred for all SNPs. We thus conclude that the high derived allele signal is real and note that this highest bin includes rs497116 in the *CASP12* gene, the derived allele of which is found to be positively selected and is discussed in chapter 4. Nonsense-SNPs with a DAF above 0.70 (within the three highest bins in Figure 12) are displayed in Table 6 below.

Gene	SNP	DAF
RBPJ	rs5007634	0.999
HERC6	rs4413373	0.994
*CASP12**	rs497116	0.962
THSD7B	rs12622896	0.960
*FMO2**	rs6661174	0.959
SLAIN1	rs17777179	0.945
*NPPA**	rs5065	0.848
KRTAP13-2	rs877346	0.718

Table 6 Genes with nonsense-SNP derived allele frequency above 0.7. *Genes discussed in more detail in text.

As expected, the previously mentioned *FMO2* gene shows up with a high derived allele frequency of nonsense-SNP rs6661174 (DAF = 0.959).

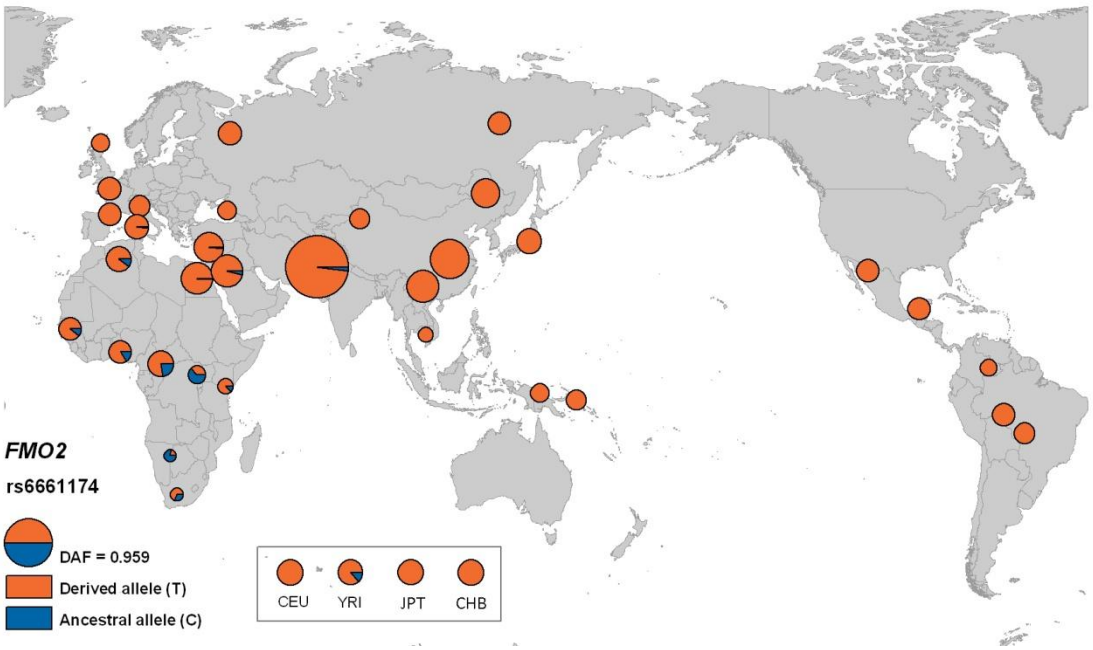


Figure 13 Geographical distribution of derived (stop) and ancestral (normal) alleles of rs6661174 in *FMO2*. The derived allele is nearly fixed in the human population but the ancestral allele has the highest frequency in Africa (especially the San) and is found at low frequencies in Sardinians, Druze, Bedouin, Palestinians and Pakistanis.

Previous studies have revealed that the derived truncated version of this gene is fixed in European and Asian populations while the ancestral functional allele has been found in African-Americans and Hispanics (Dolphin et al. 1998; Krueger et al. 2004; Veeramah et al. 2008). Indeed we found similar population distributions for this SNP (Figure 13), with the derived (stop) allele nearly fixed in human populations and the ancestral (normal) allele found at various frequencies in all African populations and at low frequencies in three populations from Israel (Bedouin, Druze and Palestinians), one population from Italy (Sardinians) and in Pakistan. If carriers of the functional allele are exposed to thioureas (which are present in a wide range of industrial, household and medical products) they are at increased risk of pulmonary toxicity (Veeramah et al. 2008). As exposure to these chemicals is now widespread it is interesting to consider whether they might also have been present in the pre-industrial environment and whether the stop allele might have reached its high frequency because of positive selection. A recent study used the LRH test to search for selection signals and did not find evidence for

unusually long haplotypes around either allele (nor did we for that matter, see section 3.1.6). However, as the authors note, the mutation was dated at ~500 thousand years ago (Veeramah et al. 2008) and was therefore likely to be too old for the LRH test to pick up a signal of positive selection.

In addition, the stop allele in *NPPA* (DAF = 0.848) has previously been reported at a high frequency in human populations and was shown to be associated with a decreased risk of stroke recurrence (Rubattu et al. 2004). Stroke is a disease of old age and may not itself have exerted a strong selective pressure in the past, but the association with a phenotype raises the possibility that the allele might be linked to other advantageous phenotypes as well and could thus be susceptible to positive selection.

The frequencies of ancestral and derived alleles in different populations are displayed in Figure 14A and Figure 14B for the nonsense- and synonymous SNPs, respectively. Some SNP studies have shown higher variability in European populations, which can be explained by the ascertainment bias caused by the SNPs being first identified in non-African populations (Tishkoff and Kidd 2004). Figure 14 illustrates the overall lower frequency of the nonsense-SNP derived allele and the lack of strong geographical bias in the distribution of either category. Furthermore, this addresses the ascertainment bias issue as the SNPs are not only ascertained (and variable) in Europeans.

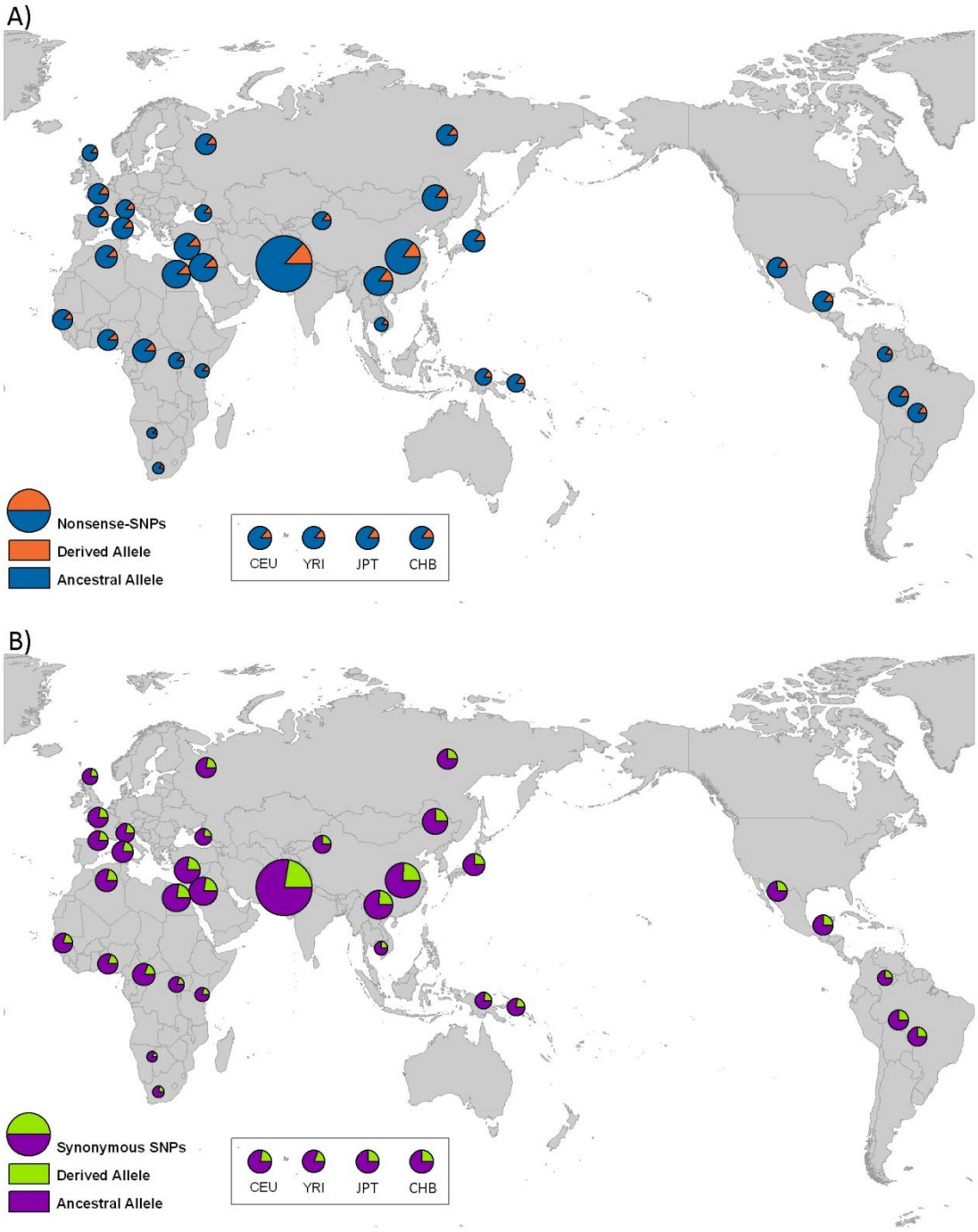


Figure 14 Mean frequencies of ancestral and derived alleles in different populations for A) Nonsense-SNPs and **B)** Synonymous-SNPs. Populations look similar, but derived alleles are more frequent in synonymous SNPs than nonsense SNPs.

3.1.1.2 Frequency of Homozygotes and Heterozygotes

Previous reports have suggested that nonsense-SNPs may be expected to be deleterious in the homozygous state but kept as a result of heterozygote neutrality or advantage (Dean et al. 2002). Therefore, we wished to know whether the high

derived allele frequencies were simply carried in the populations in the heterozygous state and whether or not stop homozygotes were present; and if so, whether they were found at the expected frequency. To this end, we examined the frequencies of the nonsense-SNPs as heterozygotes and homozygotes in our population samples. For 99 nonsense-SNPs (59%), at least one stop homozygous sample was found (Figure 15), showing that both copies of these genes may be truncated in our sampled individuals. These stop alleles are thus not only carried around in a heterozygous state and in fact we do not find exceptionally high frequencies of heterozygotes.

Three of the eight SNPs found in the HGMD database were not found as stop homozygotes. However, the other five were present as stop homozygotes and two (unsurprisingly in the *NPPA* and *FMO2* genes) were found at a high frequency.

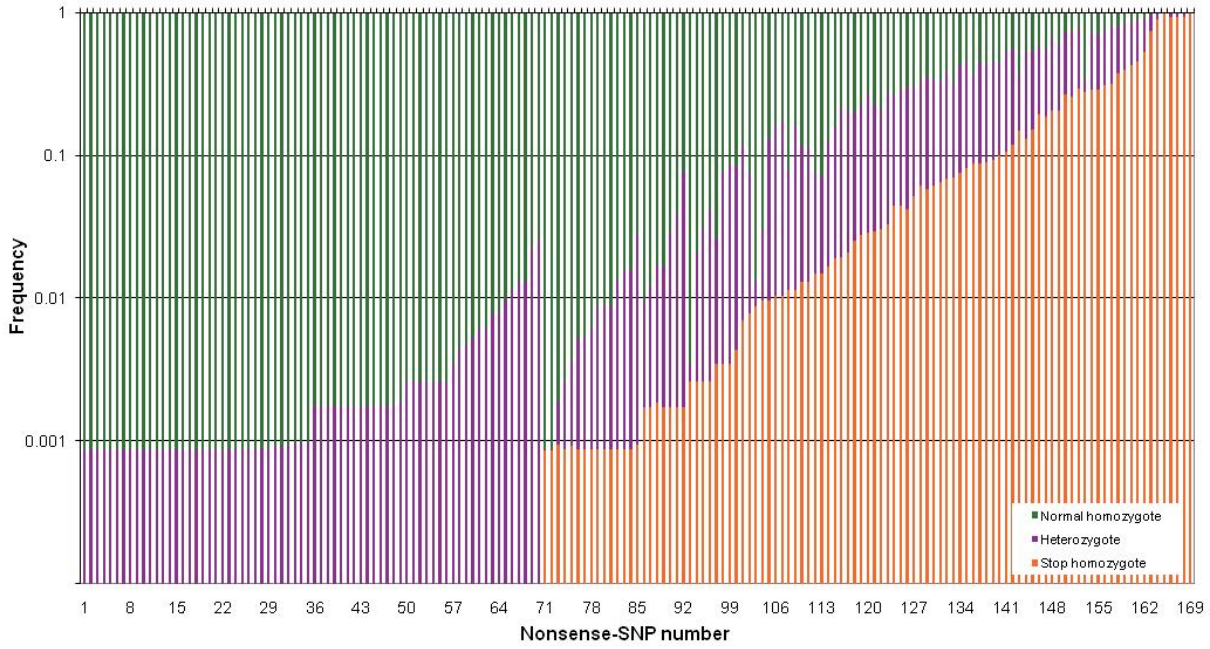


Figure 15 Proportions of stop homozygotes, normal homozygotes and heterozygotes for each nonsense-SNP. The genotype frequencies of normal homozygotes (green), heterozygotes (purple) and stop homozygotes (orange) were plotted on a logarithmic scale. The nonsense-SNPs were sorted along the X-axis according to the frequency of the stop homozygotes and the identifying number can be found on the accompanying CD.

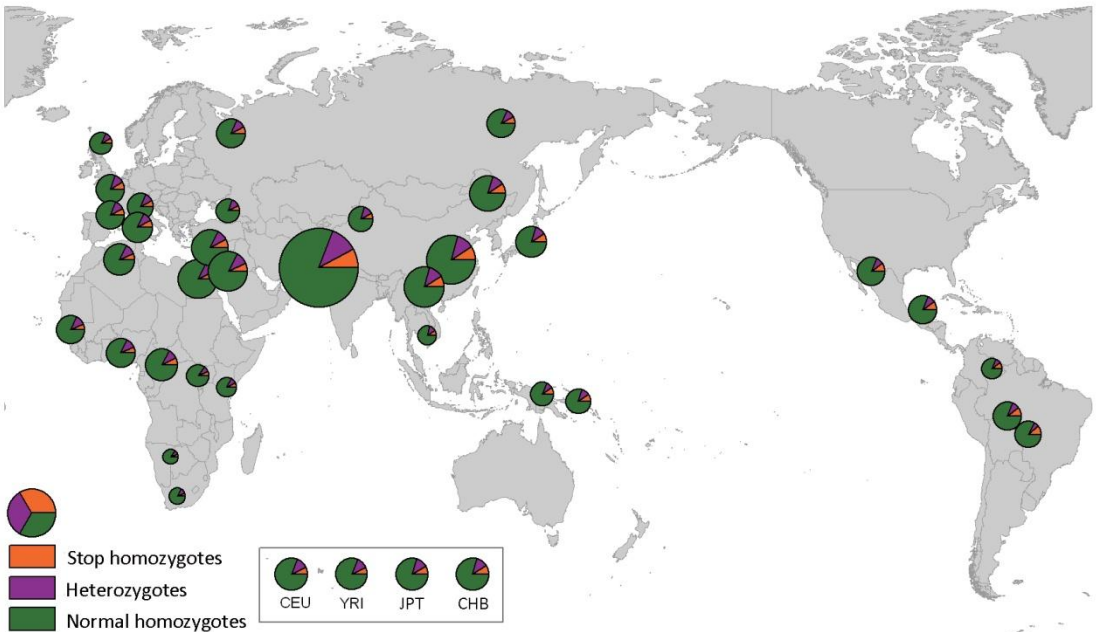


Figure 16 Proportions of stop homozygotes, heterozygotes and normal homozygotes in different populations.

Interestingly, the overall proportion of the different allelic states in the geographically distinct populations is very similar (Figure 16).

3.1.2 Stop that Nonsense! Protein Truncations and NMD

We wished to understand the effects these 169 stop alleles might be having on the gene product: the stop allele could result simply in a shorter protein, but truncated peptides are prevented and are usually eliminated by a NMD.

We mapped the position of the nonsense-SNP within the transcript and found that the truncations were distributed evenly throughout the polypeptide length (Figure 17). ~49% of the nonsense-SNPs led to the deletion of >50% of the amino acid sequence, an extensive truncation that might alter the protein structure and function radically.

In addition, 55% nonsense-SNPs were predicted to cause transcript degradation by NMD (in at least one transcript) which would result in loss of the gene function, while the rest of the nonsense-SNPs (45%) are expected to result in the production of a truncated protein. Either way, these nonsense-SNPs could be having severe effects on the gene product.

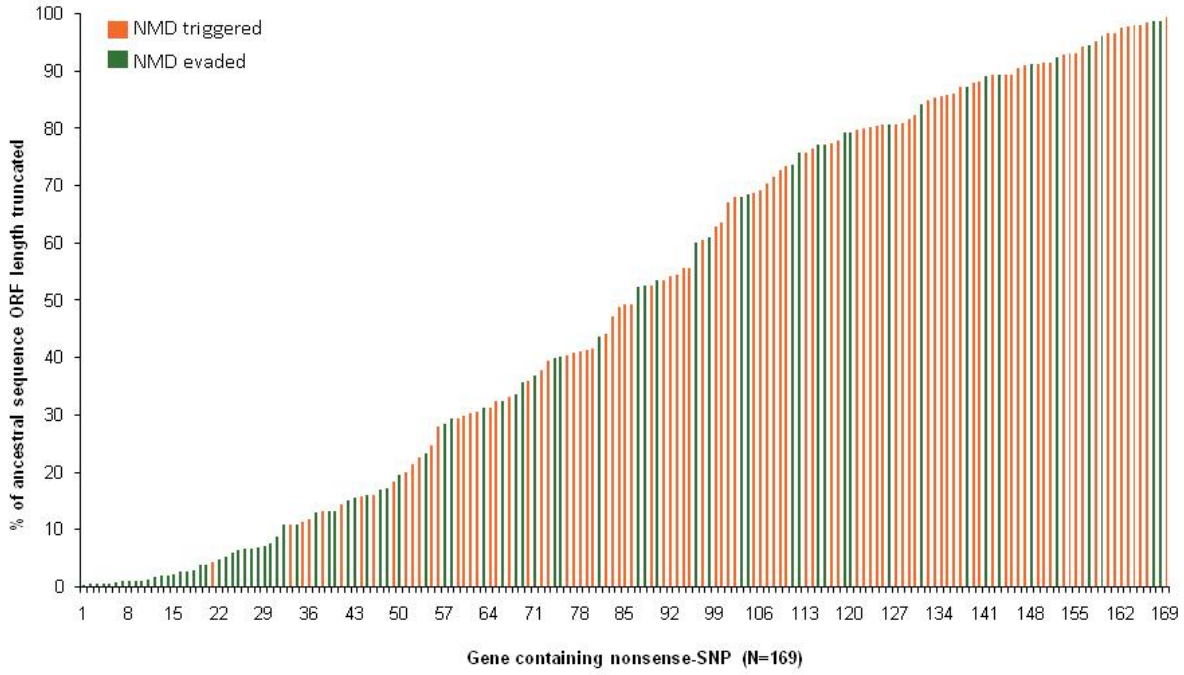


Figure 17 Even distribution of truncation positions in the protein. Truncations were calculated as the percentage of the ancestral ORF length. The 167 genes containing 169 nonsense-SNPs were sorted along the X-axis according to the amount of peptide truncation, starting at 1 for the lowest truncation and ending at 169 for the highest truncation. The identifying number of the SNP displayed in the figure can be found in the accompanying CD. Orange labels transcripts where NMD is predicted to be triggered which could result in the complete loss of the gene product, whereas green refers to transcripts where NMD is evaded because the nonsense-SNP is located either in the last exon or less than 50 nucleotides upstream of the last exon-exon boundary.

Table 7 displays the nonsense-SNPs that cause more than 90% of the peptide to be truncated and some of these will now be discussed. *DGCR8* is located in the DiGeorge syndrome chromosomal region and has the highest truncation in our dataset (99.4%); however, its detailed function is unknown. The nonsense-SNP in *AMPD1* causes a severe truncation (98.4%) resulting in the complete loss of the AMPD activity. AMPD activity determines the adenylate energy charge and energy metabolism in the cell and the *AMPD1* gene is found to be expressed at a high levels skeletal muscle. This SNP was typed in individuals with AMPD deficiency and was found to cause decreased enzyme activity (Morisaki et al. 1992) and was thus thought to be disadvantageous. Indeed, the stop allele is found at a low frequency of 4.4% in our samples and is mostly found in heterozygote state. However, a recent study has shown that individuals carrying the stop allele develop a greater and

faster blood flow response to high intensity exercise than normal homozygotes (Norman et al. 2008). Thus, while perhaps only advantageous to a selected few, this is another example (in addition to *ACTN3* mentioned in our introduction*)* of a gene loss that could contribute to athletic performance.

SNP ID	Gene ID	Truncated (%)	Retained (%)	NMD candidate
rs2106143	DGCR8*	99.4	0.6	YES
rs13343184	OR5AK2*	98.7	1.3	NO
rs3733689	PCDHB10	98.6	1.4	NO
rs17602729	AMPD1*	98.4	1.6	YES
rs17582155	RORC	98.1	1.9	YES
rs2271286	SPTBN5	98.0	2.0	YES
rs2043211	CARD8*	97.7	2.3	YES
rs9332960	SRD5A2	97.6	2.4	YES
rs11552294	SLC25A5	96.7	3.3	YES
rs2400941	Q96NA9_HUMAN	96.6	3.4	YES
rs16930998	OR51I1*	96.2	3.8	NO
rs16982743	SIGLEC12	95.1	4.9	YES
rs1459101	OR4C16*	94.5	5.5	NO
rs11539065	Q9H579-2	94.2	5.8	YES
rs11546516	ENOPH1	93.1	6.9	YES
rs12048007	ARHGEF19	93.0	7.0	YES
rs9567515	KIAA1704	93.0	7.0	YES
rs11542462	NP_660151.2	92.4	7.6	NO
rs17292725	STARD6	91.4	8.6	YES
rs7532205	C1orf105	91.4	8.6	YES
rs2233091	MATN4	91.2	8.8	YES
rs7120775	OR4X2*	91.1	8.9	NO
rs6671527	MOBKL2C	91.1	8.9	YES
rs2292830	CLCA3	90.4	9.6	YES

Table 7 Nonsense-SNPs that cause the most extreme truncations, leading to a degraded transcript if NMD is triggered or a severely truncated peptide when NMD is not triggered. *Genes discussed in more detail in text.

Table 7 contains four genes involved in odor perception (*OR5AK2*, *OR51I1*, *OR4C16*, *OR4X2*) which is not very surprising as the olfactory receptor gene family is the largest in the mammalian genome with more than 60% reported as disrupted and functionally inactivated in humans, a high proportion compared to non-human primates (Gilad et al. 2003b; Glusman et al. 2001). It has been suggested that, as humans lost the need to depend on their sense of smell for survival, loss of olfactory receptor genes became selectively neutral and so losses accumulated (Gilad et al. 2003a). We would therefore expect to find several of them in this survey of nonsense-SNP in the human genome.

The *CARD8* gene is implicated in the regulation of apoptosis and inflammation. A number of studies have reported conflicting results on the possible association of the nonsense-SNP (rs2043211) with inflammatory bowel disease, with some associating the stop allele with Crohn's disease and ulcerative colitis (Fisher et al. 2007), others finding an association of the normal allele with Crohn's disease (McGovern et al. 2006) and yet others reporting no association with either allele (Franke et al. 2007). The rs2043211 nonsense-SNP has been observed in alternatively spliced transcripts. A recent study found that expression of the different transcripts in Crohn's disease patients homozygous for the stop allele was also observed in patients homozygous for the normal allele. A possible explanation given was that there was a partial rescue of the gene function by alternative initiation of translation or splicing which then evaded the effects of NMD and led to a functionally compromised but near full-length protein (Bagnall et al. 2008). The stop allele was found at 34% frequency in our samples and results (in one transcript) in a severe truncation (97.7%) of the *CARD8* protein product with the expectation of NMD being triggered.

3.1.3 Gene Expression

More than half of the nonsense-SNPs were predicted to trigger NMD which would result in the gene product not being expressed as the transcript would be degraded. In order to test this assumption we used available gene expression data to see if there was a correlation between the genotypes (stop/stop, normal/normal and stop/normal) and expression levels.

Gene expression information was obtained for 14,456 probes showing expression variation in lymphoblastoid cell lines (Stranger et al. 2007b). We first attempted to test our set of 169 nonsense-SNPs against this data, but found that only 57 of them mapped in genes corresponding to the 14,456 probes and only 19 were polymorphic and typed in the HapMap (The International HapMap Consortium 2005). This gave us little power to draw any conclusions and we thus resorted to using all available

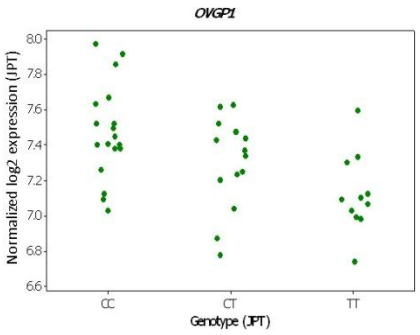
nonsense-SNPs (dbSNP126). This gave us a starting dataset of 1,624 SNPs instead of our 169. In the end 588 of these were typed in HapMap and 105 of those could be mapped in the genes corresponding to the expression probes. 11 in total were found to be significant (FDR, P<0.01), four of which were in our dataset. A summary of these results is reported in Table 8.

We expected to find a correlation between the genotype and expression level of the nonsense SNP based on the assumption that when the stop allele triggers NMD, it will lead to lower or zero expression of the gene product. Indeed, we find that in most cases (six out of eight) where the SNP was expected to trigger NMD, the stop allele was in turn associated with lower expression. Three of the SNPs that resulted in significant correlations were not expected to trigger NMD, and thus the expression of either allele (stop or normal) could be higher. However, we see two cases where the SNP should trigger NMD but the stop allele shows a correlation with higher expression levels. There are several possible explanations for this. Firstly, the transcripts in which the nonsense-SNP was identified in dbSNP could be alternatively spliced and thus different from those measured in the gene expression analysis. Secondly, prediction of NMD is imperfect and might be misleading in these cases. Thirdly, they could be false discoveries, this is but less likely because only 1 in 100 is expected.

SNP ID	Gene ID	Population	% Retained	% Truncated	NMD triggered	Stop allele expression[α]
rs10009430*	LOC132321	YRI	72	28	Y	lower
rs1227794	hmm7037	YRI	47	53	Y	lower
rs1264887	OVGP1	JPT	18	82	Y	lower
rs16982007*	FLJ10922	CHB	51	49	Y	lower
rs2273865	LGALS8	JPT	59	41	Y	lower
rs4148974	NDUFV3	CEU-CHB-JPT	42	58	Y	lower
rs1059307	LOC389414	CHB - YRI	33	67	Y	higher
rs7188975	LCMT1	CEU-CHB-JPT-YRI	16	84	Y	higher
rs12435565*	C14orf129	JPT	57	43	N	higher
rs1801876*	KIAA0748	CEU-CHB-JPT-YRI	97	3	N	higher
rs745961	FLJ14640	CEU	99	1	N	higher

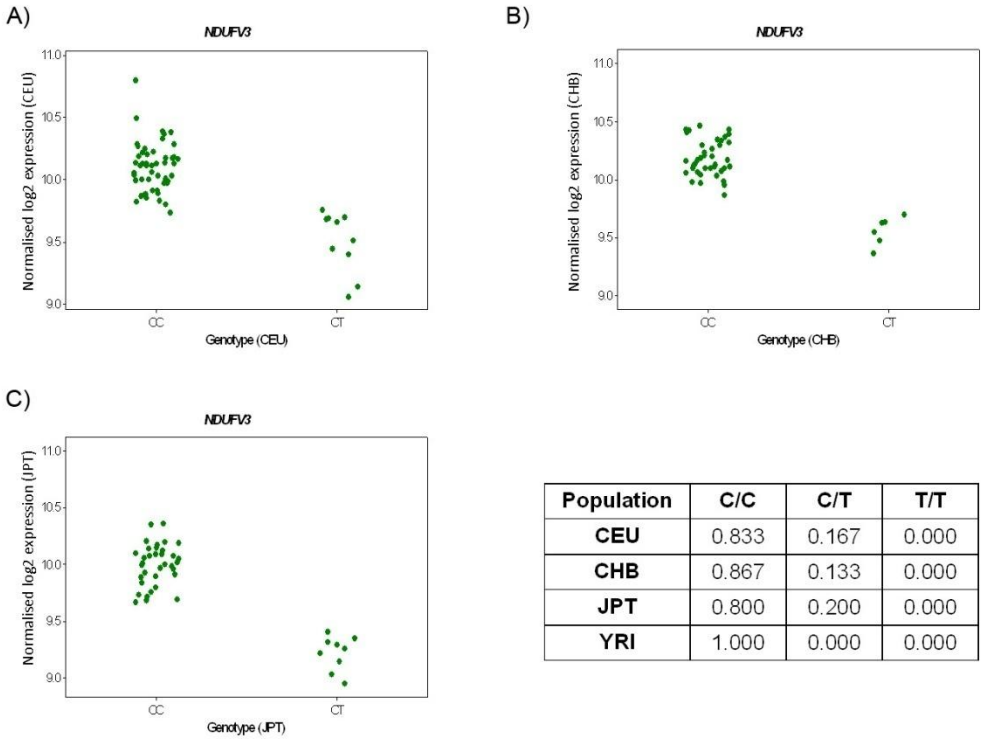
Table 8 Summary of nonsense-SNPs observed with significant (P<0.01) correlations between genotype and expression level. [α]Expression of stop allele relative to normal allele. *Nonsense-SNP present in our dataset.

Figure 18 gives an example of the expression levels for the different genotypes of rs1264887 (in *OVGP1*), a SNP that is expected to trigger NMD and exhibited an association (-log10(p-value) = 3.187, permutation threshold P<0.01) between the genotype and expression levels in the JPT population. This correlation, like several others (Table 8) was only found to be significant in the JPT. One possible explanation for a correlation only being significant in one population could be the low frequency of the stop allele. For rs1264887 the stop homozygotes (T/T) indeed are at the highest frequency in the JPT (where the significant correlation was found), being at a much lower frequency in the other populations.



Population	C/C	C/T	T/T
CEU	0.879	0.121	0.000
CHB	0.200	0.667	0.133
JPT	0.409	0.341	0.250
YRI	0.424	0.492	0.085

Figure 18 Plot of expression levels vs. possible genotypes for rs1264887 (in *OVGP1*) reported with a significant correlation (significant at permutation threshold P<0.01) in the JPT population. The stop allele is T and has lower expression levels than the normal allele (C). The genotype frequencies in the all HapMap populations are shown in the table beside the plot.



Population	C/C	C/T	T/T
CEU	0.833	0.167	0.000
CHB	0.867	0.133	0.000
JPT	0.800	0.200	0.000
YRI	1.000	0.000	0.000

Figure 19 Plot of expression levels vs. genotypes for rs4148974 (in *NDUFV3*) reported with a significant correlation P<0.01) in the three populations, A) CEU, B) CHB and C) JPY. The stop allele T has a lower expression in heterozygote state than the normal homozygote (C/C). Inset are the genotype frequencies in all HapMap populations.

Another example of significant correlation can be seen in Figure 19 where the genotypes of rs4148974 (*NDUFV3*) were significant in three populations, CEU, CHB and JPT, but only in relation to the normal homozygote versus the heterozygote as no stop homozygotes are observed. As can be seen from the frequency table (inset in Figure 19), YRI only have the C/C genotype and could thus not yield significant association.

Only 11 out of the 105 nonsense-SNP/gene associations tested were found with a significant difference in expression levels with respect to the genotype. However, we explored only those 14,456 probes out of a total of 47,296 that, after normalisation, had previously showed expression level variation (Stranger et al. 2007b). Our nonsense-SNPs may have a great effect on the expression levels, and indeed 55% of them are expected to trigger NMD and cause mRNA degradation. Therefore, if the intitial probes tested included genes with a low expression of the normal allele and

no expression with the stop allele, then it is very likely that this probe would not have made it into the 14,456 probe set used here as the expression difference between low and none is not big enough to satisfy our "differentially expressed" criterion.

Furthermore, there is a general expectation for stop alleles to be rarer than normal alleles, as their disruptive nature might be harmful and therefore eliminated from the population by negative selection. Indeed, we found this to be true for the majority of our 169 nonsense SNPs (see section 3.1.1.1). Because of this it might be difficult to pick up a signal of association between gene expression levels and genotypes of nonsense-SNPs and this may be the major reason why we observed only 11 significant signals. From such a small amount of data it is difficult to generalise about the effects nonsense-SNPs triggering NMD are having on the gene product. Despite this we do find it reassuring that six out of eight significant SNPs were behaving as expected.

3.1.4 Gene Ontology Enrichment Analysis

To further understand the functional and physiological consequences of our nonsense-SNPs as a class, we used Gene Ontology (GO) information to determine whether there was an enrichment of molecular function or biological process terms in these "lost" genes. The GO terms found to be strongly enriched ($P<0.05$) in our set of nonsense-SNPs are shown in Table 9. Amongst these we found olfactory receptor (OR) genes to be overrepresented, which was not surprising as it has previously been reported that humans have a reduced sense of smell (Gilad et al. 2003a; Gilad et al. 2003b). We also detected a significant overrepresentation of genes involved in the nervous system which was unexpected as genes in the nervous system have generally been shown to be very conserved (Nielsen et al. 2005).

Term	Count	%	P-value	Bonferroni
Biological Process				
GO:0007608~sensory perception of smell	12	8.11%	0.0002	0.6800
GO:0007606~sensory perception of chemical stimulus	12	8.11%	0.0005	0.9100
GO:0032501~multicellular organismal process	40	27.03%	0.0014	1.0000
GO:0003008~system process	20	13.51%	0.0032	1.0000
GO:0050877~neurological system process	16	10.81%	0.0114	1.0000
GO:0019320~hexose catabolic process	4	2.70%	0.0178	1.0000
GO:0046365~monosaccharide catabolic process	4	2.70%	0.0184	1.0000
GO:0046164~alcohol catabolic process	4	2.70%	0.0196	1.0000
GO:0007166~cell surface receptor linked signal transduction	21	14.19%	0.0202	1.0000
GO:0007186~G-protein coupled receptor protein signalling pathway	15	10.14%	0.0203	1.0000
GO:0007154~cell communication	38	25.68%	0.0228	1.0000
GO:0009056~catabolic process	11	7.43%	0.0263	1.0000
GO:0007600~sensory perception	12	8.11%	0.0265	1.0000
GO:0022610~biological adhesion	11	7.43%	0.0327	1.0000
GO:0007155~cell adhesion	11	7.43%	0.0327	1.0000
GO:0044275~cellular carbohydrate catabolic process	4	2.70%	0.0368	1.0000
GO:0006118~electron transport	8	5.41%	0.0406	1.0000
GO:0016052~carbohydrate catabolic process	4	2.70%	0.0430	1.0000
GO:0016337~cell-cell adhesion	6	4.05%	0.0434	1.0000
GO:0007165~signal transduction	34	22.97%	0.0436	1.0000
GO:0050878~regulation of body fluid levels	4	2.70%	0.0486	1.0000
Molecular Function				
GO:0004984~olfactory receptor activity	12	8.11%	0.0003	0.5700
GO:0030246~carbohydrate binding	10	6.76%	0.0005	0.7600
GO:0004872~receptor activity	28	18.92%	0.0011	0.9600
GO:0004499~flavin-containing monooxygenase activity	3	2.03%	0.0013	0.9800
GO:0004888~transmembrane receptor activity	20	13.51%	0.0033	1.0000
GO:0005529~sugar binding	7	4.73%	0.0037	1.0000
GO:0060089~molecular transducer activity	30	20.27%	0.0045	1.0000
GO:0004871~signal transducer activity	30	20.27%	0.0045	1.0000
GO:0001584~rhodopsin-like receptor activity	13	8.78%	0.0126	1.0000
GO:0050661~NADP binding	3	2.03%	0.0150	1.0000
GO:0004930~G-protein coupled receptor activity	14	9.46%	0.0151	1.0000
GO:0016709~oxidoreductase activity	3	2.03%	0.0196	1.0000
GO:0046983~protein dimerization activity	7	4.73%	0.0271	1.0000
GO:0042803~protein homodimerization activity	5	3.38%	0.0287	1.0000

Table 9 GO terms strongly enriched (P<0.05) in the set of nonsense-SNP genes. The table displays the enriched terms associated with the list of nonsense-SNP genes, the number of genes involved in the term, the percentage (involved genes/total genes). The P-value (modified Fisher-Exact, EASE score) and the Bonferroni correction is given (see discussion on multiple correction issue of this type of analysis in section 2.3.7 in Materials and Methods).

Considering the disruptive effects of nonsense-SNPs, it is possible that the overrepresentation of certain GO categories is simply reflecting a higher number of paralogs for genes containing nonsense SNPs. If this were true, it might result from the paralogs serving as a "backup system" for the disrupted genes, reducing the

negative selection pressure on them. We noted that 51% of the nonsense-SNP genes have at least one paralog whereas in comparison only 35% of all human genes in Ensembl (release 50) are reported with a paralog. This difference was found to be moderately significant (Fisher exact, $P<0.05$), and was even more significant in an earlier version (Ensembl release 46, Fisher exact, $P<0.001$), so it is possible that their function is "backed up" by duplicated paralogs in the human genome.

To further investigate the effects paralogs might be having on the GO results, we wished to repeat the GO enrichment analysis using the set of genes with zero paralogs versus the set of genes with one or more paralog. Unfortunately, the set of genes with zero paralogs had less than 80% of the genes annotated and therefore it was not possible to perform any enrichment analysis on this group in DAVID. However, the number of genes with one or more paralog was sufficient and the resulting GO terms that came up significant were similar to those that came out of the whole dataset.

To give us some idea of the number of paralogs the genes in these enriched GO categories have, we counted the paralogs found for each gene belonging to two enriched GO categories, namely olfactory receptor activity (GO:0004984) and neurological system process (GO:0050877) (see Table 10). For the former category a total of six genes were found to have zero or one paralogs, four had more than one paralog and one gene, *OR7G3* (ENSG00000170920), was found with a total of 20 paralogs. In the case of nervous system genes, eight genes were found with zero or only one paralog and six had more than one, albeit two genes showed extreme numbers, *GRIK5* (ENSG00000105737) with 17 and the same gene as before *OR7G3* again with 20.

Number of paralogs	GO:0004984 ~olfactory receptor activity	GO:0050877 ~neurological system process
0	4	5
1	3	3
>1	4	6
Total number of genes	**11**	**14**

Table 10 Number of paralogs for genes in two enriched GO categories.

3.1.5 Population Differentiation

As geographically separated populations may be subject to distinctive selective environments, selection can increase population differentiation at a selected locus. We used F_{ST} (Weir and Cockerham 1984) as a measure of population differentiation and found that when samples were grouped into 37 populations, most SNPs (both nonsense and synonymous) had low F_{ST} values within the 0.00-0.19 bin (Figure 20) as might be expected for human SNPs. Previous estimates of the empirical distribution of F_{ST} values across human populations have revealed average values of 0.11 (Barreiro et al. 2008), 0.12 (Akey et al. 2002; The International HapMap Consortium 2005) and 0.13 (Weir et al. 2005) and thus we consider those nonsense-SNPs with F_{ST} values in the extreme tails of the distribution as potentially interesting outliers that may have been affected by natural selection.

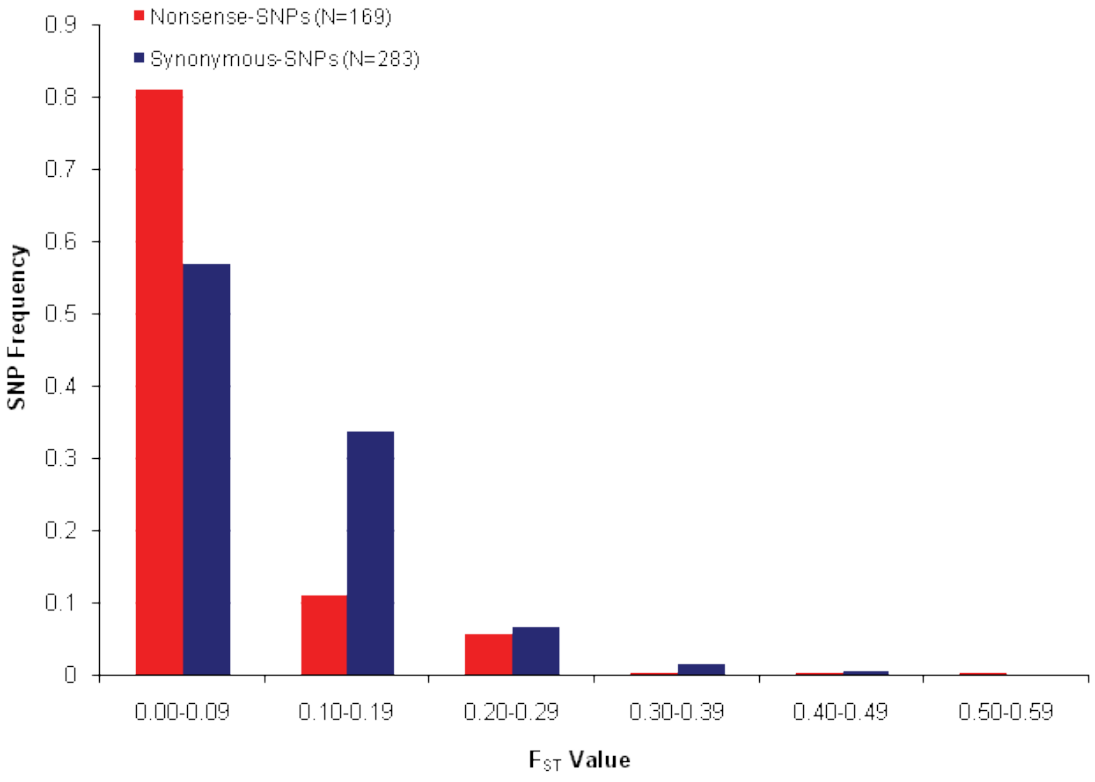


Figure 20 Comparison of F_{ST} values between nonsense- and synonymous-SNPs for 37 populations. The F_{ST} values were sorted into six bins and most of the SNPs (both nonsense and synonymous) fell in the lowest bin (0.00-0.19). On average, nonsense-SNPs (red) had significantly lower F_{ST} values than synonymous-SNPs (blue) (Kolmogorov-Smirnov, $P<<0.001$) with a mean of ~0.06 and ~0.10, respectively. The highest outlier (F_{ST}=0.54) was found in a nonsense-SNP (rs1343879) within the *MAGEE2* gene.

Variability in F_{ST} values between SNPs was to be expected as it has previously been shown that F_{ST} is different between regions of the genome (Weir et al. 2005). On average, nonsense-SNPs had significantly lower F_{ST} values than synonymous-SNPs (Kolmogorov-Smirnov, $P<<0.001$). This is in accordance with a recent study (Barreiro et al. 2008) that showed an excess of low F_{ST} values for non-synonymous SNPs compared to other classes such as synonymous SNPs. Furthermore, by taking the effect of ascertainment bias out of the equation and matching the F_{ST} values to the minor allele frequency, the authors came to the conclusion that the low values observed were a signal of purifying rather than balancing selection as it represented an excess of rare but not intermediate variants (Barreiro et al. 2008). To test for this in our data, we plotted the F_{ST} values of nonsense-SNPs against their MAF (Figure 21) and found no significant correlation between the two. However, we also find that the majority of low F_{ST} values are in SNPs with low MAF. We therefore suspect that the excess of low F_{ST} values observed for the nonsense-SNPs here is also the consequence of purifying selection acting against deleterious mutations.

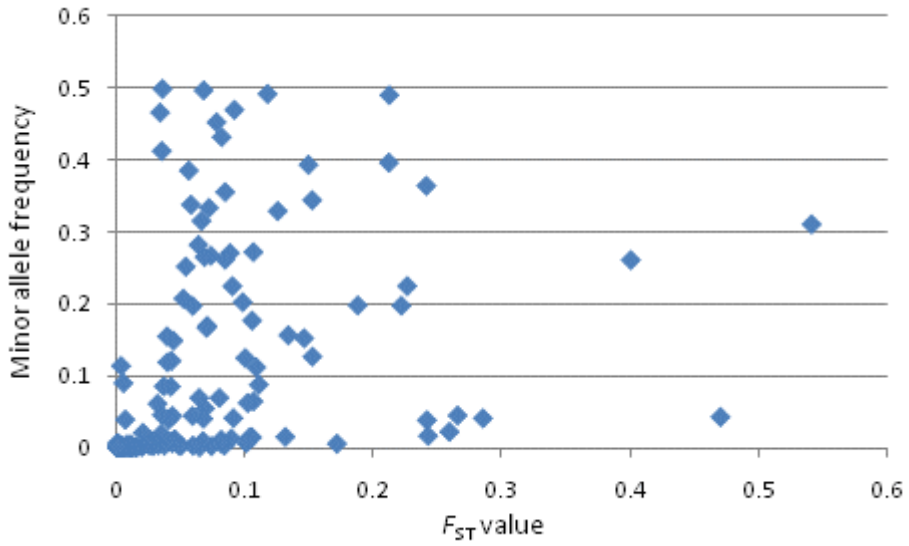


Figure 21 F_{ST} plotted against the MAF in the nonsense-SNPs. No significant linear correlation was found.

The highest F_{ST} value (0.54) was found in a nonsense-SNP (rs1343879) within the *MAGEE2* gene, a melanoma-associated antigen, with the alleles at intermediate

frequencies. This gene is on the X chromosome which has a smaller effective population size compared to the autosomes and might therefore be more sensitive to demographic events. In fact the X chromosome has previously been shown to have higher levels of differentiation compared to the autosomes (Akey et al. 2002; The International HapMap Consortium 2005). However, previous measures of empirical distributions of F_{ST} values on the X chromosome have revealed average values of 0.21 (The International HapMap Consortium 2005) and 0.195 (Akey et al. 2002), and therefore the SNP in the *MAGEE2* gene (F_{ST}=0.54) should still be considered an extreme outlier. In addition to the nonsense-SNP in *MAGEE2*, several more are reported with an F_{ST} value above 0.2 (Table 11). While the listed *MAGEE2* and *CASP12* will be explored in more detail in chapter 4 it is worth noting that *FMO2* turns up again.

SNP ID	Gene ID	F_{ST}^{α}	Heterozygosity$^{\beta}$	Above 99th percentile (32 pops)	Above 99th percentile (5 pops)
rs1343879	*MAGEE2**	0.54	0.429	+	+
rs12471298	*SEMA4C**	0.47	0.082	+	+
rs2293766	*ZAN**	0.40	0.386	+	+
rs6661174	*FMO2**	0.28	0.079		
rs7532205	*C1orf105*	0.26	0.086		
rs11089781	*APOL3*	0.26	0.043		
rs3211938	*CD36*	0.24	0.032		
rs497116	*CASP12*	0.24	0.074		
rs1801876	*KIAA0748*	0.24	0.463	+	+
rs4723884	*Q8N8G3_HUMAN*	0.23	0.349		
rs16982743	*SIGLEC12*	0.22	0.317		
rs1052972	*REG4*	0.21	0.500	+	+
rs1476860	*OR1B1*	0.21	0.479	+	+

Table 11 Nonsense-SNPs with highest F_{ST} values (above 0.2) for 37 population division. αCalculated according to reference (Weir and Cockerham 1984) across the 37 populations used in this study. ᵝCalculated according to (Nei 1987). The last two columns have a + if the nonsense-SNP was observed as significant in empirical comparison (in Figures 23 and 24). *Gene labelled in Figure 25.

In order to assess the significance of these high F_{ST} values we compared them to the empirical distribution of F_{ST} values we had calculated for the HapMap data and HGDP-CEPH data (divided into both 32 and 5 populations, see description in section 2.3.8.1). The comparison with the HapMap F_{ST} data revealed no more outliers

above the 95th or 99th percentiles than would be expected by chance. Although the *MAGEE2* is still the highest F_{ST} value observed and the only nonsense-SNP value above the 99th percentile (Figure 22).

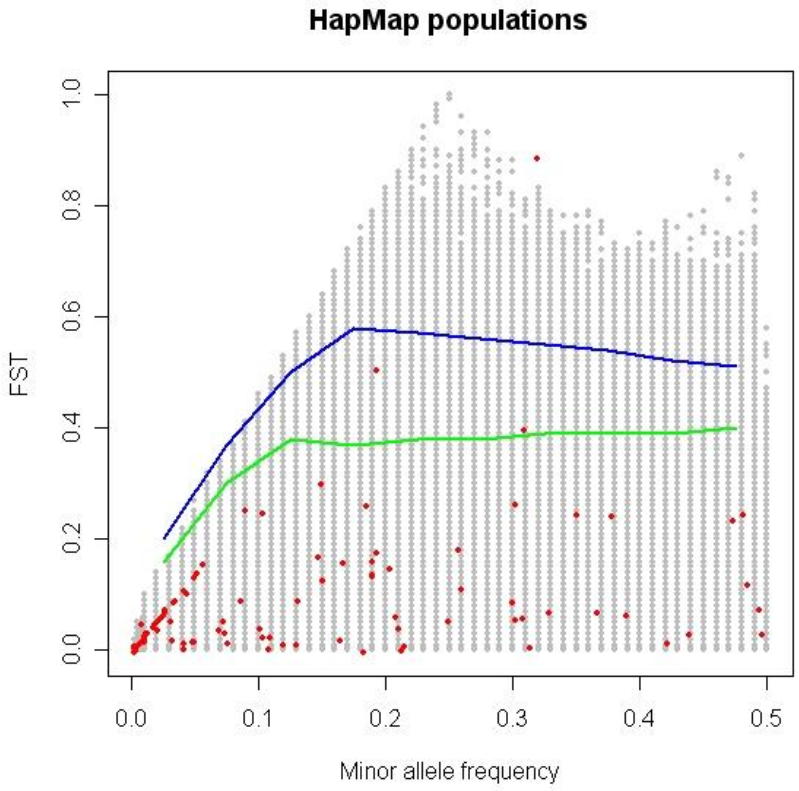


Figure 22 Empirical distribution for HapMap F_{ST} values. Values were calculated for SNPs from most of the HapMap data (in grey), and for the nonsense-SNPs polymorphic and present in HapMap (N=104). Green and blue lines represent the 95th and 99th percentiles, respectively Only one nonsense-SNP is above the 99th percentile so no more than would be expected by chance.

The empirical distribution for the HGDP-CEPH data showed more significance for the high F_{ST} nonsense-SNPs as six of them were reported above the 99th percentile for both the 32 population division (similar to that used in this thesis) as well as the population division into 5 major geographic regions (K=5 in Rosenberg et al. 2002) (see Table 11 and Figure 23 and 24). In addition to the SNPs reported in Table 11, one additional nonsense-SNP was calculated with a high F_{ST} value above the 99th percentile in the 32 HGDP-CEPH populations, rs12520799 in *C5orf20*.

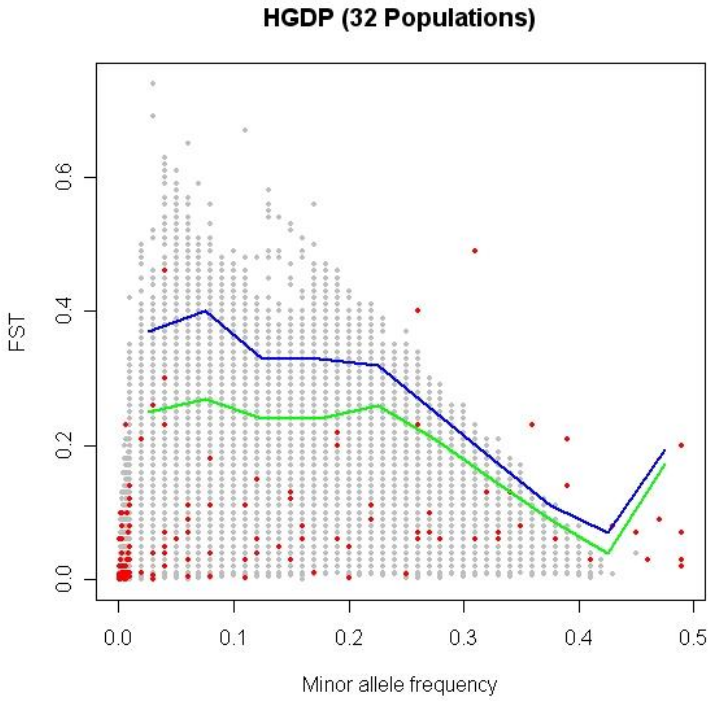


Figure 23 Empirical distribution for HGDP-CEPH F_{ST} values calculated for the 32 HGDP-CEPH populations. Green and blue lines represent the 95th and 99th percentiles, respectively. Nonsense-SNPs are plotted in red and seven (six of which are in Table 11) are observed as outliers above the 99th percentile.

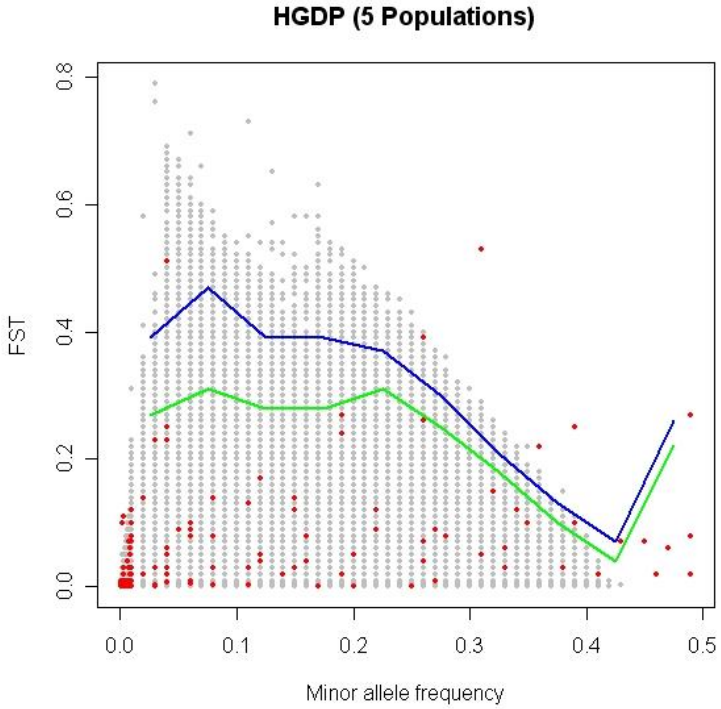


Figure 24 Empirical distribution for HGDP-CEPH F_{ST} values calculated for the HGDP-CEPH populations divided into five major geographic regions. Green and blue lines represent the 95th and 99th percentiles, respectively. Nonsense-SNPs are plotted in red and six (see Table 11) are observed as outliers above the 99th percentile.

While F_{ST} can pick up positive selection as measured by population differentiation, heterozygosity can assess the genetic diversity within a population. A selective sweep can reduce genetic diversity while balancing selection can increase the diversity. Therefore, loci with a high F_{ST} and low heterozygosity might indicate a selective sweep in a single population, while regions with high F_{ST} as well as high heterozygosity could be the result of population-specific balancing selection (Walsh et al. 2006). As always, random drift could also lead to the same results. We plotted F_{ST} versus heterozygosity and found no correlation between the variables (Figure 25). However, we note that the several of the nonsense-SNP displayed in Table 11 also show outlier behaviour in terms of heterozygosity. The SNPs in *SEMA4C* and *FMO2* have high F_{ST} but a low heterozygosity, which could indicate a selective sweep. The SNPs in *MAGEE2* and *ZAN*, on the other hand, have high F_{ST}s as well as high levels of heterozygosity which could be a sign of balancing selection.

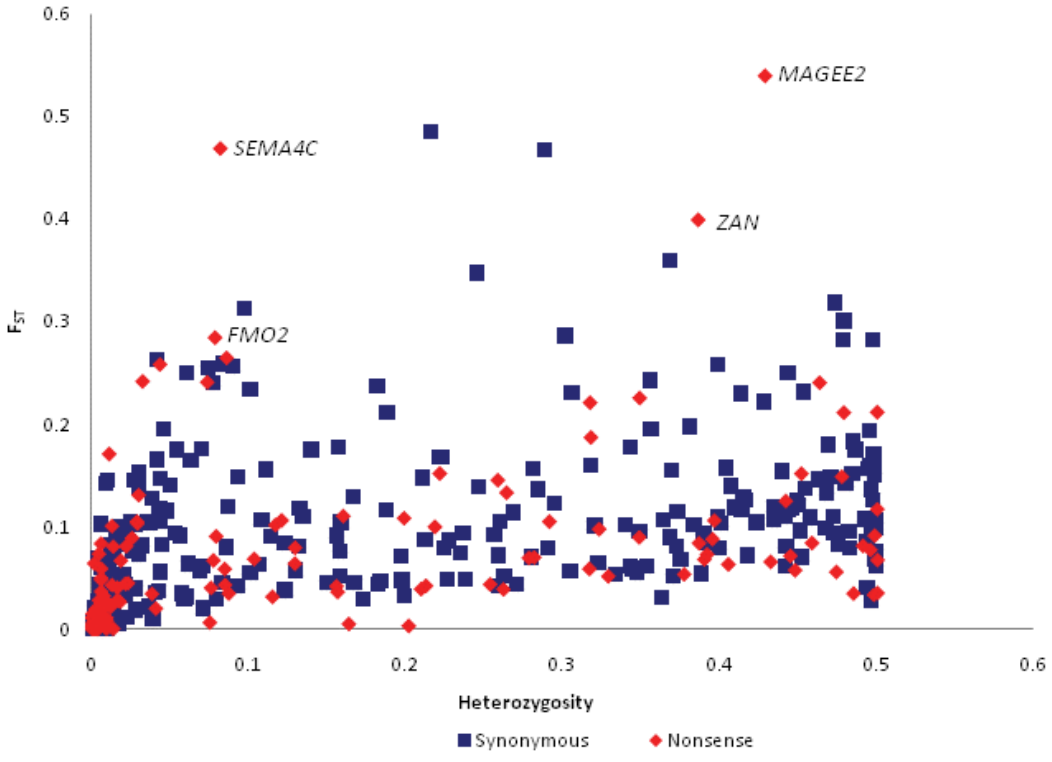


Figure 25 F_{ST} versus heterozygosity. No linear correlation was observed.

A previous study found that the mean heterozygosity at nonsense-SNP sites (2.7%) was significantly lower than that at synonymous-SNP sites (28.3%) in the

same gene which could imply that negative selection has affected allele frequencies at SNP loci in humans as heterozygosity is lowest at sites that are expected to have the greatest impact on protein structure (Hughes et al. 2003).

To better visualize the population differentiation observed we plotted the geographical distribution of the stop and normal alleles. The geographical distribution of *FMO2* has already been displayed in section 3.1.1.1 and that of *MAGEE2* will be shown in chapter 4. The geographical distribution of the *SEMA4C* nonsense-SNP is illustrated in Figure 26, and while the stop allele is at a low frequency worldwide (4%), it is only found in the Americas where it is present at 50% or greater in several samples. This high degree of continental specificity is very striking and merits further investigation of the possibility of regional selection. The stop allele in *ZAN* is found in various populations but has a strong east-west geographical structure (Figure 27) that again is worth investigating further. In contrast, Figure 28 gives an example of a nonsense-SNP with a random allelic distribution, as may have been evident from its low F_{ST} value (0.03).

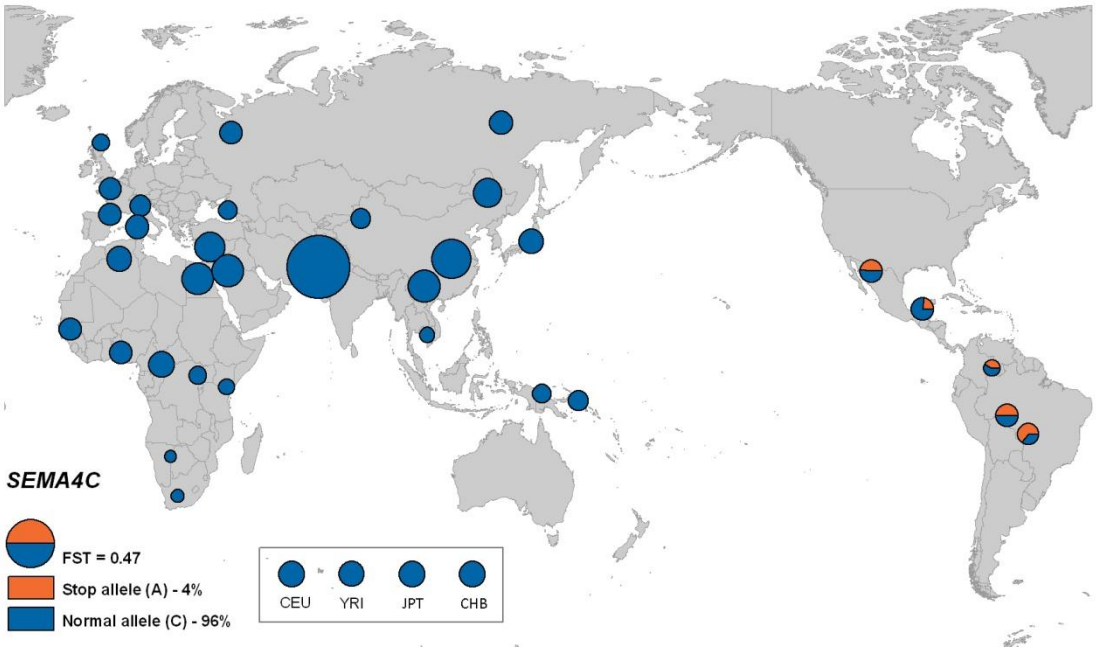


Figure 26 Geographical distribution of alleles of nonsense SNP rs12471298 in *SEMA4C*.

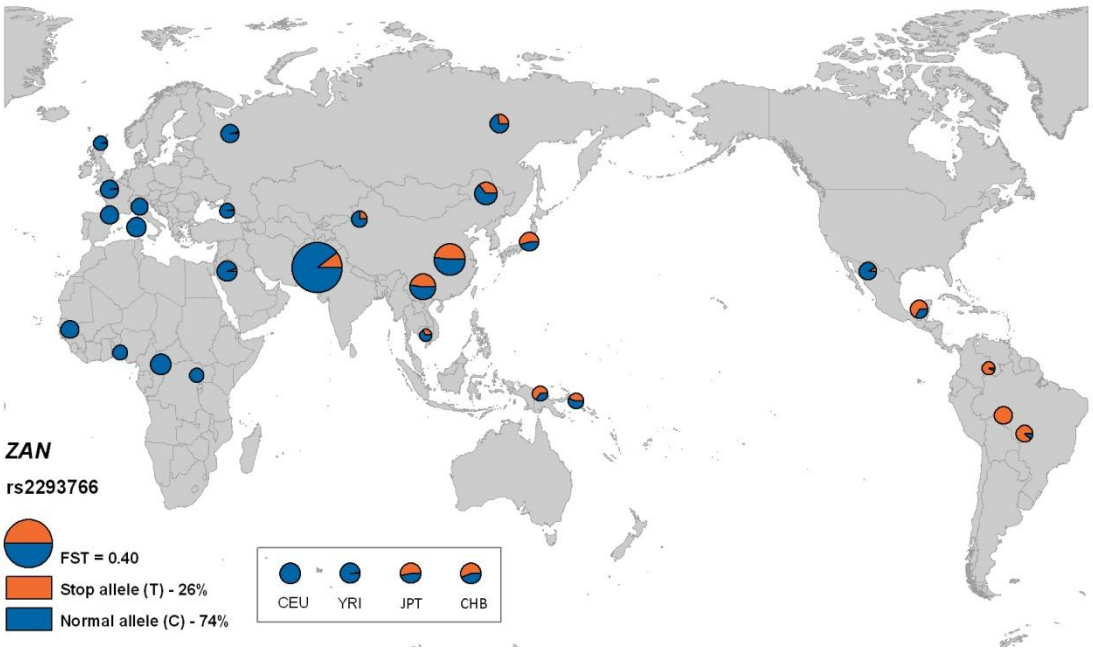


Figure 27 Geographical distribution of alleles of nonsense SNP rs2293766 in *ZAN*.

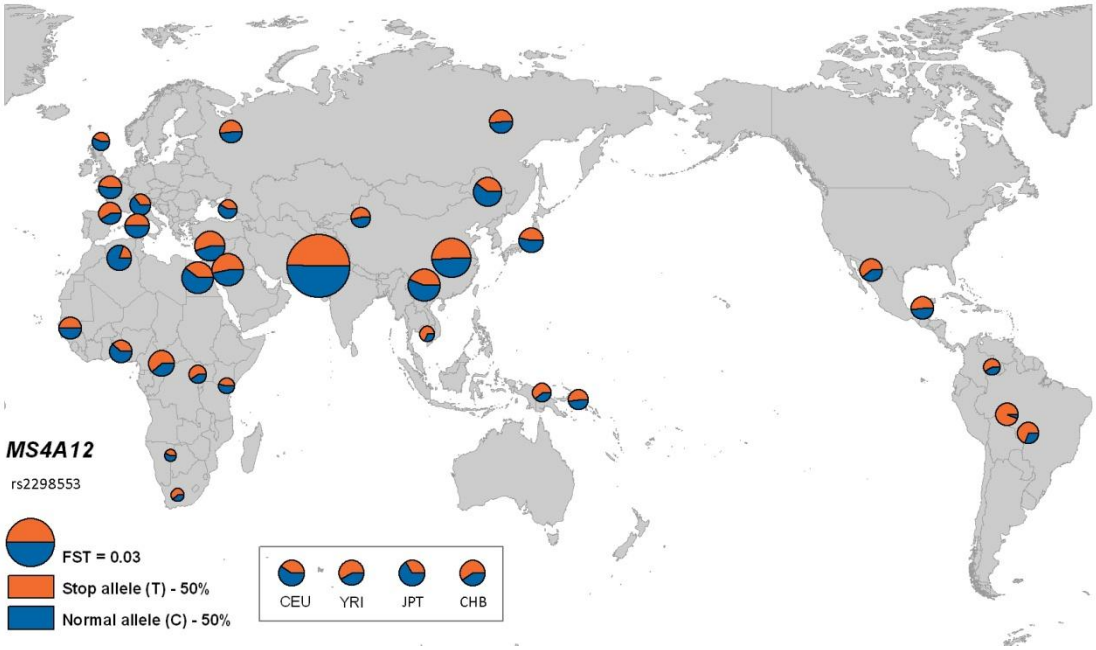


Figure 28 Geographical distribution of alleles of nonsense SNP rs2298553 in *MS4A12*, given as an example of random distribution.

Genotyping errors can theoretically be a source of unusually high F_{ST} values (Barreiro et al. 2008; Xue et al. submitted). While this is a known problem in large public SNP databases, we believe that our quality control methods were satisfactory and that such errors should therefore not influence our conclusions.

3.1.6 Extended Haplotypes

The REHH test (Sabeti et al. 2002) is a good indicator of some forms of recent positive selection as it detects long haplotypes that are more frequent than expected. Our nonsense-SNPs were selected as cores and we then tested regions of 100 kb on either side.

We found that our nonsense-SNPs as a class did not exhibit unusually extended haplotypes, but identified a few outliers above the 95th percentile (Figure 29). Among them are *NPPA*, which was mentioned before with a high frequency of stop homozygotes (DAF = 0.85), *LPL* that encodes lipoprotein lipase and has been implicated in disorders of lipoprotein metabolism, *CD36* which is a thrombospondin receptor and *HPS4* which encodes the Hermansky-Pudlak syndrome 4 protein.

A previous study observed a significant excess of long-range haplotypes among those non-synonymous SNPs with high F_{ST} values (Barreiro et al. 2008). However, only *CD36* identified here was also reported with a high F_{ST} value (F_{ST} = 0.24), the others had values below 0.15. Reassuringly, *CD36*, had also been identified with a long-range haplotype in the HapMap study (The International HapMap Consortium 2005). It should be noted, however, that *MAGEE2* was not included in the REHH analysis because it was on the X chromosome.

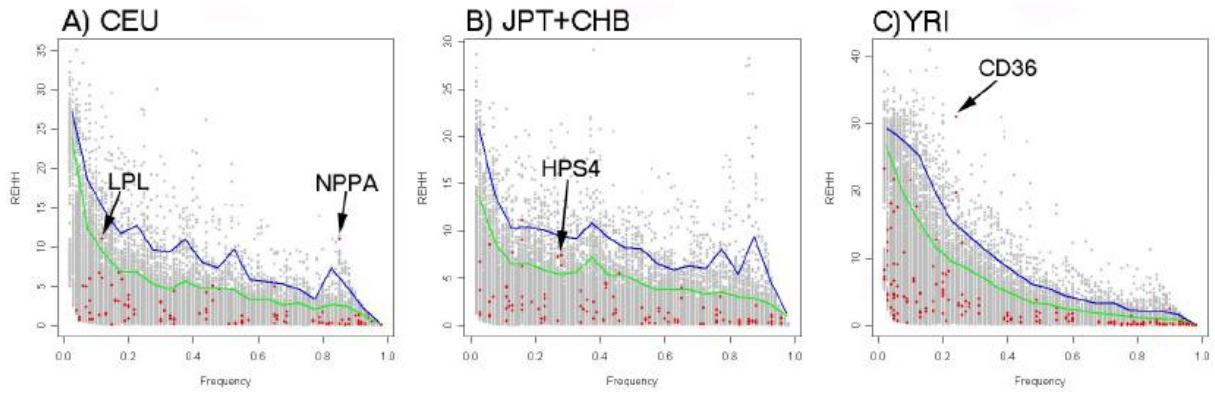


Figure 29 Relative extended haplotype homozygosity (REHH) versus frequency distribution. REHH is plotted against the frequency of each SNP in each HapMap sample, **A)** CEU, **B)** JPT+CHB and **C)** YRI. The grey dots represent the controls (30 ENCODE random regions while the red dots are the stop alleles. Green and blue lines represent the 95th and 99th percentiles, respectively. Some outliers are shown with their relevant gene name.

3.2 CONCLUSIONS

In conclusion, we find that nonsense-SNPs are surprisingly prevalent in the general human population, in contrast to previous reports of such SNPs being infrequent in the human genome (Sawyer et al. 2003). Although they are in the main rarer than near-neutral synonymous-SNPs, they lead to substantial variation in gene content between healthy individuals. People differ by 24 genes, on average, because of nonsense-SNPs, over 0.05% of their gene number.

3.2.1 The Issue of Ascertainment Bias

As the SNPs used in this study were selected from public databases (dbSNP), they will undoubtedly be affected by an ascertainment bias towards higher frequency variants, and thus reduce representation of rare and population-specific variation. As a result, this bias could in itself mimic or complicate any analysis based on the allele frequency spectrum (Walsh et al. 2006), such as F_{ST} and DAF. The REHH results should, however, be less sensitive to ascertainment bias, as LD is expected to be largely unaffected by frequency ascertainment bias (The International HapMap Consortium 2005)

To overcome the problem of ascertainment bias, statistical tests have been tailored to incorporate a known SNP ascertainment scheme in their simulations (Nielsen et al. 2004; Schaffner et al. 2005; Walsh et al. 2006). Unfortunately, it is not possible to apply this type of correction to our data as we cannot be certain of the way the SNPs used were originally ascertained. However, all is not lost because the nonsense-SNPs and synonymous-SNPs used here were picked from the same source, so should be affected in the same way. In addition, as the nonsense- and synonymous-SNP data origins are largely overlapping, it is possible to identify nonsense-SNP outliers relative to the synonymous-SNPs. Therefore, we believe that any signal picked up by these methods should be treated as potentially interesting, which can then be followed up by the more sophisticated tests based on re-sequenced data as will be discussed in the next chapter.

3.2.2 Allele Frequency Spectra

While most stop alleles were rare in our world-wide sample of populations, some interesting outliers were observed, such as *FMO2*, *NPPA* and *CASP12*. The most exciting possible explanation for a high frequency stop allele is that it has become advantageous to lose its gene and that the allele has been driven to high frequency by positive selection. Other explanations have been noted elsewhere (Savas et al. 2006). For example, the nonsense-SNP might not be deleterious because the protein product is still functional (e.g. if the stop SNP is close to the natural stop codon and NMD is not triggered) or the protein may simply not be essential for the fitness of humans. Figure 17 displays a number of nonsense-SNPs not expected to trigger NMD while causing truncations and this may indicate that they are not having severe effects on their carriers. Indeed, the derived allele frequency spectrum for nonsense-SNPs expected to trigger NMD and those not expected to do so were significantly different (Kolmogorov-Smirnov, $P<0.01$), with a mean DAF of 0.106 and 0.163, respectively. Another possibility is that the nonsense-SNP is sometimes deleterious but is tolerated because of heterozygote advantage. While this has probably been the case for some alleles such as those that reduce G6PD activity and increase resistance to malaria, we do not see evidence for this in our data, as we do not see an excess of heterozygotes. In principle, epistatic interactions with other mutations, either in the same or different genes, might compensate for the slightly deleterious effects of the nonsense-SNPs. This possibility would need further investigation into the function of these genes and variants in surrounding regions.

As noted before, most stop alleles were found at a low frequency (see Figure 12) and the most likely explanation for the overall pattern is that these alleles are generally deleterious and subjected to negative selection. Deciding whether an individual allele is disadvantageous or neutral would require further studies into the function of the gene and association of variations found within. If the nonsense mutation is relatively new in the human population, it would start at a low

frequency and in order to establish the degree of negative selection it would help to estimate the age of each mutation. Lastly, a low allele frequency could be expected in cases where the sample size is not large enough or when there are errors in the genotyping. We are satisfied with the quality controls used for determining the genotypes in this study and believe the population size and geographical distribution of the samples is sufficient. Therefore, we do not believe that this last possibility has had an effect on our data.

3.2.3 Population Differentiation

While some have argued that population differentiation, as measured by the F_{ST} statistic, is not a good estimator for population-specific positive selection (see e.g. Gardner et al. 2007) a recent study re-analysed a set of SNPs previously identified with high population differentiation in the HapMap data/samples and found support for the opposite. The SNPs were genotyped in a larger set of populations, extended haplotype homozygosity was measured and a proportion of genes (containing the SNP) was fully re-sequenced to allow for sequence-based neutrality tests. When technical artefacts, such as genotype errors, had been excluded, the conclusion was that high F_{ST} values (with the support of other lines of evidence) were consistent with a sign of population-specific positive selection rather than random genetic drift (Xue et al. submitted).

Overall, our set of nonsense-SNPs was found to have lower F_{ST} values than synonymous SNPs. This is in accordance with a previous study (Barreiro et al. 2008) that showed that nonsynonymous-SNPs had an excess of low values compared to synonymous-SNPs. The same study reported that SNPs with low values were shown to be more frequent in genes known to modulate disease. As the majority of our nonsense-SNPs were reported with low F_{ST} values we assume that, given their expected deleterious nature, this is most likely a result of negative selection.

However, some were found with higher levels of population differentiation and these deserve further investigation. Two SNPs with high F_{ST} values were found in

genes with interesting biological features according to the literature. The highest was that of the *MAGEE2* gene (F_{ST} = 0.54) which will be discussed in greater detail in Chapter 4. The other gene, *SIGLEC12*, was not such an extreme outlier (F_{ST}=0.22) but it was highly truncated (~95%) and belongs to a family of sialic acid-binding immunoglobulin-like lectins (SIGLECs), some of which are enriched in the brain and in epithelial cells (Hayakawa et al. 2005) and many of which have become inactivated in humans (Angata et al. 2001). This gene loss event has resulted in humans being unable to produce a sialic acid called N-glycolylneuraminic acid (Neu5Gc). With such an interesting candidate we intended to follow *SIGLEC12* up, along with *MAGEE2*, but unfortunately the low quality of the sequence traces generated prevented us from making any use of the data.

3.2.4 Extended Haplotypes

Apart from a few noticeable outliers, we did not find extensive evidence for unusually extended haplotypes in the nonsense-SNPs, which further indicates (perhaps unsurprisingly) that the majority of these SNPs are not positively selected. While the REHH has on several occasions provided good evidence for selection (Sabeti et al. 2002; Sabeti et al. 2007; Voight et al. 2006), the test may be limited in its ability to detect only recent selective events (occurring less than 30 KYA , perhaps less than 10 KYA) and in its power when an allele is at high or low frequency, or when selection has acted on more than one haplotype. For example, the test failed to show an REHH signal for the Duffy Fy*O locus (Walsh et al. 2006) which has been shown biologically to have been under strong positive selection as a result of its association with resistance to malaria (Hamblin and Di Rienzo 2000; Hamblin et al. 2002; Livingstone 1984).

3.2.5 Overrepresented Functions

The GO analysis revealed an excess of genes involved in olfactory reception and the nervous system. As noted, the first category was expected to show up as previous

studies indicate that humans have a reduced sense of smell (Gilad et al. 2003a; Gilad et al. 2003b). Indeed, a recent study on nonprocessed pseudogenes inactivated in the human lineage reported an overrepresentation of genes involved in chemoreception (which olfactory receptors belong to) and immune response (Wang et al. 2006). The latter, however, was not observed in our study. Finding an overrepresentation of genes involved in the nervous system was, however, unexpected as such genes have generally been shown to be very conservative (Nielsen et al. 2005).

Genes related to the senses, including a group of olfactory receptor genes have been identified in CNV regions (Wong et al. 2007), suggesting that these have duplicated paralogs and are therefore backed up. However, as was demonstrated with the *ACTN3* gene in section 1.5.4, while the function of a lost gene may be compensated by a closely related gene, its loss can still have significant consequences.

Overall, the analyses described here have identified the general characteristics of the class of nonsense-SNPs, and also pinpointed a small number of nonsense-SNPs that appeared to be exceptional. These included Tables 6, 7, and 11, and from this list, *CASP12*, *MAGEE2* and *SIGLEC12* were chosen for more detailed study, although as mentioned, only the first two yielded sufficiently high-quality data to allow detailed follow-up. These results are described in the next chapter.

4 Detailed Analyses of Individual Genes

The main indicators of positive selection used so far in this thesis, high frequency derived alleles and high levels of population differentiation in individual nonsense-SNPs, are indirect and can readily arise in other ways as well. This chapter describes data from resequencing two examples of interesting genes, *CASP12* and *MAGEE2*, so that additional tests could be used to investigate whether the unusual characteristics were found in extended regions of DNA surrounding each nonsense-SNP and if so were likely to have arisen by neutral processes, or whether positive selection would provide the best explanation.

The *MAGEE2* gene came up as an interesting outlier in our genome-wide survey of nonsense-SNPs described in chapter 3. The *CASP12* gene was, however, analysed before embarking on the main part of this study and the results were published by Xue et al (2006). The genotyping of nonsense-SNP rs497116 in *CASP12* and the subsequent analysis was performed by myself, while the resequencing part was performed by Yali Xue who then performed the analysis on the variation data.

In this chapter, I will refer to the different nonsense-SNP alleles in *CASP12* as "inactive" and "active" (instead of "stop" and "normal") as the functional consequences of the mutation is known. For the *MAGEE2*, for which I have no functional information, I will continue to refer to the stop and normal allele at the nonsense-SNP as was done in the previous chapter.

4.1 Results

4.1.1 CASP12

The human caspase-12 gene (*CASP12*) is on chromosome 11 and has been shown to modulate inflammation and innate immunity in humans (Saleh et al. 2004). The variation at the nonsense-SNP, rs497116, in the *CASP12* gene produces two versions of the protein which exist in human populations (see Figure 30), a truncated inactive

version and a full-length active version (Fischer et al. 2002). We find that the gene is truncated by ~64% and that the nonsense-SNP is expected to trigger NMD. A previous study (Saleh et al. 2004) found that the active ancestral version (considered by them an unusual 'long' variant) was only present in populations of African descent and was associated with a reduction in levels of cytokines after stimulation by bacterial lipopolysaccharides, leading to a lower initial immune response. However, carrying the active allele was found to increase susceptibility to developing severe sepsis, a later over-reaction of the immune system, as well as resulting in higher mortality rates once sepsis had developed. The truncated derived version, on the other hand, was associated with lower levels of severe sepsis and was found to be nearly fixed in human populations.

As there was a limited amount of information available on the evolutionary history of *CASP12*, we decided to investigate whether the inactive form had spread by neutral genetic drift or whether this was a case of selective advantage associated with gene loss. The *CASP12* nonsense-SNP did not show up as an extreme outlier in our survey of nonsense-SNP, except for a high DAF and moderate F_{ST} value of 0.24, which was not found to be significant when compared to empirical distributions.

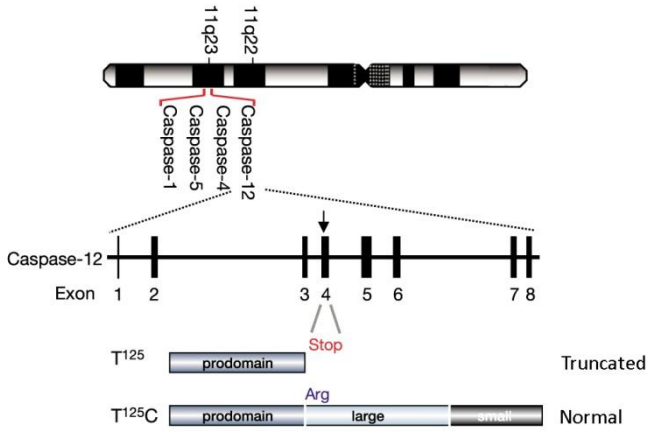


Figure 30 Map of the region at 11q23 containing the *CASP12* gene. The exon–intron organization of *CASP12* is shown. The arrow indicates the nonsense-SNP (rs497116) which changes an Argenine residue into a premature stop codon. The two products resulting, a truncated inactive version caused by the stop allele and a full-length active version, are displayed. This figure is adapted from (Saleh et al. 2004).

In accordance with previous results, we found the inactive allele to be nearly fixed in the human species with an overall frequency of 96%, while the active allele was mainly found in African populations (see Figure 31). Mbuti Pygmies and San have the highest frequencies of the active allele – 60% and 57%, respectively. Outside Africa, the active allele was very rare but was detected at low frequencies in Israel, Pakistan and China. No disagreement with HWE was observed in individual populations, but the pooled sample departed significantly from HWE (Chi-squared, P<<0.001), reflecting population subdivision.

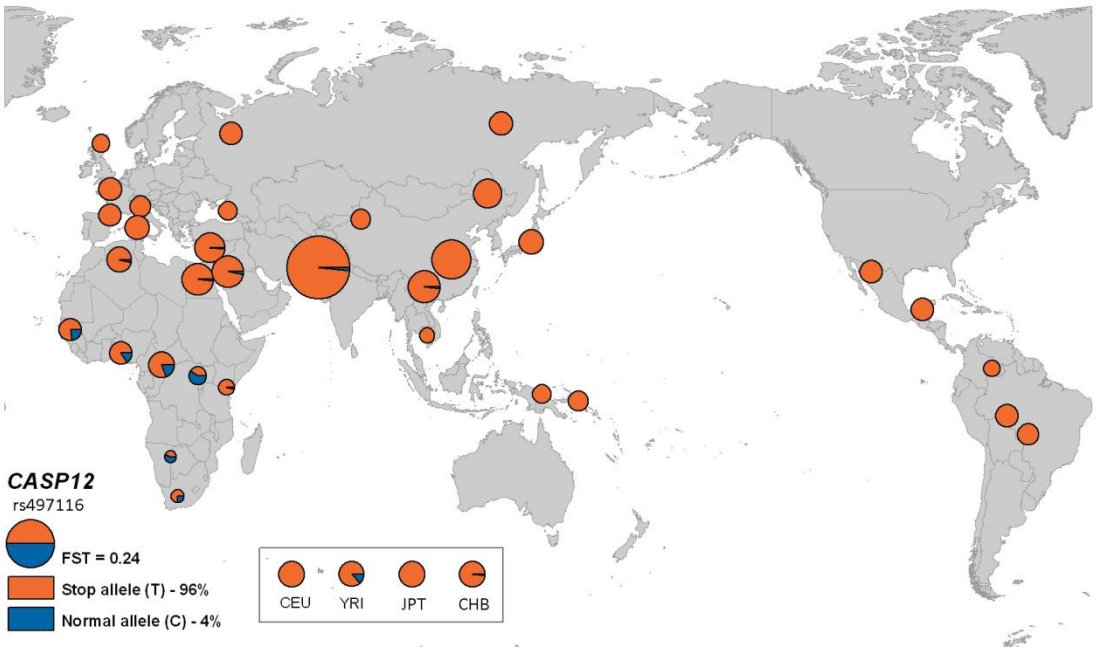


Figure 31 Geographical distribution of inactive ("stop") and active ("normal") alleles of the CASP12 nonsense-SNP (rs rs497116). The stop allele is nearly fixed in the human population and the normal allele is mainly found in African populations. The stop allele is represented in orange and the normal allele in blue. Pies are proportional to sample sizes.

4.1.1.1 Sequence Variation in CASP12

We next wanted to determine whether the observed predominance of the inactive allele was due to positive selection, or if it was the result of a neutral variant rising in frequency, for example because of the bottleneck associated with the human migration out of Africa. To this end, we resequenced a 13.3-kb region that covers the whole *CASP12* gene and an additional ~0.7 kb on each side of it in 77 individuals

from three HapMap populations (26 YRI, 26 CHB and 25 CEU). Our sample thus consisted of 155 chromosomes (154 from the HapMap samples in addition to the reference sequence which is of unknown origin). Eight were found to carry the active allele – six YRI, one CHB and the reference sequence, which is similar to the worldwide geographical distribution observed in Figure 31. The remaining 147 chromosomes carried the inactive allele. A total of 123 SNPs were detected (Appendix F.1.) but these were distributed very unevenly among the different versions of the gene and populations. We inferred the haplotypes from the SNP data and found that the active genes were much more diverse: the eight chromosomes carried 61 SNPs and showed a nucleotide diversity (π) of 19.7×10^{-4}, whereas the 147 inactive chromosomes carried 76 SNPs and had a nucleotide diversity almost 10 times lower, 2.0×10^{-4} (see summary in Table 12). This led to higher diversity in the YRI ($\pi = 9.1\times10^{-4}$) than in the other populations, 1.9×10^{-4} and 0.5×10^{-4} in the CHB and CEU, respectively—a ratio more extreme than any encountered in a study of 132 genes in African American and European American populations (Akey et al. 2004). The inactive version was also more diverse in Africa than outside ($\pi = 4.4\times10^{-4}$ and $\pi = 0.7\times10^{-4}$) which is in accordance with most other studies on diversity within and outside Africa (Prugnolle et al. 2005; Rosenberg et al. 2002; The International HapMap Consortium 2005). The low diversity of the inactive form, particularly outside Africa, provided the second indication that their spread might have been rapid and thus due to positive selection.

4.1.1.2 Long-Range Haplotype Tests (CASP12)

We performed the REHH test (Sabeti et al. 2002) on four HapMap populations, CEU, CHB+JPT and YRI (described in section 2.3.8.4), and did not find any evidence for unusually extended haplotypes at high frequencies in *CASP12*. This test compares the suspected positively selected allele to the other allele at the same position and therefore relies on the SNP being polymorphic. As the inactive allele is fixed in CEU it is not possible to calculate REHH for this population. The inactive allele is also

nearly fixed in the Asian populations as only one active allele is reported in the combined populations of CHB and JPT. Again it is impossible to make inferences from such a low frequency. The YRI, however, have 14 copies of the active allele and thus it is possible to visualize the haplotypes in Haplotter (Figure 32) which is based on a LRH test which calculates the Integrated Haplotype Score (iHS) (Voight et al. 2006). In Figure 32A, a continuous block of the same colour represents a haplotype, and if it is shared by many chromosomes it will be thick. Indeed, there is some indication of such a block for the inactive allele (represented in red), but this is seen mainly on one side and does not include all chromosomes. However, while we see some indication of a long haplotype as was reported by Xue et al (2006), it must be compared with the ancestral allele haplotypes (blue). This also shows a long haplotype on one side for a proportion of chromosomes and thus LD seems to be similar for the active and inactive allele (Figure 32B).

We therefore conclude that LRH tests do not give us any evidence for positive selection of the *CASP12* nonsense-SNP. The low frequency of the active allele does not allow us to apply such tests in any population except the YRI, where it appears that either no such signature was formed, or enough time has passed for recombination to break up the long range structure. In fact, the age of the inactive allele was estimated between ~100-500 KYA in Xue *et al* (2006) which is too old for the LRH tests to detect a signal.

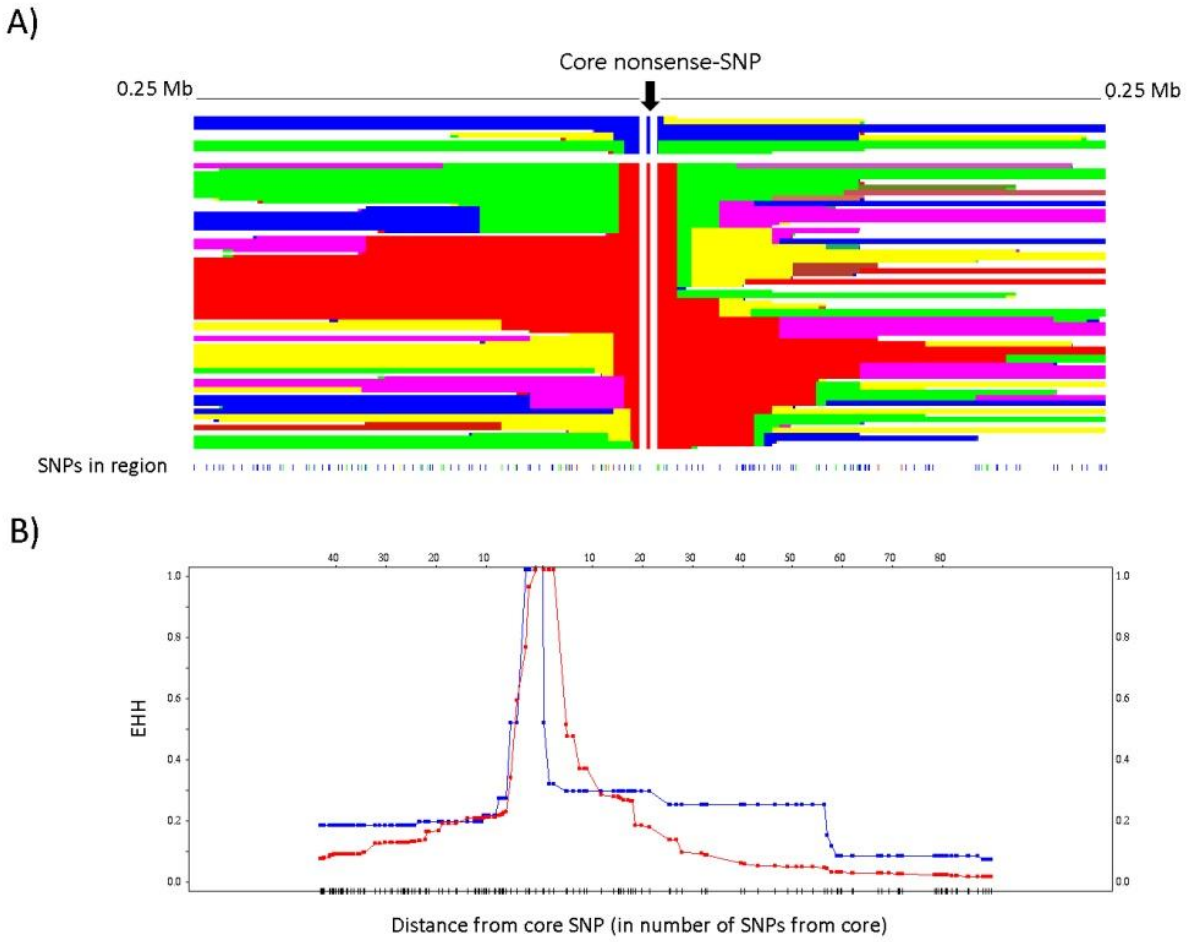


Figure 32 *CASP12* haplotypes in YRI A) Haplotypes at different distances from the nonsense-SNP (core) at the centre. Each horizontal line represents the haplotype of each chromosome. The blue vertical line represents the ancestral state (active allele) and the derived state (inactive allele) is represented in red. The distances over which the haplotypes are spread is displayed at the top of the graph. The total region size displayed on the top is 0.5 Mb and the SNPs in the region are showed at the bottom. **B) The decay of Extended Haplotype Homozygosity (EHH) at different distances from the nonsense-SNP (core).** The decay starts increasing at a short distance from the nonsense-SNP, for both the inactive (red) and active (blue) allele.

4.1.1.3 Neutrality Tests (CASP12)

Neutral models of evolution provide predictions of expected allele-frequency characteristics, and observed patterns can be compared with these. We calculated Tajima's *D* (Tajima 1989c), Fu and Li's *D* and *F* (Fu and Li 1993), and Fay and Wu's *H* (Fay and Wu 2000). The results are summarized in Table 12. Neutrality is rejected for *CASP12* by all tests in the combined populations. In individual populations, neutrality is similarly rejected by all tests for the CHB, but only by Tajima's *D* and Fay and Wu's *H* for the YRI and by Tajima's *D* for the CEU. These results can readily

be understood in terms of a selective sweep that has proceeded to different stages in the different populations, as will be discussed later.

Location	Sample characteristics			Allele frequency distribution tests				Haplotype test
	Sample size (chr)	Number of polymorphic sites	Nucleotide Diversity (π) ($\times 10^4$)	Tajima's *D*	Fu & Li's *D*	Fu & Li's *F*	Fay & Wu's *W*	Fu's F_s
All populations	155	123	4.5	-2.32*	-2.75*	-3.06**	-46.2**	-27.7**
YRI	52	99	9.1	-1.59*	-1.05	-1.54	-28.7*	-5.8
CEU	50	7	0.5	-1.57*	-1.17	-1.54	-0.9	-6.6**
CHB	52	47	1.9	-2.60**	-3.20**	-3.59**	-33.5**	-5.2
Active (all$^{\alpha}$)	8	61	19.7					
Inactive (all$^{\alpha}$)	147	76	2.0					
Inactive (African)	46	57	4.4					
Inactive (non-African)	101	21	0.7					

Table 12 Summary statistics for CASP12. $^{\alpha}$All samples (YRI, CEU, CHB and reference sequence). *P<0.05 **P<0.01 (one-sided tests).

Another type of neutrality test examines haplotypes rather than single variable positions. A total of 36 haplotypes were identified, but one haplotype carrying the inactive allele occurred 99 times and accounted for 64% of the sample (and 76% of non-African chromosomes). Fu's F_s test (Fu 1997) shows that significantly fewer haplotypes are found in the whole sample and in CEU than expected under neutrality (Table 12).

We conclude that sequence variation in *CASP12* is significantly different from that expected under neutrality. Departures from neutral expectation at a single locus can arise by demographic processes; however, the neutral model used in these evaluations incorporated the best-fit demographic model. Thus the most likely explanation for all these deviations is positive selection.

4.1.1.4 CASP12 Network

A median-joining network was constructed to show the relationships between the inferred haplotypes (Figure 33). The eight haplotypes carrying the active allele were all different from one another and from those carrying the inactive allele. All of the inactive allele haplotypes clustered together, with 99 chromosomes at the center of the cluster, 29 one step away, 6 two steps away, and a few more distant. Outside

Africa, the most distant inactive haplotype was only three steps from the center, whereas there was more diversity among the inactive haplotypes in Africa, and not all radiated directly from the central haplotype.

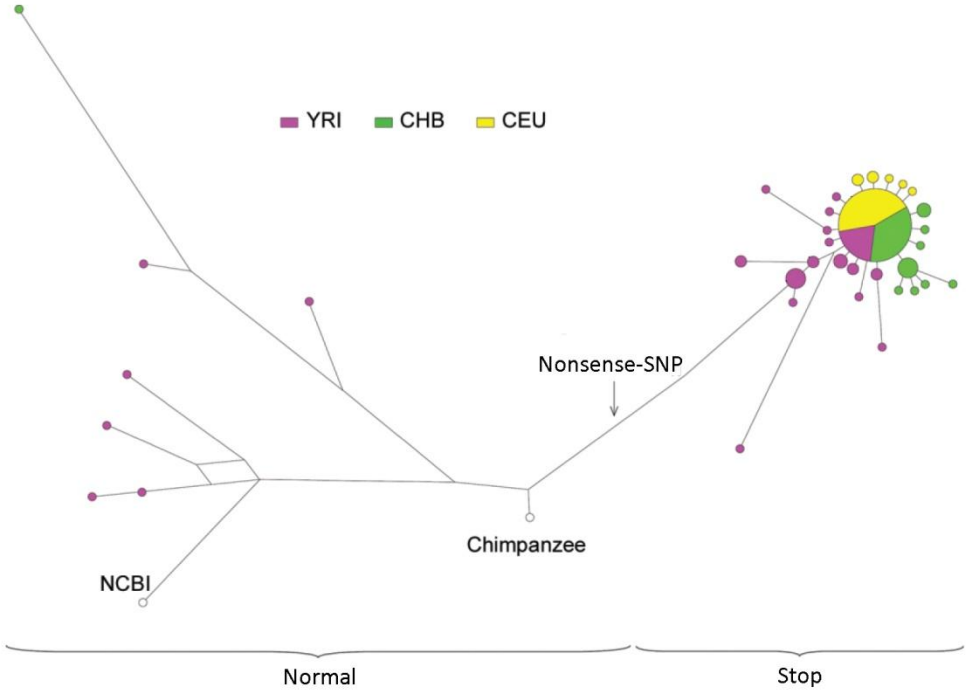


Figure 33 Median-joining network of inferred *CASP12* haplotypes. Circle areas are proportional to the haplotype frequency and are colour coded according to population; YRI (purple), CHB (green), and CEU (yellow). Lines represent mutational steps between them; the shortest lines equal one mutation. Location of nonsense-SNP (rs497116) is indicated with an arrow and the cluster of active ("normal") and inactive ("stop") haplotypes is labelled at the bottom. The NCBI reference sequence and the chimpanzee outgroup are labelled.

4.1.2 MAGEE2

The *MAGEE2* gene is a melanoma-associated antigen which belongs to a family of *MAGE* genes that are found predominantly on the X chromosome. Several members of the *MAGE* gene family (including *MAGEE2*) are expressed in tumour cells but are silent in normal adult tissues except in the male germ line, leading to an alternative name for these genes, cancer-testis genes. Because of their specific expression on tumour cells, these antigens are potential targets for cancer immunotherapy (Chomez et al. 2001; Ross et al. 2005), but their normal function is completely unknown.

The nonsense-SNP (rs1343879) in *MAGEE2* was identified with the highest F_{ST} value (0.54) in our set of world-wide populations which might be a sign of positive selection. When the F_{ST} value was compared to other empirical F_{ST} values (section 2.3.8.3), it was found to be significantly high (above the 99th percentile) in the HGDP-CEPH populations and in the HapMap. In addition, the geographical distribution of the stop allele (Figure 34) showed an interesting pattern and so taken together this provoked our curiosity about the evolutionary history of the gene. The nonsense-SNP in our dataset was found to truncate the gene by ~77% and yet NMD is not expected to be triggered.

The stop allele (A) has the highest frequency in Asian and Central American populations and is virtually absent from European and African populations. The geographical distribution reveals an east-west gradient of the derived stop allele which may have arisen in the east before the Asian ancestral populations migrated into the Americas less than 20 KYA.

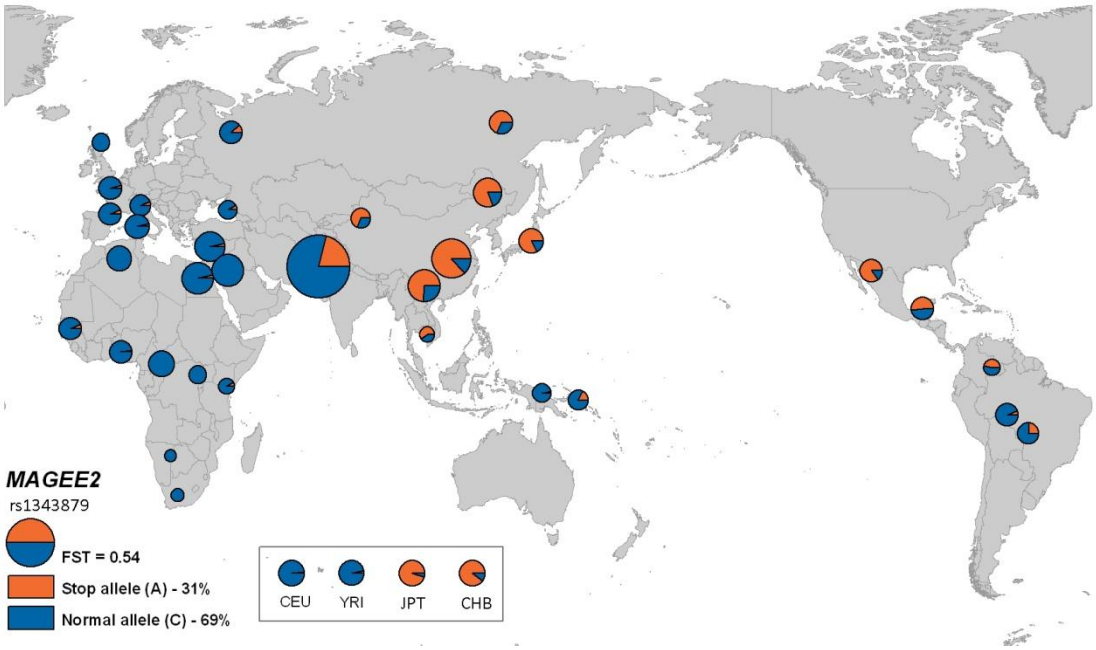


Figure 34 Geographical distribution of stop (orange) and normal (blue) alleles in *MAGEE2*. HapMap populations are displayed separately as they do not have precise geographic locations. Pies are proportional to sample sizes.

4.1.2.1 Sequence Variation at MAGEE2

Next, we wished to explore whether the higher frequency of the *MAGEE2* stop allele in certain populations was the result of population-specific selection or simply of random genetic drift. Therefore, we resequenced a ~12 kb region that covers the whole *MAGEE2* gene and an additional ~5 kb on each side of it in 91 individuals from three HapMap and one extended HapMap population (23 YRI, 23 CHB, 22 CEU and 23 LWK) together with one chimpanzee. 32 chromosomes were found to carry the stop allele – 1 YRI, 28 CHB, 1 CEU and 2 LWK, and these are similar proportions to the worldwide geographical distribution observed in Figure 34. The remaining chromosomes carried the normal allele. A total of 43 SNPs were detected (Appendix F.2.) in the *MAGEE2* gene. We inferred the haplotypes from the SNP data and found that the haplotypes carrying the stop allele were much less diverse than the normal ones, but that the normal were not as diverse as one might have expected (Sachidanandam et al. 2001). Among the 79 chromosomes carrying the normal allele we identified 36 SNPs and a nucleotide diversity (π) of 3.7×10^{-4},, while the 32 inactive haplotypes only carried 8 SNPs and had a nucleotide diversity of 0.8×10^{-4} (see

summary in Table 13), which was even lower than the diversity of the stop allele chromosomes in *CASP12* 2.0×10^{-4}. Again we see a higher diversity in the African populations ($\pi = 4.3\times10^{-4}$ in YRI and $\pi = 4.7\times10^{-4}$ in LWK) compared to the CEU ($\pi = 2.9\times10^{-4}$) and CHB ($\pi = 1.6\times10^{-4}$), but this ratio is not as extreme as that found in the *CASP12* gene (see section 4.1.1.1). The lower diversity observed for the truncated version is consistent with positive selection, but to explore this possibility further we needed to apply more tests.

4.1.2.2 Long-Range Haplotype Test (MAGEE2)

Unfortunately, the *MAGEE2* nonsense-SNP was not included in our REHH analysis of nonsense-SNPs reported in section 2.3.8.4. This is because *MAGEE2* lies on the X chromosome and the process of phasing haplotypes needs to be done differently to that for autosomal SNPs because of the different copy number of the X in males and females. Therefore, we were unable to use the phased HapMap data for this SNP; male X chromosomes are perfectly phased, but were too few in number. To compensate for these factors we made use of Haplotter (Voight et al. 2006) again to see if this nonsense-SNP was associated with unusually long haplotypes. Again our results are affected by the polymorphic distribution of the SNP in the HapMap populations. The stop allele was virtually absent from CEU and YRI (2.2% and 1.1% respectively) while nearly fixed in the combined CHB+JPT (91%). The results are displayed in Figure 35. The low frequency of the stop allele in YRI made it difficult to make any inferences from the LRH figures which are thus not shown. However, we looked at a region spanning 2.5 Mb around the nonsense SNP in CEU (Figure 35A and B) and CHB+JPT (Figure 35C and D). We found haplotype blocks surrounding the high frequency normal allele in CEU that seems to decay rapidly in about half the chromosomes. In the CHB+JPT, the haplotypes associated with both alleles are long, but decay is noticeably slower for the stop allele and is also much slower than is observed for the normal allele in CEU, which could indicate that recent positive selection has acted on the stop allele in the CHB+JPT populations.

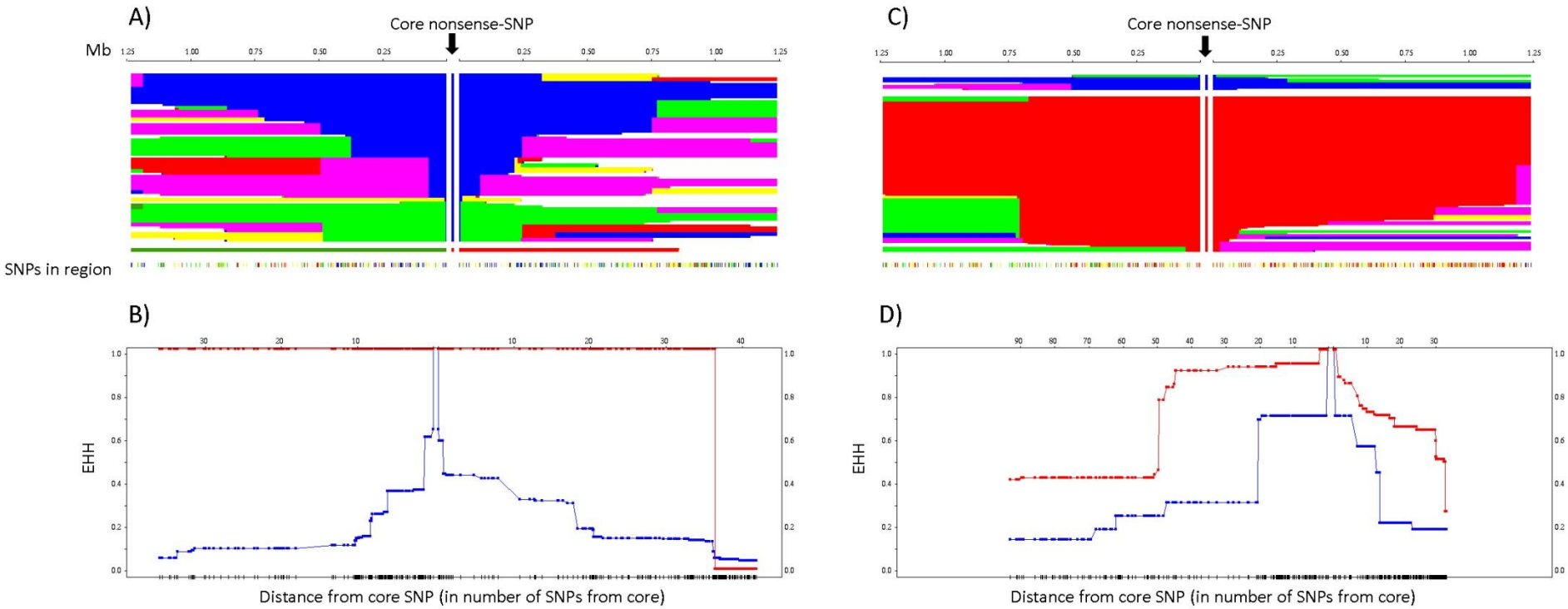


Figure 35 Long-range haplotypes analysed for region surrounding core nonsense-SNP in *MAGEE2* in CEU and CHB+JPT. Haplotypes at different distances from the nonsense-SNP (core) at the centre are displayed for A) CEU and C) CHB+JPT. Each horizontal line represents the haplotype of each chromosome. The blue vertical line represents the ancestral state (normal allele) and the derived state (stop allele) is represented in red. The distances over which the haplotypes are spread is displayed at the top of the graph. The total region size displayed is 2.5 Mb and the SNPs in the region are showed at the bottom**. The decay of Extended Haplotype Homozygosity (EHH) at different distances from the nonsense-SNP (core) is displayed for B) CEU and D) CHB+JPT.**

4.1.2.3 Neutrality tests (MAGEE2)

We calculated Tajima's *D* (Tajima 1989c), Fu and Li's *D*, *D**, *F* and *F** (Fu and Li 1993), and Fay and Wu's *H* (Fay and Wu 2000) and neutrality could not be rejected for any of these in either the combined or individual populations, except for Fay and Wu's *H* in the CHB population (see Table 13). Fay and Wu's *H* uses the derived allele frequency spectrum to search for evidence of departures from neutrality (Fay and Wu 2000). This could suggest a population-specific selective sweep acting on the stop allele in the CHB population but not in the others. Therefore, we decided to look at the nucleotide diversity of the inactive haplotype in CHB alone, but found that this was not different from that of the combined populations ($\pi = 1.4\times10^{-4}$).

We then analysed the haplotypes more closely with Fu's F_s test (Fu 1997) and found that significantly fewer haplotypes are found in the whole sample than would be expected under neutrality but not for the individual populations (Table 13).

Overall, most of these tests imply that sequence variation in *MAGEE2* is not significantly different from that expected under neutrality and, except for Fay and Wu's *H*, would be consistent with the idea that the stop allele has risen in frequency in the Asian populations as a consequence of genetic drift.

	Sample characteristics			Allele frequency distribution tests						Haplotype test
Location	Sample size (chr.)	Number of polymorphic sites	Nucleotide Diversity (π) ($\times10^4$)	Tajima's D	Fu & Li's D	Fu & Li's D*	Fu & Li's F	Fu & Li's F*	Fay & Wu's H	Fu's Fs
All$^{\alpha}$	111	43	4.2	-1.24	-2.20	-2.28	-2.16	-2.23	0.42	-27.03**
YRI	26	22	4.3	-0.49	-0.07	0.00	-0.26	-0.18	3.10	-4.25
LWK	21	21	4.7	-0.24	0.26	0.29	0.12	0.16	2.93	-4.15
CEU	33	17	2.9	-0.68	-1.58	-1.35	-1.54	-1.34	1.05	-2.36
CHB	31	11	1.6	-1.10	-0.14	-0.58	-0.54	-0.87	-8.32**	-3.54
Active (all$^{\alpha}$)	79	36	3.7							
Inactive (all$^{\alpha}$)	32	8	0.9							
Inactive (CHB)	28	7	0.8							

Table 13 Summary statistics for MAGEE2. **P<0.01 (one-sided tests, empirical distribution from the best-fit model). αAll samples (YRI, LWK, CEU, and CHB).

4.1.2.4 MAGEE2 Network

A median-joining network was constructed from the inferred haplotypes of *MAGEE2* (Figure 36). As was seen in the geographical distribution of the nonsense-SNP (Figure 34) there is a clear east-west division for the haplotypes which is caused by the nonsense-SNP. All haplotypes carrying the inactive form in the CHB population cluster together (inside red circle in Figure 36) where there is one high-frequency haplotype with the other nonsense-allele haplotypes only one or two steps away. This pattern helps to explain the significantly negative value of Fay and Wu's *H* in the CHB sample by illustrating the moderately high frequency of a derived haplotype cluster specific to the CHB.

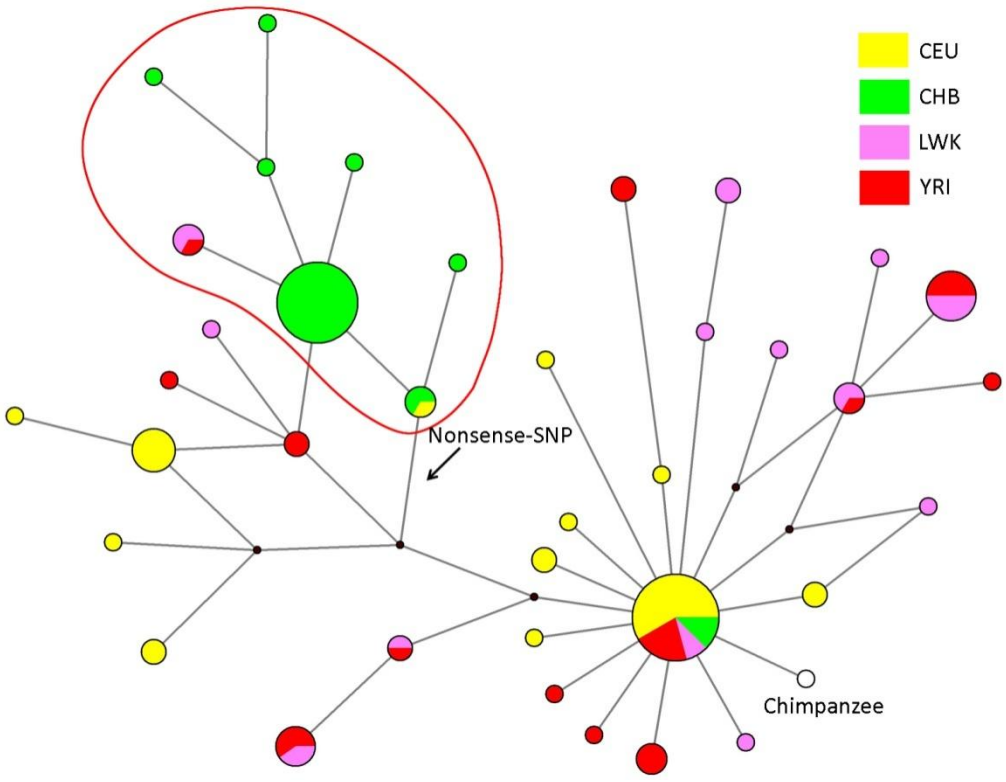


Figure 36 Median-joining network of inferred *MAGEE2* haplotypes. Circle areas are proportional to the haplotype frequency and are colour-coded according to population; CEU in yellow, CHB in green, LWK in pink and YRI in red. Lines represent mutational steps between them (one or two steps, according to length). Arrow shows the location of nonsense-SNP (rs1343879).

4.2 CONCLUSIONS

These results illustrate the information that can be obtained by more detailed resequencing studies of individual genes. *CASP12* provides one of the clearest examples of positive selection thus far identified in the human genome. Selection has carried the stop allele to a high frequency in the YRI, so significant values of the summary statistics Tajima's *D* and Fay and Wu's *H* are seen. This allele is present at even higher frequency (although not fixed) in the CHB, so shows even more highly significant values of the statistics. In the CEU sample, it is fixed, with the result that diversity is low and the summary statistics have less power, leading to less significant values. The selective sweep is thus proposed to have proceeded to different stages in the different populations. Long Range Haplotype tests show no evidence for selection, a finding that does not conflict with the allele frequency tests but rather is a consequence of the time of selection (too ancient for an LRH signal) and high frequency of the selected allele (too high for an LRH signal). As carriers of the stop allele are protected against severe sepsis, it is reasonable to propose that avoidance of sepsis, and survival if sepsis develops, has been the selective factor.

MAGEE2 shows a quite different pattern, with only limited evidence for a departure from neutrality and thus for positive selection. While it could be argued that the observations could be explained by drift, it is also worth considering a scenario involving selection. Here, the low frequency of the stop allele in all sequenced samples except the CHB precludes any signal from summary statistics in most populations. In the sequenced CHB, the network shows evidence of rapid expansion of a cluster of haplotypes but the frequency of the stop allele is 31% in the combined populations and the number of SNPs in the sequenced region of the stop allele chromosomes is 8, explaining why the summary statistics are less significant than for the *CASP12* stop allele in the YRI. This low frequency might be due to a more recent origin of the nonsense-SNP in *MAGEE2* than for *CASP12*, and an LRH signal might thus be expected to reveal positive selection. The combined Asian

populations (CHB+JPT) do in fact display differential decay of LD extending over a large region surrounding the nonsense-SNP. There is another test, the XP-EHH (cross population extended haplotype homozygosity) test (Sabeti et al. 2007), which has been designed to detect selective sweeps when the selected allele is fixed in one population but remains polymorphic across other populations. However, in the case of the *MAGEE2*, the gene was included in the study of Sabeti et al. (2007) and listed on the basis of its high F_{ST} value, but was not associated with any unusual long range haplotype signal, cross-population or otherwise (Sabeti et al. 2007 Table S10).

We propose that the stop allele in *MAGEE2* may have undergone recent positive selection in Asian populations, but without any understanding of the normal function of this gene, it is impossible to speculate usefully about the reasons for selection.

5 Discussion and Future Directions

The main goal of our research was to evaluate the overall evolutionary forces acting on nonsense-SNPs in the human genome and thus provide some insights into the importance of variation in gene number for human evolution. To this end, we embarked on a genome-wide study of loss events and typed a large number of nonsense-SNPs in a set of geographically diverse populations. From this dataset, we hoped to identify candidates for positive selection (to be followed up in more detail by resequencing), and thus provide an evaluation of the relevance of the less-is-more theory for human evolution. We believe we have now accomplished these goals, and in the next few sections will discuss our main conclusions.

5.1 Prevalence and Consequences of Nonsense-SNPs

In chapter 3 we reported the prevalence of nonsense-SNPs in the human genome and their consequences for the protein product. We found that nonsense-SNPs are more prevalent in the human genome than some studies have suggested (Sawyer et al. 2003) and that they are not simply a class of deleterious disease-causing mutations slipping through the system at low frequencies in a heterozygous state. The prevalence of nonsense-SNPs was such that the individuals sampled were found to differ by 24 genes, on average, because of nonsense-SNPs. This will almost certainly be an underestimate and will increase with the findings of large-scale sequencing projects, such as personal genome sequencing projects (Levy et al. 2007; Wheeler et al. 2008) and the more systematic "1,000 genomes" project (http://www.1000genomes.org/). Nevertheless, this is still a higher difference than was reported initially for the more commonly occurring CNVs, where individuals were found to differ by only 11 genes (Sebat et al. 2004).

These nonsense SNPs are made up of a mixture of potentially deleterious variants present only in a heterozygous state (and thus maintained at low frequency

in the population; 70 SNPs, 41%), and near-neutral or advantageous variants that are found in a homozygous state (and can rise to high frequency). For 99 (59%) nonsense-SNPs, at least one stop homozygous sample was reported, showing that both copies of the nonsense-SNP containing genes can be truncated in our sample donors. However, as we have little phenotypic information on the sample donors, we cannot predict the consequences these nonsense-SNPs are having on their health, except to say that they are compatible with survival to adulthood in a state where the individual is competent to provide informed consent for the use of their sample and is sufficiently interested in helping scientists to provide the sample. Direct insights into their phenotypic consequences could potentially be obtained by detailed studies of individuals of known genotype, by the inclusion of these SNPs in association surveys, or from model organisms.

We attempted to predict some consequences of the nonsense-SNPs *in silico* by using bioinformatic information on the SNP position to predict the likely extent of truncation and the triggering, or not, of NMD. These predictions revealed that many of the nonsense-SNPs analysed will cause a large segment of the protein to be truncated (in at least one transcript), and that 55% can trigger NMD. The consequences could thus often be radical: they could lead to the complete loss of the gene product or possibly to an altered function. We therefore attempted to test the consequences by using available gene expression data. This analysis did not leave us with many significant results to make a generalisation about, but most did meet the prediction of reduced expression in cases where NMD was triggered.

With such a large set of genes to consider (167), it was difficult to study each individual gene in full detail. When we came across an interesting outlier in the genome-wide data, we generally did a literature search for information on that particular gene, but many of these genes had not been studied in enough detail to reveal the functional implications of their loss. A detailed experimental approach would be needed to evaluate the true effects of the nonsense-SNPs and then it might be possible to find out the biological effect of these losses. Future work might thus

include some functional studies. Some of the genes were found in the HGM database and have thus already been implicated in disease. This was not unexpected as nonsense-SNPs are known to be the cause of many diseases. However, in the context of human evolution and our interest in the gene loss theory, we were more interested in those that could have been advantageous for our species.

In an attempt to identify the types of genes where nonsense-SNPs can be found, we performed an analysis of the gene ontology. We found that the genes we studied were mainly overrepresented in terms related to olfactory reception and the nervous system, the latter being rather surprising. Annotation of the human genome is incomplete and so not all of our genes were represented in this analysis. We might thus be missing important categories for the less-annotated genes in our dataset. There is a suggestion that the genes containing nonsense-SNPs are more likely to have paralogs that help back up their function should one be lost. However, a study of the representation of genes in segmental duplications (and thus all with paralogs) reported an overrepresentation of genes associated with immunity and defence, membrane surface interactions, drug detoxification and growth/development (Bailey et al. 2002), none of which were found to be overrepresented in our "lost" genes.

5.2 Selective Forces

We wanted to infer the evolutionary forces acting on nonsense-SNPs, i.e. whether they were evolutionarily advantageous, disadvantageous or neutral. Our measures of derived allele frequencies, population differentiation and long-range haplotypes led us to believe that the SNPs are in the main largely neutral or slightly deleterious We did, however, find interesting outliers, some of which we followed up by resequencing. We reported results for *CASP12* and *MAGEE2*. We intended to follow up *SIGLEC12* as well, but the sequence traces were not good enough to use and we may attempt to redo this in the future. Additionally, *SEMA4C* came up as of potential interest to us because of its specificity for the Americas and may also be followed up by genotyping in a larger samples and resequencing. The resequencing

of the two genes enabled us to use neutrality tests and median-joining networks to investigate the region in greater detail in order to infer the selective forces. We found that *CASP12* gave clear evidence for positive selection in most populations by all neutrality tests used, while *MAGEE2* had an interesting phylogenetic structure and may have been subjected to selection more recently (perhaps starting 20,000 – 40,000 years ago) in Asian populations as suggested by its geographical distribution and the value of Fay and Wu's *H* in the CHB sample. While the *CASP12* results are understandable in view of its role in sepsis resistance, no functional information was available for *MAGEE2* and it would thus be interesting to perform extensive functional studies of this gene in the future. *MAGEE2* perhaps illustrates the situation that is most likely to emerge from genome-wide surveys of this kind: despite reasonable evidence for selection, no clues about the nature of the selective force.

To conclude, we do find some nonsense-SNPs that may be taken to support Olson's less-is-more hypothesis, and thus that gene loss has contributed to human evolution, but do not find evidence that such loss has been a major evolutionary force in human history.

5.3 The Effectiveness of Our Methods

When SNP data are used, one always has to be aware of ascertainment bias in their discovery as allele frequencies and distributions will depend greatly on this. Indeed, many global SNP projects, such as the HapMap, have displayed a deficit of rare and an excess of intermediate frequency SNPs (The International HapMap Consortium 2005). Furthermore, as many of the SNPs used were initially discovered in non-African populations, the HapMap data may be missing out variation within Africa. As we looked at the geographical distribution of our rare stop alleles, we did not find much difference between African and non-African populations. This might be an indication that the expected excess of variation in African populations is not found in our dataset, and thus that African-specific variants are under-represented.

An additional concern related to ascertainment is that tests that depend on allele frequency data, such as population differentiation measures (F_{ST}), should be interpreted with caution. Our genome-wide survey of nonsense-SNPs using genotype data enabled us to pick up signals (in population differentiation and otherwise) that, when followed up by resequencing, were revealed to be of evolutionary interest. So while F_{ST} should not be taken alone as evidence for selection (Xue et al. submitted), it may provide us with useful clues which can then be followed up by more trustworthy methods. Indeed, resequencing data will not be affected by ascertainment bias. Furthermore, as our two SNP classes, nonsense- and synonymous-SNPs, were chosen in the same way, they should also be subject to the same ascertainment bias. Therefore, comparison of the two classes is justifiable as a way to identify nonsense-SNP outliers compared to the assumed neutral synonymous-SNPs, as well as comparing nonsense-SNPs to other nonsense-SNPs to identify those of special interest.

However, while resequencing data are without ascertainmet bias, the neutrality tests are still potentially subject to erroneous conclusions as they rely on population genetic models that make specific (and undoubtedly too simplistic) assumptions about the demography of the populations. In particular, these models often make the assumption that population size is constant and that there is no population structure. Neutrality tests have even been shown to reject neutrality in the absence of selection (reviewed in Nielsen 2005). Indeed most interpretation of genetic diversity is highly sensitive to demographic assumptions. For example, it has been shown that Tajima's *D* (Tajima 1989c) will reject the neutral model in the presence of population growth (Simonsen et al. 1995). Population growth may give a similar effect to a selective sweep. Tests based on patterns of LD may be particularly sensitive, because they rely on assumptions about demography as well as the underlying recombination rates and these can vary greatly between regions (McVean et al. 2004). Thus we are also concerned with distinguishing between the signal given by demographic and selective processes. However, demography will have a similar

affect on the whole genome, whereas selection will have locus-specific effects. Therefore, this problem can be overcome in a number of ways: by modelling demography more realistically (Schaffner et al. 2005) or by the use of empirical comparisons and data from multiple loci as was done with our survey of nonsense-SNPs.

In the end we find that the combination of tests based on genotype (multiple loci) and resequencing (free of ascertainment bias) data currently provides the best way to distinguish a real selective signal from an apparent one based on ascertainment or demography. If accompanied by biological insights into the nature of the phenotype that might be under selection, a convincing case for selection can be made.

5.4 The importance of Knowing one's Nonsense-SNPs

The extent of gene content variation in the healthy population is starting to be appreciated, and it can be seen to be made up of copy number variation (Jakobsson et al. 2008; Redon et al. 2006; Sebat et al. 2004), nonsense-SNPs, other truncating variants such as indels and splice site alterations, and polymorphisms in regulatory elements that ablate expression (Stranger et al. 2007b). Together, these lead to substantial differences in the number of active genes carried by different healthy humans. The number of genes affected in this way is still largely unknown, but this study provides a minimum estimate of the variation due to nonsense-SNPs, and suggests that the total must be a substantial proportion of the entire gene content.

We see that the set of nonsense-SNPs documented in this thesis can be particularly significant for three areas of genetics and medicine. First, sequencing is starting to be used to survey genes or genomes for disease-associated variants, and to inform genetic risk counselling, including whole-genome resequencing for personalized medicine. Nonsense-SNPs discovered in such studies would merit particular attention, but at least the 99 found here in the homozygous state are not associated with mendelian disorders, have no overt influence on the phenotype and are compatible with healthy life. Second, there are nevertheless some situations

where generally-neutral differences in gene content have medical consequences: allogeneic hematopoietic stem cell transplantation, where a donor lacking a gene may mount an immune reaction against the tissues of a recipient with that gene, leading to graft-versus-host disease (Murata et al. 2003). Donors and recipients should be screened for potential gene differences, including those resulting from these nonsense-SNPs. Third, a general treatment for a wide variety of genetic disorders caused by nonsense-SNPs has been proposed: administration of the drug PTC124 which promotes read-through of premature but not normal termination codons (Welch et al. 2007). Such treatment would, if effective, also promote the expression of endogenous non-target genes carrying nonsense-SNPs, and the consequences of this should be evaluated. We need to understand the full extent of human genetic variation in order to reap the full benefits of present and future medicine.

Bibliography

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, and Kruglyak L. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol 2(10):e286.

Akey JM, Zhang G, Zhang K, Jin L, and Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12(12):1805-1814.

Allison AC. 1954. Protection afforded by sickle-cell trait against subtertian malarial infection. Br Med J 1:290-294.

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, and Lander ES. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407(6803):513-516.

Angata T, Varki NM, and Varki A. 2001. A second uniquely human mutation affecting sialic acid biology. J Biol Chem 276(43):40282-40287.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT and others. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25-29.

Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, Meidl P, Searle SM, Stalker J, Storey R, Trevanion S and others. 2005. The Vertebrate Genome Annotation (Vega) database. Nucleic Acids Res 33(Database issue):D459-465.

Asthana S, Schmidt S, and Sunyaev S. 2005. A limited role for balancing selection. Trends Genet 21(1):30-32.

Bagnall RD, Roberts RG, Mirza MM, Torigoe T, Prescott NJ, and Mathew CG. 2008. Novel isoforms of the CARD8 (TUCAN) gene evade a nonsense mutation. Eur J Hum Genet 16(5):619-625.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, and Eichler EE. 2002. Recent segmental duplications in the human genome. Science 297(5583):1003-1007.

Bamshad M, and Wooding SP. 2003. Signatures of natural selection in the human genome. Nat Rev Genet 4(2):99-111.

Bandelt HJ, Forster P, and Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16(1):37-48.

Barbujani G, Magagni A, Minch E, and Cavalli-Sforza LL. 1997. An apportionment of human DNA diversity. Proc Natl Acad Sci U S A 94(9):4516-4519.

Barreiro LB, Laval G, Quach H, Patin E, and Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. Nat Genet 40(3):340-345.

Beeghly-Fadiel A, Long JR, Gao YT, Li C, Qu S, Cai Q, Zheng Y, Ruan ZX, Levy SE, Deming SL and others. 2008. Common MMP-7 polymorphisms and breast cancer susceptibility: a multistage study of association and functionality. Cancer Res 68(15):6453-6459.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, and Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74(6):1111-1120.

Black FL, and Hedrick PW. 1997. Strong balancing selection at HLA loci: evidence from segregation in South Amerindian families. Proc Natl Acad Sci U S A 94(23):12452-12456.

Bloom G, and Sherman P. 2005. Dairying barriers affect the distribution of lactose malabsorption. Evol Hum Behav 26(4):301-312.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, and Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140(2):783-796.
Bubb KL, Bovee D, Buckley D, Haugen E, Kibukawa M, Paddock M, Palmieri A, Subramanian S, Zhou Y, Kaul R and others. 2006. Scan of human genome reveals no new Loci under ancient balancing selection. Genetics 173(4):2165-2177.
Burger J, Kirchner M, Bramanti B, Haak W, and Thomas MG. 2007. Absence of the lactase-persistence-associated allele in early Neolithic Europeans. Proc Natl Acad Sci U S A 104(10):3736-3741.
Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD and others. 2005. Natural selection on protein-coding genes in the human genome. Nature 437(7062):1153-1157.
Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A and others. 2002. A human genome diversity cell line panel. Science 296(5566):261-262.
Cann RL, Stoneking M, and Wilson AC. 1987. Mitochondrial DNA and human evolution. Nature 325(6099):31-36.
Cartegni L, Chew SL, and Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3(4):285-298.
Cavalli-Sforza LL, and Bodmer WFja. 1971. The genetics of human populations. San Francisco: W. H. Freeman.
Cavalli-Sforza LL, and Feldman MW. 2003. The application of molecular genetic approaches to the study of human evolution. Nat Genet 33 Suppl:266-275.
Cavalli-Sforza LL, Menozzi P, and Piazza A. 1994. The history and geography of human genes. Princeton, N.J.; Chichester: Princeton University Press.
Chan S, Seto JT, Macarthur DG, Yang N, North K, and Head S. 2008. A gene for speed: contractile properties of isolated whole EDL muscle from an {alpha}-actinin-3 knockout mouse. Am J Physiol Cell Physiol.
Chance PF, Alderson MK, Leppig KA, Lensch MW, Matsunami N, Smith B, Swanson PD, Odelberg SJ, Disteche CM, and Bird TD. 1993. DNA deletion associated with hereditary neuropathy with liability to pressure palsies. Cell 72(1):143-151.
Charlesworth B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet Res 63(3):213-227.
Charlesworth B, Morgan MT, and Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134(4):1289-1303.
Check E. 2005. Human genome: patchwork people. Nature 437(7062):1084-1086.
Chomez P, De Backer O, Bertrand M, De Plaen E, Boon T, and Lucas S. 2001. An overview of the MAGE gene family with the identification of all human members of the family. Cancer Res 61(14):5544-5551.
Crawford DC, Akey DT, and Nickerson DA. 2005. The patterns of natural variation in human genes. Annu Rev Genomics Hum Genet 6:287-312.
Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, Nickerson DA, and Stephens M. 2004. Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genet 36(7):700-706.
Darwin C. 1859. On the Origin of Species by means of Natural Selection, or the Preservation of favoured races in the struggle for life: John Murray: London.
Dean M, Carrington M, and O'Brien SJ. 2002. Balanced polymorphism selected by genetic versus infectious human disease. Annu Rev Genomics Hum Genet 3:263-292.

Dean M, Carrington M, Winkler C, Huttley GA, Smith MW, Allikmets R, Goedert JJ, Buchbinder SP, Vittinghoff E, Gomperts E and others. 1996. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Science 273(5283):1856-1862.

Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, and Lempicki RA. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4(5):P3.

Di Rienzo A, and Hudson RR. 2005. An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet 21(11):596-601.

Dolphin CT, Beckett DJ, Janmohamed A, Cullingford TE, Smith RL, Shephard EA, and Phillips IR. 1998. The flavin-containing monooxygenase 2 gene (FMO2) of humans, but not of other primates, encodes a truncated, nonfunctional protein. J Biol Chem 273(46):30599-30607.

Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R and others. 2007. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447(7148):1087-1093.

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, and Jarvela I. 2002. Identification of a variant associated with adult-type hypolactasia. Nat Genet 30(2):233-237.

Eswaran V, Harpending H, and Rogers AR. 2005. Genomics refutes an exclusively African origin of humans. J Hum Evol 49(1):1-18.

Fagundes NJ, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL, and Excoffier L. 2007. Statistical evaluation of alternative models of human evolution. Proc Natl Acad Sci U S A 104(45):17614-17619.

Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, Blaser MJ, Graham DY, Vacher S, Perez-Perez GI and others. 2003. Traces of human migrations in Helicobacter pylori populations. Science 299(5612):1582-1585.

Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen M, Steemers F, Butler SL, Deloukas P and others. 2003. Highly parallel SNP genotyping. Cold Spring Harb Symp Quant Biol 68:69-78.

Fay JC, and Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics 155(3):1405-1413.

Feuk L, Carson AR, and Scherer SW. 2006. Structural variation in the human genome. Nat Rev Genet 7(2):85-97.

Fischer A, Wiebe V, Paabo S, and Przeworski M. 2004. Evidence for a complex demographic history of chimpanzees. Mol Biol Evol 21(5):799-808.

Fischer H, Koenig U, Eckhart L, and Tschachler E. 2002. Human caspase 12 has acquired deleterious mutations. Biochem Biophys Res Commun 293(2):722-726.

Fisher SA, Mirza MM, Onnie CM, Soars D, Lewis CM, Prescott NJ, Mathew CG, Sanderson J, Forbes A, Todhunter C and others. 2007. Combined evidence from three large British Association studies rejects TUCAN/CARD8 as an IBD susceptibility gene. Gastroenterology 132(5):2078-2080.

Franke A, Rosenstiel P, Balschun T, Von Kampen O, Schreiber S, Sina C, Hampe J, Karlsen TH, Vatn MH, and Solberg C. 2007. No association between the TUCAN (CARD8) Cys10Stop mutation and inflammatory bowel disease in a large retrospective German and a clinically well-characterized Norwegian sample. Gastroenterology 132(5):2080-2081.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM and others. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164):851-861.

Frischmeyer PA, and Dietz HC. 1999. Nonsense-mediated mRNA decay in health and disease. Hum Mol Genet 8(10):1893-1900.
Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147(2):915-925.
Fu YX, and Li WH. 1993. Statistical tests of neutrality of mutations. Genetics 133(3):693-709.
Gardner M, Williamson S, Casals F, Bosch E, Navarro A, Calafell F, Bertranpetit J, and Comas D. 2007. Extreme individual marker F(ST)values do not imply population-specific selection in humans: the NRG1 example. Hum Genet 121(6):759-762.
Gilad Y, Bustamante CD, Lancet D, and Paabo S. 2003a. Natural selection on the olfactory receptor gene family in humans and chimpanzees. Am J Hum Genet 73(3):489-501.
Gilad Y, Man O, Paabo S, and Lancet D. 2003b. Human specific loss of olfactory receptor genes. Proc Natl Acad Sci U S A 100(6):3324-3327.
Glusman G, Yanai I, Rubin I, and Lancet D. 2001. The complete human olfactory subgenome. Genome Res 11(5):685-702.
Goudet J. 2005. Hierfstat, a package for R to compute and test hierarchical F-statistics. Mol Ecol Notes 5:184-186.
Hamblin MT, and Di Rienzo A. 2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66(5):1669-1679.
Hamblin MT, Thompson EE, and Di Rienzo A. 2002. Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet 70(2):369-383.
Han A, Kim WY, and Park SM. 2007. SNP2NMD: A database of human single nucleotide polymorphisms causing nonsense-mediated mRNA decay. Bioinformatics 23(3):397-399.
Harpending H, and Rogers A. 2000. Genetic perspectives on human origins and differentiation. Annu Rev Genomics Hum Genet 1:361-385.
Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, and Sherry ST. 1998. Genetic traces of ancient demography. Proc Natl Acad Sci U S A 95(4):1961-1967.
Hayakawa T, Angata T, Lewis AL, Mikkelsen TS, Varki NM, and Varki A. 2005. A human-specific gene in microglia. Science 309(5741):1693.
Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, Jonasdottir A, Sigurdsson A, Baker A, Palsson A and others. 2007. A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science 316(5830):1491-1493.
Helgason A, Palsson S, Thorleifsson G, Grant SF, Emilsson V, Gunnarsdottir S, Adeyemo A, Chen Y, Chen G, Reynisdottir I and others. 2007. Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. Nat Genet 39(2):218-225.
Hellenthal G, and Stephens M. 2006. Insights into recombination from population genetic variation. Curr Opin Genet Dev 16(6):565-572.
Henshilwood CS, d'Errico F, Yates R, Jacobs Z, Tribolo C, Duller GA, Mercier N, Sealy JC, Valladas H, Watts I and others. 2002. Emergence of modern human behavior: Middle Stone Age engravings from South Africa. Science 295(5558):1278-1280.
Hentze MW, and Kulozik AE. 1999. A perfect message: RNA surveillance and nonsense-mediated decay. Cell 96(3):307-310.
Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, Saha N, Kozlov AI, and Swallow DM. 2001. Lactase haplotype diversity in the Old World. Am J Hum Genet 68(1):160-172.
Hosack DA, Dennis G, Jr., Sherman BT, Lane HC, and Lempicki RA. 2003. Identifying biological themes within lists of genes with EASE. Genome Biol 4(10):R70.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18(2):337-338.
Hudson RR, and Kaplan NL. 1988. The coalescent process in models with selection and recombination. Genetics 120(3):831-840.
Hudson RR, and Kaplan NL. 1995. Deleterious background selection with recombination. Genetics 141(4):1605-1617.
Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, and Yeager M. 2003. Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc Natl Acad Sci U S A 100(26):15754-15757.
Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, and Lee C. 2004. Detection of large-scale variation in the human genome. Nat Genet 36(9):949-951.
Inacio A, Silva AL, Pinto J, Ji X, Morgado A, Almeida F, Faustino P, Lavinha J, Liebhaber SA, and Romao L. 2004. Nonsense mutations in close proximity to the initiation codon fail to trigger full nonsense-mediated mRNA decay. J Biol Chem 279(31):32170-32180.
Ingram CJ, Elamin MF, Mulcare CA, Weale ME, Tarekegn A, Raga TO, Bekele E, Elamin FM, Thomas MG, Bradman N and others. 2007. A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence? Hum Genet 120(6):779-788.
Inoue K, and Lupski JR. 2002. Molecular mechanisms for genomic disorders. Annu Rev Genomics Hum Genet 3:199-242.
Isken O, and Maquat LE. 2007. Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. Genes Dev 21(15):1833-1856.
Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R and others. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451(7181):998-1003.
Jobling MA, Hurles ME, and Tyler-Smith C. 2003. Human Evolutionary Genetics: Origins, Peoples & Disease. London/New York: Garland Science Publishing.
Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D. 2002. The human genome browser at UCSC. Genome Res 12(6):996-1006.
Kimura M. 1968. Evolutionary rate at the molecular level. Nature 217(129):624-626.
Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.
King PH, Waldrop R, Lupski JR, and Shaffer LG. 1998. Charcot-Marie-Tooth phenotype produced by a duplicated PMP22 gene as part of a 17p trisomy-translocation to the X chromosome. Clin Genet 54(5):413-416.
Krueger SK, Siddens LK, Martin SR, Yu Z, Pereira CB, Cabacungan ET, Hines RN, Ardlie KG, Raucy JL, and Williams DE. 2004. Differences in FMO2*1 allelic frequency between Hispanics of Puerto Rican and Mexican descent. Drug Metab Dispos 32(12):1337-1340.
Kruglyak L, and Nickerson DA. 2001. Variation is the spice of life. Nat Genet 27(3):234-236.
Kryukov GV, Pennacchio LA, and Sunyaev SR. 2007. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet 80(4):727-739.
Kuhn K, Baker SC, Chudin E, Lieu MH, Oeser S, Bennett H, Rigault P, Barker D, McDaniel TK, and Chee MS. 2004. A novel, high-performance random array platform for quantitative gene expression profiling. Genome Res 14(11):2347-2356.

Kwiatkowski DP. 2005. How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet 77(2):171-192.
Lahr MM, and Foley RA. 1994. Multiple Dispersals and Modern Human Origins. Evol Anthropol 3(2):48-60.
Lahr MM, and Foley RA. 1998. Towards a theory of modern human origins: Geography, Demography and Diversity in Recent Human Evolution. Yearb Phys Anthropol 41:137-176.
Lao O, de Gruijter JM, van Duijn K, Navarro A, and Kayser M. 2007. Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. Ann Hum Genet 71(Pt 3):354-369.
Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D and others. 2008. Correlation between genetic and geographic structure in Europe. Curr Biol 18(16):1241-1248.
Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G and others. 2007. The diploid genome sequence of an individual human. PLoS Biol 5(10):e254.
Lewis BP, Green RE, and Brenner SE. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A 100(1):189-192.
Liu H, Prugnolle F, Manica A, and Balloux F. 2006. A geographically explicit genetic model of worldwide human-settlement history. Am J Hum Genet 79(2):230-237.
Livingstone FB. 1984. The Duffy blood groups, vivax malaria, and malaria selection in human populations: a review. Hum Biol 56(3):413-425.
Lupski JR, de Oca-Luna RM, Slaugenhaupt S, Pentao L, Guzzetta V, Trask BJ, Saucedo-Cardenas O, Barker DF, Killian JM, Garcia CA and others. 1991. DNA duplication associated with Charcot-Marie-Tooth disease type 1A. Cell 66(2):219-232.
MacArthur DG, and North KN. 2004. A gene for speed? The evolution and function of alpha-actinin-3. Bioessays 26(7):786-795.
MacArthur DG, Seto JT, Raftery JM, Quinlan KG, Huttley GA, Hook JW, Lemckert FA, Kee AJ, Edwards MR, Berman Y and others. 2007. Loss of ACTN3 gene function alters mouse muscle metabolism and shows evidence of positive selection in humans. Nat Genet 39(10):1261-1265.
Maquat LE. 2004. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol 5(2):89-99.
Marth GT, Czabarka E, Murvai J, and Sherry ST. 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics 166(1):351-372.
McDougall I, Brown FH, and Fleagle JG. 2005. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature 433(7027):733-736.
McGovern DP, Butler H, Ahmad T, Paolucci M, van Heel DA, Negoro K, Hysi P, Ragoussis J, Travis SP, Cardon LR and others. 2006. TUCAN (CARD8) genetic variants and inflammatory bowel disease. Gastroenterology 131(4):1190-1196.
McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, and Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. Science 304(5670):581-584.
Mellars P. 2006. Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia. Science 313(5788):796-800.
Meyre D, Bouatia-Naji N, Tounian A, Samson C, Lecoeur C, Vatin V, Ghoussaini M, Wachter C, Hercberg S, Charpentier G and others. 2005. Variants of ENPP1 are

associated with childhood and adult obesity and increase the risk of glucose intolerance and type 2 diabetes. Nat Genet 37(8):863-867.
Morisaki T, Gross M, Morisaki H, Pongratz D, Zollner N, and Holmes EW. 1992. Molecular basis of AMP deaminase deficiency in skeletal muscle. Proc Natl Acad Sci U S A 89(14):6457-6461.
Mort M, Ivanov D, Cooper DN, and Chuzhanova NA. 2008. A meta-analysis of nonsense mutations causing human genetic disease. Hum Mut.
Mulcare CA, Weale ME, Jones AL, Connell B, Zeitlyn D, Tarekegn A, Swallow DM, Bradman N, and Thomas MG. 2004. The T allele of a single-nucleotide polymorphism 13.9 kb upstream of the lactase gene (LCT) (C-13.9kbT) does not predict or cause the lactase-persistence phenotype in Africans. Am J Hum Genet 74(6):1102-1110.
Murata M, Warren EH, and Riddell SR. 2003. A human minor histocompatibility antigen resulting from differential expression due to a gene deletion. J Exp Med 197(10):1279-1289.
Myers S, Bottolo L, Freeman C, McVean G, and Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310(5746):321-324.
Nagy E, and Maquat LE. 1998. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends Biochem Sci 23(6):198-199.
Neel JV. 1962. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am J Hum Genet 14:353-362.
Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.
Nei M, and Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3(5):418-426.
Nielsen R. 2005. Molecular signatures of natural selection. Annu Rev Genet 39:197-218.
Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ and others. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3(6):e170.
Nielsen R, Hubisz MJ, and Clark AG. 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics 168(4):2373-2382.
Norman B, Nygren AT, Nowak J, and Sabina RL. 2008. The effect of AMPD1 genotype on blood flow response to sprint exercise. Eur J Appl Physiol 103(2):173-180.
Norton HL, Kittles RA, Parra E, McKeigue P, Mao X, Cheng K, Canfield VA, Bradley DG, McEvoy B, and Shriver MD. 2007. Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. Mol Biol Evol 24(3):710-722.
Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR and others. 2008. Genes mirror geography within Europe. Nature 456(7218):98-101.
Ohashi J, Naka I, Kimura R, Natsuhara K, Yamauchi T, Furusawa T, Nakazawa M, Ataka Y, Patarapotikul J, Nuchnoi P and others. 2007. FTO polymorphisms in oceanic populations. J Hum Genet 52(12):1031-1035.
Ohno S. 1970. Evolution by gene duplication. London: Allen & Unwin.
Olson MV. 1999. When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet 64(1):18-23.
Olson MV, and Varki A. 2003. Sequencing the chimpanzee genome: insights into human evolution and disease. Nat Rev Genet 4(1):20-28.

Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y and others. 2002. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 32(4):650-654.
Perry GH, Verrelli BC, and Stone AC. 2005. Comparative analyses reveal a complex history of molecular evolution for human MYH16. Mol Biol Evol 22(3):379-382.
Plagnol V, and Wall JD. 2006. Possible ancestral structure in human populations. PLoS Genet 2(7):e105.
Pritchard JK, and Przeworski M. 2001. Linkage disequilibrium in humans: models and data. Am J Hum Genet 69(1):1-14.
Prugnolle F, Manica A, and Balloux F. 2005. Geography predicts neutral genetic diversity of human populations. Curr Biol 15(5):R159-160.
Przeworski M. 2002. The signature of positive selection at randomly chosen loci. Genetics 160(3):1179-1189.
Przeworski M, Hudson RR, and Di Rienzo A. 2000. Adjusting the focus on human variation. Trends Genet 16(7):296-302.
Ray N, Currat M, Berthier P, and Excoffier L. 2005. Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. Genome Res 15(8):1161-1167.
Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W and others. 2006. Global variation in copy number in the human genome. Nature 444(7118):444-454.
Ronald J, and Akey JM. 2005. Genome-wide scans for loci under selection in humans. Hum Genomics 2(2):113-125.
Rosenberg NA. 2006. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet 70(Pt 6):841-847.
Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, and Feldman MW. 2002. Genetic structure of human populations. Science 298(5602):2381-2385.
Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M, Howell GR, Burrows C, Bird CP and others. 2005. The DNA sequence of the human X chromosome. Nature 434(7031):325-337.
Rozas J, Sanchez-DelBarrio JC, Messeguer X, and Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19(18):2496-2497.
Rozen S, and Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365-386.
Rubattu S, Stanzione R, Di Angelantonio E, Zanda B, Evangelista A, Tarasi D, Gigante B, Pirisi A, Brunetti E, and Volpe M. 2004. Atrial natriuretic peptide gene polymorphisms and risk of ischemic stroke in humans. Stroke 35(4):814-818.
Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ and others. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419(6909):832-837.
Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, and Lander ES. 2006. Positive natural selection in the human lineage. Science 312(5780):1614-1620.
Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R and others. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449(7164):913-918.

Sabeti PC, Walsh E, Schaffner SF, Varilly P, Fry B, Hutcheson HB, Cullen M, Mikkelsen TS, Roy J, Patterson N and others. 2005. The case for selection at CCR5-Delta32. PLoS Biol 3(11):e378.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL and others. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409(6822):928-933.

Saleh M, Vaillancourt JP, Graham RK, Huyck M, Srinivasula SM, Alnemri ES, Steinberg MH, Nolan V, Baldwin CT, Hotchkiss RS and others. 2004. Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. Nature 429(6987):75-79.

Sandhu MS, Weedon MN, Fawcett KA, Wasson J, Debenham SL, Daly A, Lango H, Frayling TM, Neumann RJ, Sherva R and others. 2007. Common variants in WFS1 confer risk of type 2 diabetes. Nat Genet 39(8):951-953.

Savas S, Tuzmen S, and Ozcelik H. 2006. Human SNPs resulting in premature stop codons and protein truncation. Hum Genomics 2(5):274–286.

Sawyer SL, Berglind LC, and Brookes AJ. 2003. Negligible validation rate for public domain stop-codon SNPs. Hum Mutat 22(3):252-254.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, and Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15(11):1576-1583.

Schneider JA, Pungliya MS, Choi JY, Jiang R, Sun XJ, Salisbury BA, and Stephens JC. 2003. DNA variability of human genes. Mech Ageing Dev 124(1):17-25.

Schneider S, Roessli D, and Excoffier L. 2000. Arlequin ver. 2.000: A software for population genetics data analysis. Version 2.000: Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M and others. 2004. Large-scale copy number polymorphism in the human genome. Science 305(5683):525-528.

Simonsen KL, Churchill GA, and Aquadro CF. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141(1):413-429.

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S and others. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445(7130):881-885.

Slatkin M, and Hudson RR. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129(2):555-562.

Smith JM, and Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res 23(1):23-35.

Solberg OD, Mack SJ, Lancaster AK, Single RM, Tsai Y, Sanchez-Mazas A, and Thomson G. 2008. Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. Hum Immunol 69(7):443-464.

Speakman JR. 2006. Thrifty genes for obesity and the metabolic syndrome--time to call off the search? Diab Vasc Dis Res 3(1):7-11.

Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, Masson G, Jakobsdottir M, Thorlacius S, Helgason A and others. 2007. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 39(7):865-869.

Stedman HH, Kozyak BW, Nelson A, Thesier DM, Su LT, Low DW, Bridges CR, Shrager JB, Minugh-Purvis N, and Mitchell MA. 2004. Myosin gene mutation correlates with anatomical changes in the human lineage. Nature 428(6981):415-418.
Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG and others. 2005. A common inversion under selection in Europeans. Nat Genet 37(2):129-137.
Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, and Cooper DN. 2003. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21(6):577-581.
Stephens M, and Donnelly P. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73(5):1162-1169.
Stephens M, Smith NJ, and Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68(4):978-989.
Strachan T, and Read AP. 2004. Human molecular genetics 3. London ; New York: Garland Press.
Stranger BE, and Dermitzakis ET. 2006. From DNA to RNA to disease and back: the 'central dogma' of regulatory disease variation. Hum Genomics 2(6):383-390.
Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavare S and others. 2005. Genome-wide associations of gene expression variation in humans. PLoS Genet 1(6):e78.
Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C and others. 2007a. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315(5813):848-853.
Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D and others. 2007b. Population genomics of human gene expression. Nat Genet 39(10):1217-1224.
Swallow DM. 2003. Genetics of lactase persistence and lactose intolerance. Annu Rev Genet 37:197-219.
Tajima F. 1989a. DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-subpopulation model. Genetics 123(1):229-240.
Tajima F. 1989b. The effect of change in population size on DNA polymorphism. Genetics 123(3):597-601.
Tajima F. 1989c. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585-595.
Takahata N, Satta Y, and Klein J. 1995. Divergence time and population size in the lineage leading to modern humans. Theor Popul Biol 48(2):198-221.
Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, and Visscher PM. 2007. Recent human effective population size estimated from linkage disequilibrium. Genome Res 17(4):520-526.
The International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437(7063):1299-1320.
The International Human Genome Mapping Consortium. 2001. A physical map of the human genome. Nature 409(6822):934-941.
The International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature 431(7011):931-945.
Tishkoff SA, and Kidd KK. 2004. Implications of biogeography of human populations for 'race' and medicine. Nat Genet 36(11 Suppl):S21-27.
Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M and others. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39(1):31-40.

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J and others. 2001. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293(5529):455-462.

Usdin K. 2008. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. Genome Res 18(7):1011-1019.

Vallender EJ, and Lahn BT. 2004. Positive selection on the human genome. Hum Mol Genet 13 Spec No 2:R245-R254.

Veeramah KR, Thomas MG, Weale ME, Zeitlyn D, Tarekegn A, Bekele E, Mendell NR, Shephard EA, Bradman N, and Phillips IR. 2008. The potentially deleterious functional variant flavin-containing monooxygenase 2*1 is at high frequency throughout sub-Saharan Africa. Pharmacogenet Genomics 18(10):877-886.

Voight BF, Kudaravalli S, Wen X, and Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol 4(3):e72.

Walsh EC, Sabeti P, Hutcheson HB, Fry B, Schaffner SF, de Bakker PI, Varilly P, Palma AA, Roy J, Cooper R and others. 2006. Searching for signals of evolutionary selection in 168 genes related to immune function. Hum Genet 119(1-2):92-102.

Wang X, Grus WE, and Zhang J. 2006. Gene losses during human origins. PLoS Biol 4(3):e52.

Weir BS, Cardon LR, Anderson AD, Nielsen DM, and Hill WG. 2005. Measures of human population structure show heterogeneity among genomic regions. Genome Res 15(11):1468-1476.

Weir BS, and Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370.

Welch EM, Barton ER, Zhuo J, Tomizawa Y, Friesen WJ, Trifillis P, Paushkin S, Patel M, Trotta CR, Hwang S and others. 2007. PTC124 targets genetic disorders caused by nonsense mutations. Nature 447(7140):87-91.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT and others. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452(7189):872-876.

Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, and Bustamante CD. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci U S A 102(22):7882-7887.

Wolfe ND, Dunavan CP, and Diamond J. 2007. Origins of major human infectious diseases. Nature 447(7142):279-283.

Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE and others. 2007. A comprehensive analysis of common copy-number variations in the human genome. Am J Hum Genet 80(1):91-104.

Wright S. 1931. Evolution in Mendelian Populations. Genetics 16(2):97-159.

Xue Y, Daly A, Yngvadottir B, Liu M, Coop G, Kim Y, Sabeti P, Chen Y, Stalker J, Huckle E and others. 2006. Spread of an inactive form of caspase-12 in humans is due to recent positive selection. Am J Hum Genet 78(4):659-670.

Xue Y, Zhang X, Huang N, Daly A, Gillson CJ, Yngvadottir B, Nica AC, Woodwark C, Chen Y, Ayub Q and others. submitted. Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation.

Yang N, MacArthur DG, Gulbin JP, Hahn AG, Beggs AH, Easteal S, and North K. 2003. ACTN3 genotype is associated with human elite athletic performance. Am J Hum Genet 73(3):627-631.

Yngvadottir B, and Carvalho-Silva DR. 2008. Reconstructing Human History Using Autosomal, Y-Chromosomal and Mitochondrial Markers. In: (ELS) EoLS, editor. Encyclopedia of Life Sciences (ELS). Chichester: John Wiley & Sons, Ltd.

Yu N, Jensen-Seaman MI, Chemnick L, Kidd JR, Deinard AS, Ryder O, Kidd KK, and Li WH. 2003. Low nucleotide diversity in chimpanzees and bonobos. Genetics 164(4):1511-1518.

Zhang J, and Maquat LE. 1997. Evidence that translation reinitiation abrogates nonsense-mediated mRNA decay in mammalian cells. EMBO J 16(4):826-833.

Zhao Z, Jin L, Fu YX, Ramsay M, Jenkins T, Leskinen E, Pamilo P, Trexler M, Patthy L, Jorde LB and others. 2000. Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc Natl Acad Sci USA 97(21):11354-11358.

Zhu J, Sanborn JZ, Diekhans M, Lowe CB, Pringle TH, and Haussler D. 2007. Comparative genomics search for losses of long-established genes on the human lineage. PLoS Comput Biol 3(12):e247.

Appendix A

A list of all sample names used in genotyping (A.1) and resequencing (A.2) is available on the accompanying CD.

Appendix B

B.1. The Illumina primers used for genotyping are on accompanying CD.

B.2. Long-Range PCR primers for *MAGEE2* and *SIGLEC12*

Primers	Primer Sequence	Product Size
Magee2_Hs_MS1_F	CAATGGCTATTCTTGTGTCTTCC	
Magee2_Hs_MS1_R	CCATCAGAGCACCTATTTTCATC	7046
Magee2_Hs_MS2_F	CCTCAATCAGAGACAACCCATAG	
Magee2_Hs_MS2_R	GTGAACACCCACATCCTTTACAT	7053
Siglec12_Hs_MS1_F	TGTGTTCAAAGTCATTGATGGAG	
Siglec12_Hs_MS1_R	CACCCTTTTAATTTTGTGCTCTG	7019
Siglec12_Hs_MS2_F	AATGCAAATCAAAGCCTTAATGA	
Siglec12_Hs_MS2_R	CAGAACAAGGATTGTACCCTGAG	7059

B.3. Nested PCR primers for *MAGEE2* and *SIGLEC12*

STS_name	forward_primer	reverse_primer	STS_size
Siglec12_1	TTCTGTCCATAATGATCCTGCTT	TACACGTATGCTGTGTGTCTCCT	507
Siglec12_2	TGTCTCTCATTCAGTTTCTCGGT	TGAGGTTTGGACCAGTGTTAGTT	503
Siglec12_3	ACAGGAGTGGCATTTCCTAGAAT	GTGGAAGATGAGGTGATTCAGAT	484
Siglec12_4	AACTAACACTGGTCCAAACCTCA	CAGCCAAGCAAGACAGATACTTT	486
Siglec12_5	ATCTGAATCACCTCATCTTCCAC	CAAGTCCAAGCCTAAGTCTTCCT	575
Siglec12_6	AAAGTATCTGTCTTGCTTGGCTG	CTTCTGGTCTTGTCCTCAGGTCT	469
Siglec12_7	TATGGAGATGTTGAGGACTCTCG	GTGAAGAAATCTGAGAAGCTGGA	583
Siglec12_8	AGGGTCTCTTTGGAGGGTACTG	AAATATTTGAGCTGACGGATGTG	496
Siglec12_9	GATGCTGTAAACTTATGGCCAAG	TCCTTTAGAATTAACAGCCCTCC	578
Siglec12_10	CAACACGGATGAACCTAGAGAAC	GAGGAAGGAGAGATCCAGTATGC	458
Siglec12_11	TAAGTTCTGATGTCCTGTTGCAC	CGCAGGGTACTTAACCATTTCTA	588
Siglec12_12	GAATGGAGGAAGAGAAGGGAGT	TATTGCCACCTCCTTCAGTTAAA	461
Siglec12_13	CATTAACTTCCAACAACAATTCCA	AGGTAACAGATCAACAGGTTCCA	519
Siglec12_14	TAACTGAAGGAGGTGGCAATAAA	TAAAGTTCTGAAGGTCACAAGCC	596
Siglec12_15	AGGAGGAGGGCATAGATAGATTG	AGATGTGGTGTGTGTGAATGT	458
Siglec12_16	GGCTTGTGACCTTCAGAACTTTA	GGAAGTCGTCAGTTTATTGATGG	581
Siglec12_17	GTCTTGCAGTCTTCTCCTTCTTG	GAATTGCCAATAGATTGGGTGTA	529
Siglec12_18	GCCATCAATAAACTGACGACTTC	AGAAGTTGAGCCTGTGTGTGAAT	572
Siglec12_19	TACACCCAATCTATTGGCAATTC	TTAGTGATGTTTGAGCACAGGAA	491
Siglec12_20	TTAATCATGGTCTCTTGCACAAA	TCATTAAGGCTTTGATTTGCATT	513
Siglec12_21	GGAGTGCTCTTCAATCTCACAGT	TGGGTGTATATCCACAAGTAGGG	531
Siglec12_22	AATGCAAATCAAAGCCTTAATGA	CCACATTGTAAGGTGTGACAAGA	445
Siglec12_23	TGTGACATTTGTGGGAATGTAAA	ACATTCGTTGAACAATAACTCCG	514
Siglec12_24	ATTCACAATAACCAAGAGGTGGA	CCATCCATCCAGCCATCTAC	455
Siglec12_25	GGGTGAACCTTGAGGATATTTGT	GTCTGTCCATCCATCCATCC	538
Siglec12_26	GTAGATGGCTGGATGGATGG	GATATGGCCAAACCAATCAACTA	553

STS_name	forward_primer	reverse_primer	STS_size
Siglec12_27	GGATGGATGGATGGACTGAC	CAGTGTCATCACCAACAAGGTTA	565
Siglec12_28	TGAATTAATGAATGAAGGGATGG	TCTCTCGAGTAGCTGGGATTACA	592
Siglec12_29	CAAATCTCAGCAATGAAGATGAA	TCAGCAGAGCACTGTCACTAATC	446
Siglec12_30	GCTCGTGTATGTAATCCCAGCTA	GGGATTACTTCCTCACTGTCCTT	485
Siglec12_31	AACCCAGATTAGTGACAGTGCTC	ACTGAAAGGCTCTCTGGTCTCTT	481
Siglec12_32	GAAAGGAAGGACAGTGAGGAAGT	GGTCTTCCTGTACTTCTGCATCA	591
Siglec12_33	GTCTGAGGTTTGCCACAGACAT	TTTAAATAAAGGCAGACTGCACC	482
Siglec12_34	TGATGCAGAAGTACAGGAAGACC	CTGGACTAGTGTGAGGCAAGTG	442
Siglec12_35	ACAGAAAGAACAAGGACGGAAG	AACGAGTACACAGGTGGGTAGG	476
Siglec12_36	GTTCTTAAATTGTGTGCCCAAAG	ACTTTGTCTCTCTCTGCCCTTCT	580
Siglec12_37	CCTACCCACCTGTGTACTCGTT	TTATCCTACAGCACAAACACTGC	593
Siglec12_38	GGTTGTGGATGCTGTAGAGAAAG	GATCTCAGAGGTGGTTTGATGTC	453
Siglec12_39	ATATGCACGTTCACTGCTCACT	AGCTCAAGGATTTGAGGACTCTT	529
Siglec12_40	GAGGGTGCAAACAGTCTCACTAC	GAACTTGACCATGACTGTCTTCC	525
Siglec12_41	ATAGGTGTGTGTGTGGAGGAGTT	CAACATATCCTGTGAGTTCTGGG	477
Siglec12_42	AGGAGGATCTGGAACAGAAAGAC	GAGGAGTAAGACCAGAGCCTGAG	594
Siglec12_43	ATCGAGGAGCGAGTGATAGTG	ACTACTTCCAGGTGGAGAGAGGA	444
Siglec12_44	AGGCTCAGGCTCTGGTCTTACT	TCCCAGGACCTACTGTCAAGATA	521
Siglec12_45	CCTCTCTCCACCTGGAAGTAGTA	TGAGTTCCTATAGCATGTGGGTT	560
Siglec12_46	CCAGCCTGTATCTTGACAGTAGG	GTCCTCGGAGATCCACATTTAG	474
Siglec12_47	AAGAGAAACTGCAAACCCACAT	TTATCTCCACACATTGTCAAACG	444
Siglec12_48	CTAAATGTGGATCTCCGAGGAC	CAAATGTGATGAGGGCTTTAGG	556
MAGEE2_1	ATTTCAAAGAAATGGCTCTCAGG	TTATCATTCCAGGTGGTAAAGGA	555
MAGEE2_2	AAACTCTAAGGATGTCCATGGCT	GGAGGATAGAGAAAGGGAAGACA	432
MAGEE2_3	ATCTCAGGGTGATCTCCTTTACC	AAGCAGACCTCAGTGTCCTACAG	482
MAGEE2_4	CAATGGCTATTCTTGTGTCTTCC	GATTTCAGAACTGTGACCCAAAC	561
MAGEE2_5	ATTCTTGATGCACCTAAACCTCA	ATCTTTAGGAATTCTCTGCAGGC	516
MAGEE2_6	CAGCATTAAGTGGCCTAGAAAGA	TTTCAGGGATGAGTGGTTAGAAA	572
MAGEE2_7	AACCCTGTTCAACGAGAAATCTT	CATCAATGTCCACCAAGTATTGA	486
MAGEE2_8	TTTCTAACCACTCATCCCTGAAA	GAGTACTCCCATCCTTTAGGCAC	440
MAGEE2_9	TTCCATCTTCCAAATGAGTTCAC	GTATGCATTTCTGTCCAGTCTCC	562
MAGEE2_10	GGCTATTTCCCACTTCTGTTCTT	TGAGGCTTAACTAATATGCCCAA	490
MAGEE2_11	GGAGACTGGACAGAAATGCATAC	CTGGACAACGATGATCAATGTAA	430
MAGEE2_12	TGCAGCCTTTCTTTGTATATTCC	TGTGCTAAACAGCAATTCTCTCA	512
MAGEE2_13	ACAGTGCCTGACGCATATTTAGT	TACAGCCCAGTCTCAGTGATCTT	483
MAGEE2_14	TGAGAGAATTGCTGTTTAGCACA	CACTTATTACAGAGGGATGGCTG	562
MAGEE2_15	GAAACGATTGTGTCGTTCTCTTC	CTTTAAAGGAGGAGTGGGAGAAA	540
MAGEE2_16	GCAAACACTCACCTTAACAGCTT	GGCATAATAGGCATAAGCTCAGA	520
MAGEE2_17	GTCACTAAAGCACATCTTGTCCC	GATTGCTCTCCCTTTGTGTCTAC	502
MAGEE2_18	ATTGCTTGTTTACTTCCATGAGC	AATTAGAATCCAGGTAGGGCTTG	565
MAGEE2_19	TTTGTTGTGTGTGTGTCATTGTT	CAGAACAATATAGGGAGGCTGTG	437
MAGEE2_20	CACAATTTGTACTTCCCACACAA	AATCAGAGGAAGAGCAAGTGATG	539
MAGEE2_21	CTGGATTAGAGTAGGACAGTGGC	TTGGTCCAGTTGGCTATTAGTGT	430

STS_name	forward_primer	reverse_primer	STS_size
MAGEE2_22	CATCACTTGCTCTTCCTCTGATT	GAACACAAGGAACCTCCTCACTA	517
MAGEE2_23	AGTCAGGCTTCTTTCTTTGGTTT	GTTATTGATCCTCAGGCTGACAC	465
MAGEE2_24	CAGTAGTGAGGAGGTTCCTTGTG	GACCCTCTAGAAGACAGGTCGAT	430
MAGEE2_25	GTGTCAGCCTGAGGATCAATAAC	CATGTCTCTGGTAAGCCAGAATG	459
MAGEE2_26	AATCGACCTGTCTTCTAGAGGGT	ATGAGAGACAAAGCATGGAAGAC	559
MAGEE2_27	AGAGACATGGTTCCAGGAGACAG	AATAATTCTAATCGTTAGCGCCC	566
MAGEE2_28	GAGAAGAGAATTCGAGAAATGGA	ATGACCCAGTTTCCTCATCTGTA	511
MAGEE2_29	AAAGCTTCCATTCACTCAACAGA	ACTCCTTCTTGAGGCTTTATTGG	556
MAGEE2_30	TCCAAAGTCAAGCAGCAATAAAT	TGTCTAGCAACCTTCATCCTCAT	428
MAGEE2_31	CCAATAAAGCCTCAAGAAGGAGT	TGCAACCATCATAGTCTTCAACTT	484
MAGEE2_32	AGTATGTCAAAGCTTGTTGTGGC	TTCTACCGTCTAGAACATGCACA	573
MAGEE2_33	TTCAGTAATCTTATGCCCTGGAA	GGAGGGAGTAGTAGGAGGTTCAA	495
MAGEE2_34	GTGTGGTTTGTATGCTGATCCTT	TCAATATGGCCAGGAAGAAGATA	467
MAGEE2_35	TTGAACCTCCTACTACTCCCTCC	TGCTGTAATGGTCTGTTTCTTGA	549
MAGEE2_36	TCATGCTCTCCATTCTGAACTTT	CAGTTGACAGGGTAAGTTGTGGT	495
MAGEE2_37	CAATCTCTTCCCACTCATTTGTC	GTGTGGACAAGAATACAGAGCAA	509
MAGEE2_38	CCACAACTTACCCTGTCAACTGT	ATTGTTTCCATCTCCTTCCATTT	520
MAGEE2_39	TTGCTCTGTATTCTTGTCCACAC	AATGTTCACTTGAGCACCAATCT	513
MAGEE2_40	AAATGGAAGGAGATGGAAACAAT	AAACAAATATCTCGGGTGGGT	535
MAGEE2_41	AGATTGGTGCTCAAGTGAACATT	GACAGAGGGTCTTCAAATGAGTG	511
MAGEE2_42	GAATATCCATGATTCCACCCAC	AAGCTATAAGGCAAACAATCGAA	448
MAGEE2_43	CATATTTGTATGACAAGCAGCCA	GACATGGAGAAACCAGAACACTC	559
MAGEE2_44	CATTCGATTGTTTGCCTTATAGC	ATCCAAGTGGCTGATAAACACAT	441
MAGEE2_45	TTTGAAGAATTCCCAGAAGGAGT	GAAGGCACAGAAATACCTCATTG	540
MAGEE2_46	TTTGTCAAATATATGTACTGCAAATG	AAACAACCTGAGAAATGGAACAA	496
MAGEE2_47	GCCTTCTTCAAATACGATTTATTG	CCAAGGCTTCTCAAGTATGAATG	465
MAGEE2_48	AGCTTTCATCATGGCATTGTAGT	CCCTTTCACATCCACTTACTGAG	523

Appendix C

All scripts used are available on the accompanying CD.

Perl:

createfstatinput

hgdp2sweep

merge_sts

pcroverlap

phase2fasta

phase2network

snptab2phase

Java:

DelimitedFileTransformer

InputFileTransformer

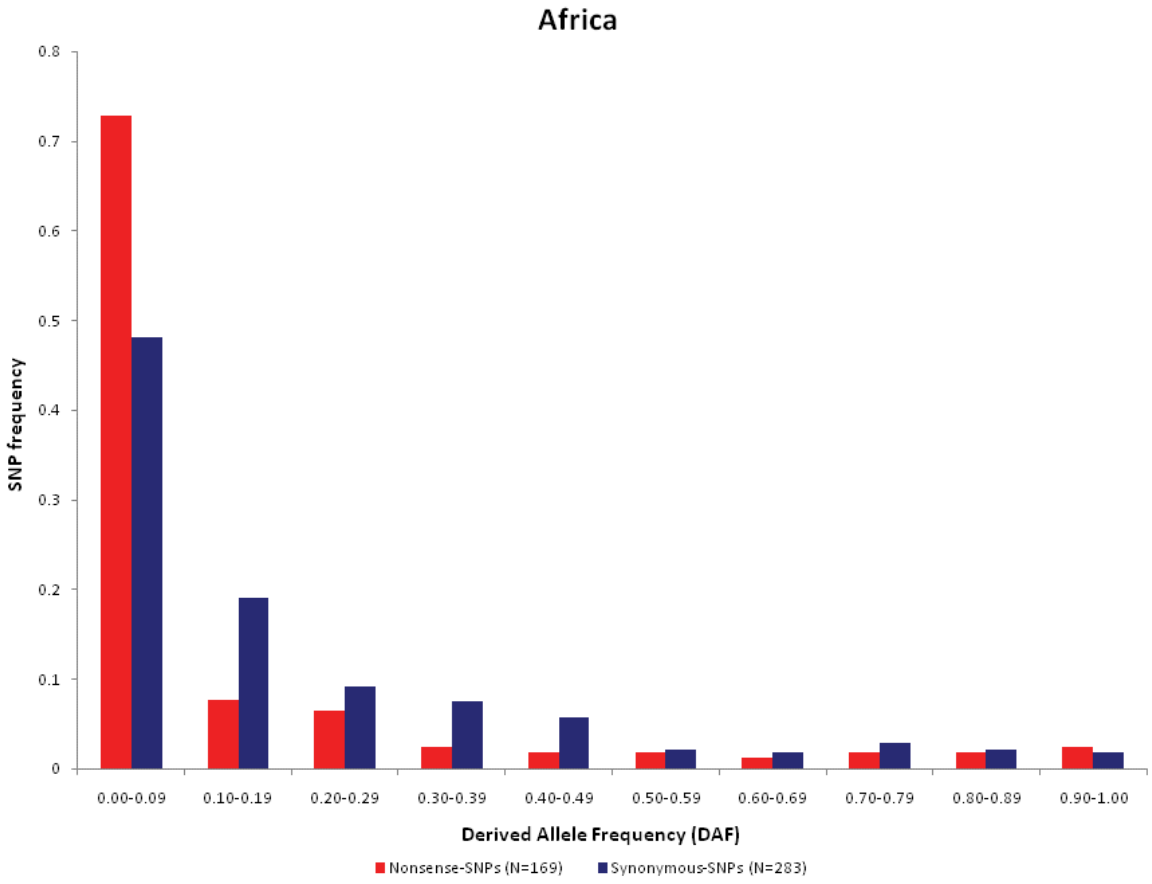
SweepFileConversion

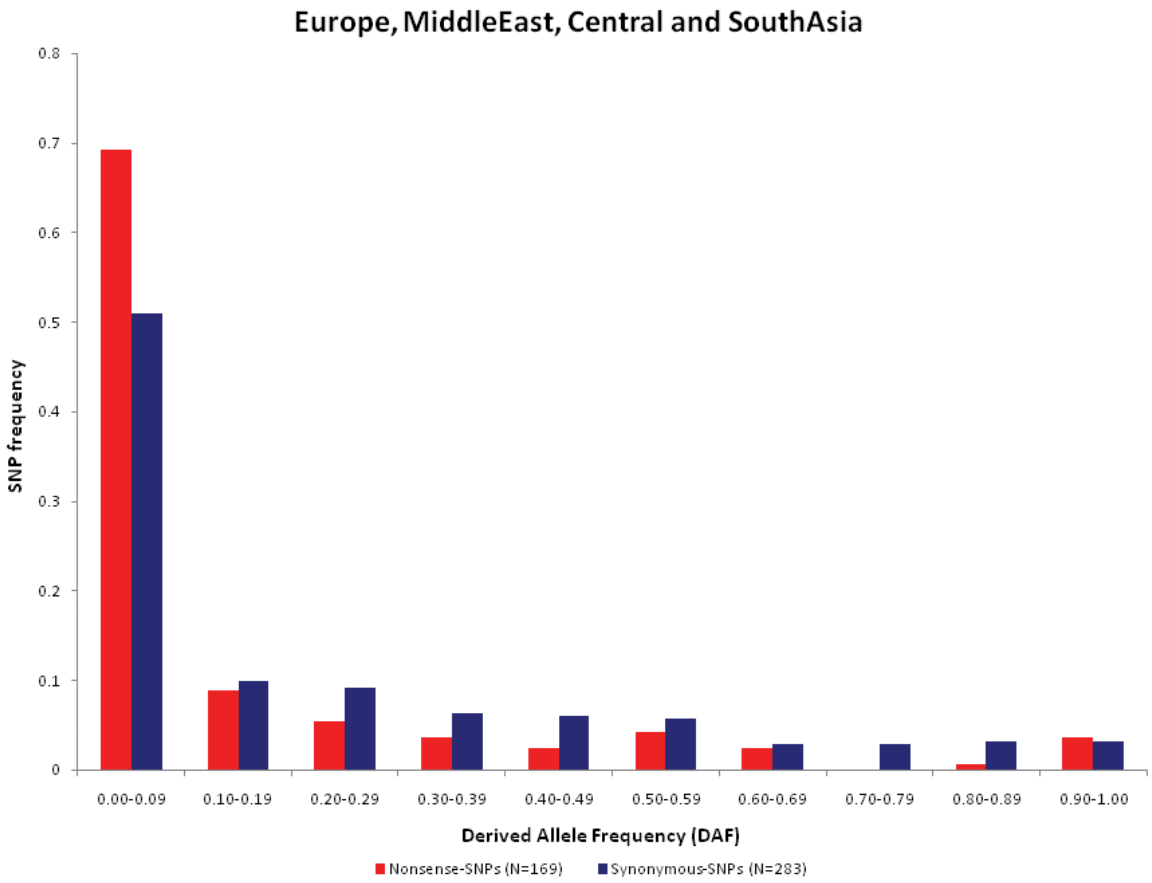
Appendix D

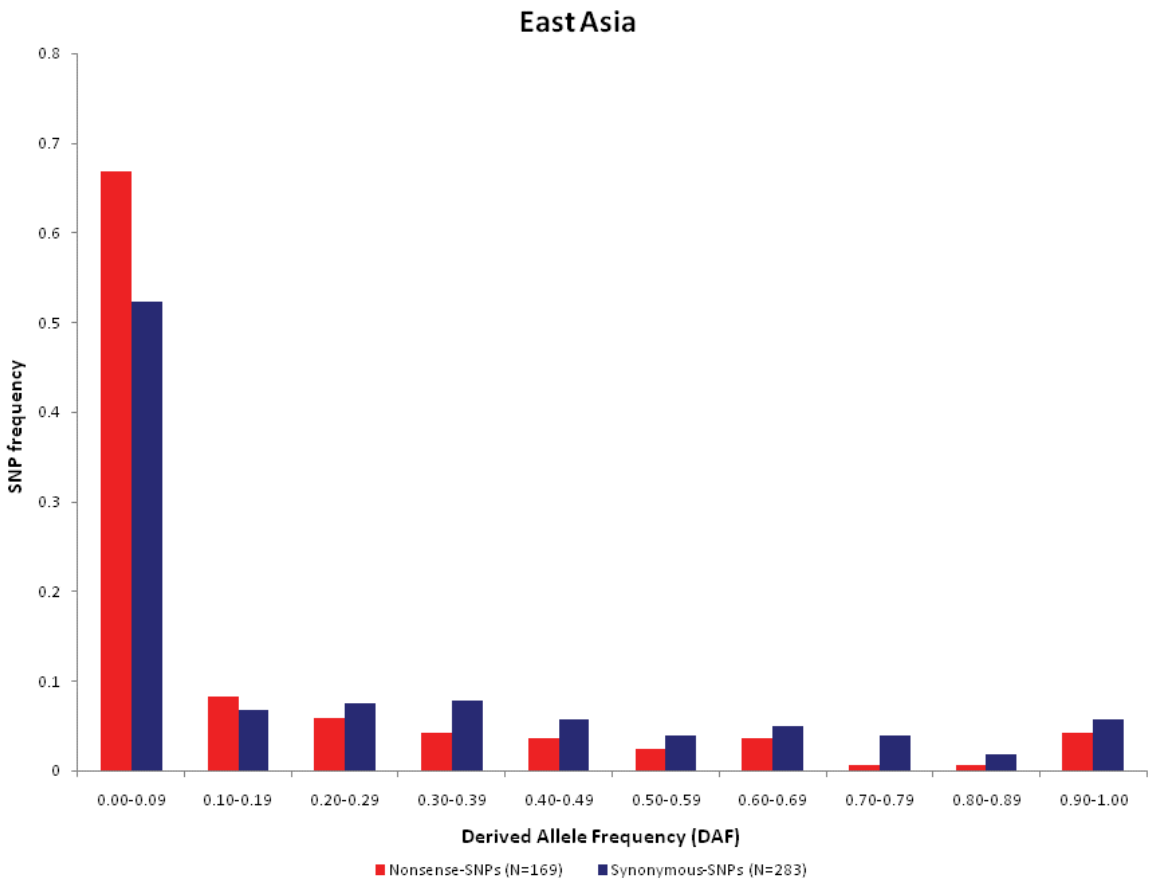
All genotype data is available in a tab delimted text file on the accompanying CD.

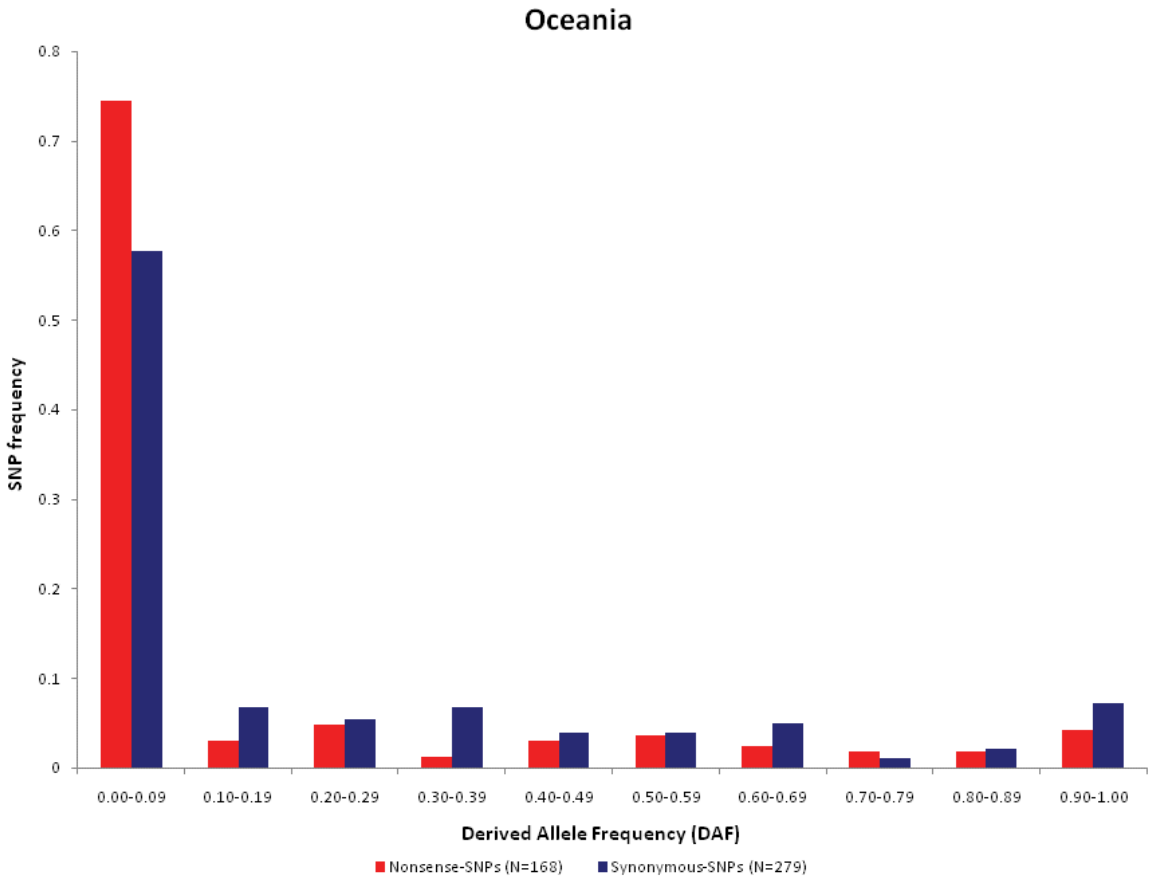
Appendix E

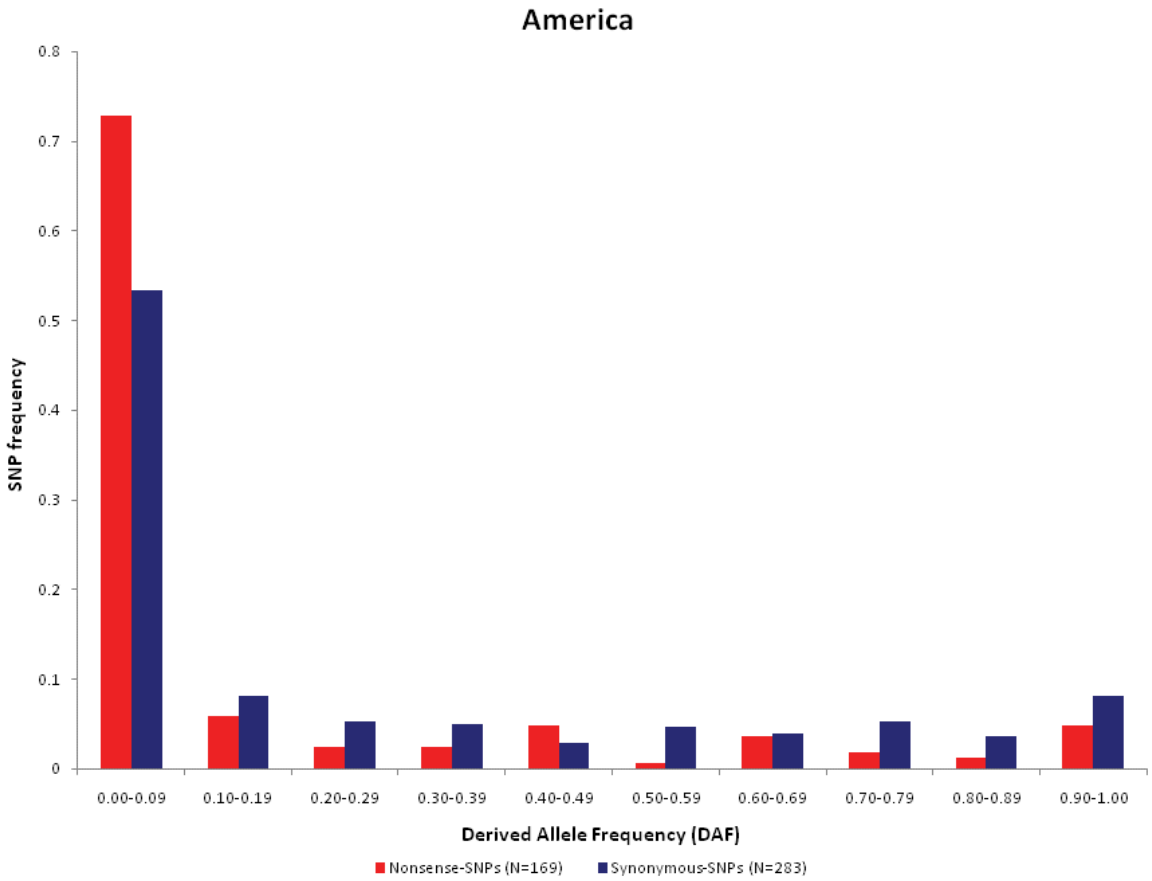
The derived allele frequency spectrum for the nonsense- and synonymous-SNPs plotted for the five populations (according to K=5 in Rosenberg et al. 2002) separately. The distribution was found to be similar to that of the combined populations.




Europe, MiddleEast, Central and SouthAsia
SNP frequency
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.00-0.09
0.10-0.19
0.20-0.29
0.30-0.39
0.40-0.49
0.50-0.59
0.60-0.69
0.70-0.79
0.80-0.89
0.90-1.00
Derived Allele Frequency (DAF)
Nonsense-SNPs (N=169)
Synonymous-SNPs (N=283)


East Asia
SNP frequency
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.00-0.09
0.10-0.19
0.20-0.29
0.30-0.39
0.40-0.49
0.50-0.59
0.60-0.69
0.70-0.79
0.80-0.89
0.90-1.00
Derived Allele Frequency (DAF)
Nonsense-SNPs (N=169)
Synonymous-SNPs (N=283)


Oceania
SNP frequency
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.00-0.09
0.10-0.19
0.20-0.29
0.30-0.39
0.40-0.49
0.50-0.59
0.60-0.69
0.70-0.79
0.80-0.89
0.90-1.00
Derived Allele Frequency (DAF)
Nonsense-SNPs (N=168)
Synonymous-SNPs (N=279)


America
SNP frequency
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.00-0.09
0.10-0.19
0.20-0.29
0.30-0.39
0.40-0.49
0.50-0.59
0.60-0.69
0.70-0.79
0.80-0.89
0.90-1.00
Derived Allele Frequency (DAF)
Nonsense-SNPs (N=169)
Synonymous-SNPs (N=283)

Appendix F

Summary of results from genome-wide survey of nonsense SNPs (also available on accompanying CD). The displayed includes: External Gene ID (normally from Hugo), the SNP ID, chromosome, chromosomal position (Build36), %truncated, prediction of whether NMD is triggered (YES/NO), heterozygosity calculated according to Nei (1987), the minor allele frequency (MAF) and the derived allele frequency (DAF), FST value calculated according to Weir and Cockerham (1984) accross the combined 37 populations used in this study (HGDP-CEPH and HapMap), across the HGDP-CEPH populations divided into 5 geographical regions according to K=5 in Rosenberg et al (2002), and in the genotypes in the publicly available HapMap data (The International HapMap Consortium 2005). The table is sorted according to Gene ID.

Gene ID	SNP ID	Chr	Position (B36)	Truncated (%)	NMD	Heterozygosity	MAF	DAF	F_{ST} (37 pops in study)	F_{ST} (5 pop division)	F_{ST} (HapMap,4 pops)
	rs11727979	4	9014324	28.4	NO	0.01	0.01	0.02	0.01	0.01	
	rs34358	5	75000878	18.3	YES	0.47	0.53	0.09	0.50	0.06	0.07
	rs6997135	8	57125768	52.6	NO	0.01	0.01	0.08	0.02	0.09	0.09
ABCA10	rs10491178	17	64661568	14.4	YES	0.11	0.11	0.11	0.20	0.14	0.01
ADAM10	rs1801973	15	56689928	16.0	YES	0.00	0.00	0.00	0.00	0.00	
AKAP9	rs2285686	7	91525621	75.8	YES	0.00	0.00	0.00	0.00	0.00	
ALOX15	rs11870258	17	4488681	70.3	YES	0.00	0.00	0.00	0.00	0.00	0.00
AMPD1	rs17602729	1	115037580	98.4	YES	0.04	0.04	0.06	0.08	0.06	0.13
ANKRD1	rs1130407	10	92670013	89.4	YES	0.00	0.00	0.06	0.00	0.00	0.00
APOL3	rs11089781	22	34886714	85.6	YES	0.02	0.02	0.26	0.04	0.15	
ARHGEF19	rs12048007	1	16407971	93.0	YES	0.00	0.00	0.00	0.00	0.00	
ASCC1	rs11000217	10	73634249	80.7	YES	0.09	0.09	0.11	0.16	0.15	0.17

Gene ID	SNP ID	Chr	Position (B36)	Truncated (%)	NMD	Heterozygosity	MAF	DAF	F_{ST} (37 pops in study)	F_{ST} (5 pop division)	F_{ST} (HapMap,4 pops)
BARHL2	rs1335726	1	90952878	44.1	YES	0.00	0.00	0.00	0.00	0.00	0.00
BEX1	rs11550088	X	102204783	79.4	NO	0.00	0.00	0.01	0.00	0.00	0.00
BRCA2	rs11571833	13	31870626	2.7	NO	0.01	0.01	0.04	0.02	0.01	0.00
C14orf129	rs12435565	14	95918572	43.6	NO	0.00	0.00	0.00	0.00	0.00	0.00
C14orf149	rs3177474	14	59015670	37.7	YES	0.00	0.00	0.01	0.00	0.00	0.00
C16orf24	rs12931094	16	711847	55.5	YES	0.00	0.00	0.00	0.00	0.00	
C19orf44	rs3826726	19	16481328	40.7	YES	0.00	0.00	0.04	0.01	0.01	0.00
C1orf105	rs7532205	1	170688829	91.4	YES	0.04	0.04	0.26	0.09	0.25	
C1orf157	rs11803208	1	202273170	10.9	YES	0.01	0.01	0.00	0.01	0.00	0.00
C21orf111	rs12329656	21	45588725	10.9	NO	0.01	0.01	0.09	0.03	0.08	0.07
C4orf33	rs10009430	4	130250213	28.0	YES	0.01	0.01	0.04	0.01	0.02	0.07
C5orf20	rs12520799	5	134810349	52.2	NO	0.39	0.39	0.15	0.48	0.11	0.24
C6orf148	rs16883571	6	74076059	85.7	YES	0.04	0.04	0.04	0.08	0.02	0.02
C8orf49	rs809203	8	11656913	16.0	NO	0.27	0.27	0.11	0.40	0.10	0.14
CARD8	rs2043211	19	53429518	97.7	YES	0.34	0.34	0.06	0.45	0.07	0.07
CASP12	rs497116	11	104268327	63.7	YES	0.04	0.96	0.24	0.07	0.24	0.10
CD22	rs25677	19	40523826	54.2	YES	0.00	0.00	0.00	0.00	0.00	
CD2BP2	rs11547274	16	30273069	84.8	YES	0.00	0.00	0.00	0.00	0.00	0.01
CD36	rs3211938	7	80138385	31.3	YES	0.02	0.02	0.24	0.03	0.09	0.25
CD99	rs4268274	X	2647719	76.5	YES	0.00	0.00	0.00	0.00	0.00	
CDKL1	rs7148089	14	49872477	0.5	NO	0.41	0.59	0.03	0.48	0.02	0.03
CDKL1	rs11570829	14	49879445	22.5	YES	0.00	0.00	0.05	0.01	0.02	0.05
CEL	rs13287310	9	134936348	29.4	YES	0.00	0.00	0.00	0.00	0.00	
CLCA3	rs2292830	1	86873963	90.4	YES	0.45	0.55	0.08	0.50	0.08	0.12
CLEC7A	rs16910526	12	10162354	5.9	NO	0.05	0.05	0.03	0.09	0.04	0.06
CMA1	rs13306254	14	24046497	86.0	YES	0.00	0.00	0.01	0.01	0.02	0.00
CPN2	rs4974538	3	195543601	6.8	NO	0.20	0.20	0.06	0.32	0.02	0.00
CRHR2	rs8192492	7	30659687	7.0	NO	0.00	0.00	0.01	0.00	0.00	0.00

Gene ID	SNP ID	Chr	Position (B36)	Truncated (%)	NMD	Heterozygosity	MAF	DAF	F_{ST} (37 pops in study)	F_{ST} (5 pop division)	F_{ST} (HapMap,4 pops)
CST2	rs6049157	20	23753918	35.9	YES	0.01	0.01	0.10	0.01	0.08	0.00
CYFIP2	rs7705781	5	156700710	29.8	YES	0.00	0.00	0.02	0.01	0.03	0.00
DEPDC1	rs12759438	1	68717221	2.0	NO	0.09	0.09	0.04	0.16	0.04	0.03
DGCR8	rs2106143	22	18453499	99.4	YES	0.00	0.00	0.00	0.00	0.00	
DHDH	rs10423255	19	54137586	30.4	YES	0.04	0.04	0.01	0.08	0.00	0.00
DKC1	rs2853347	X	153647431	85.2	YES	0.00	0.00	0.00	0.00	0.00	0.00
DMN	rs5030689	15	97463526	87.2	YES	0.00	0.00	0.00	0.00	0.00	
DSCR8	rs2836172	21	38450325	13.3	NO	0.06	0.06	0.10	0.12	0.10	0.09
ELP4	rs3026403	11	31761622	1.9	NO	0.00	0.00	0.01	0.00	0.00	0.00
ENOPH1	rs11546516	4	83571058	93.1	YES	0.00	0.00	0.00	0.00	0.01	0.00
FAM19A5	rs3752466	22	47531852	77.2	NO	0.00	0.00	0.07	0.00	0.01	0.00
FCGR2A	rs9427397	1	159742828	80.4	YES	0.00	0.00	0.00	0.00	0.00	0.00
FCN3	rs15544	1	27573458	81.7	YES	0.00	0.00	0.01	0.00	0.00	
FLJ41766	rs12446322	16	21234774	87.3	NO	0.16	0.16	0.13	0.26	0.13	0.30
FMO2	rs2020866	1	169439745	47.2	YES	0.00	0.00	0.00	0.00	0.00	0.01
FMO2	rs6661174	1	169444714	11.8	YES	0.04	0.96	0.28	0.08	0.24	0.14
FMO6P	rs1736565	1	169379114	80.6	YES	0.36	0.64	0.08	0.46	0.10	0.13
FUT2	rs1800028	19	53898629	41.3	YES	0.00	0.00	0.00	0.00	0.00	
GLUD2	rs10657	X	120010633	15.6	NO	0.00	0.00	0.00	0.00	0.00	
GPNMB	rs11537976	7	23280348	1.1	NO	0.00	0.00	0.00	0.00	0.00	0.00
GRIK5	rs1143143	19	47201928	30.3	YES	0.00	0.00	0.00	0.00	0.00	
HERC6	rs4413373	4	89582627	0.2	NO	0.01	0.99	0.02	0.01	0.02	0.02
HPS4	rs3747129	22	25192041	53.5	YES	0.20	0.20	0.10	0.32	0.10	0.13
IDI2	rs1044261	10	1055710	36.8	NO	0.04	0.04	0.04	0.08	0.04	0.02
IL17RB	rs1043261	3	53874316	3.8	NO	0.15	0.15	0.04	0.25	0.03	0.01
INMT	rs6966017	7	30761567	53.4	NO	0.00	0.00	0.03	0.00	0.01	
KIAA0748	rs1801876	12	53630291	3.8	NO	0.36	0.36	0.24	0.46	0.22	0.40
KIAA1704	rs9567515	13	44461859	93.0	YES	0.00	0.00	0.00	0.00	0.00	0.00

Gene ID	SNP ID	Chr	Position (B36)	Truncated (%)	NMD	Heterozygosity	MAF	DAF	F_{ST} (37 pops in study)	F_{ST} (5 pop division)	F_{ST} (HapMap,4 pops)
KRT7	rs11558308	12	50913627	80.0	YES	0.00	0.00	0.00	0.00	0.00	
KRTAP13-1	rs1985418	21	30690365	73.7	NO	0.00	0.00	0.08	0.01	0.02	0.00
KRTAP13-2	rs877346	21	30665998	23.3	NO	0.28	0.72	0.06	0.41	0.09	0.09
LCE5A	rs2282298	1	150750869	33.6	NO	0.01	0.01	0.04	0.02	0.04	0.02
LCN10	rs9886752	9	138754316	21.3	YES	0.22	0.22	0.09	0.35	0.09	0.02
LGALS1	rs4887	22	36404553	49.3	YES	0.00	0.00	0.00	0.00	0.00	
LPL	rs328	8	19864004	0.4	NO	0.09	0.09	0.04	0.16	0.04	0.02
MAGEE2	rs1343879	X	74921254	77.1	NO	0.31	0.31	0.54	0.43	0.53	0.88
MATN4	rs2233091	20	43366762	91.2	YES	0.00	0.00	0.00	0.00	0.00	0.00
MCTP2	rs2289010	15	92712024	72.6	YES	0.00	0.00	0.00	0.00	0.00	0.00
MLLT11	rs11546017	1	149287905	6.4	NO	0.00	0.00	0.00	0.00	0.00	
MOBKL2C	rs6671527	1	46853266	91.1	YES	0.34	0.34	0.15	0.45	0.12	0.24
MOSPD3	rs1053507	7	100048504	78.0	YES	0.00	0.00	0.01	0.00	0.00	
MS4A12	rs2298553	11	60021578	73.5	YES	0.50	0.50	0.03	0.50	0.02	0.03
MST1R	rs9819888	3	49910507	55.7	YES	0.00	0.00	0.00	0.00	0.01	
NAT1	rs5030839	8	18124395	35.7	NO	0.00	0.00	0.02	0.01	0.00	
NLRP8	rs306457	19	61191091	0.9	NO	0.25	0.25	0.05	0.38	0.03	0.01
NOP5_HUMAN	rs15160	2	202870470	24.7	YES	0.00	0.00	0.00	0.00	0.00	
NP_001073929.1	rs13062420	3	171023372	41.5	YES	0.00	0.00	0.01	0.00	0.00	
NP_064546.2	rs2176186	2	228184384	6.6	NO	0.33	0.67	0.07	0.44	0.04	0.05
NP_438169.2	rs1128610	13	31876483	68.1	NO	0.00	0.00	0.01	0.00	0.00	
NP_660151.2	rs11542462	16	80591311	92.4	NO	0.07	0.07	0.08	0.13	0.10	
NP_775760.2	rs1023840	5	41097472	88.0	YES	0.26	0.26	0.08	0.39	0.08	0.16
NP_899231.1	rs2407221	4	152432053	4.8	NO	0.20	0.20	0.19	0.32	0.24	0.09
NPPA	rs5065	1	11828655	0.7	NO	0.15	0.85	0.15	0.26	0.14	0.26
OR10X1	rs863362	1	156816116	84.2	NO	0.47	0.53	0.03	0.50	0.02	0.01
OR1B1	rs1476860	9	124431062	39.8	NO	0.40	0.40	0.21	0.48	0.25	0.26
OR2D2	rs16919417	11	6870116	79.3	NO	0.01	0.01	0.04	0.02	0.06	0.03

Gene ID	SNP ID	Chr	Position (B36)	Truncated (%)	NMD	Heterozygosity	MAF	DAF	F_{ST} (37 pops in study)	F_{ST} (5 pop division)	F_{ST} (HapMap,4 pops)
OR4C16	rs1459101	11	55096228	94.5	NO	0.33	0.33	0.12	0.44	0.16	0.05
OR4X1	rs10838851	11	48242807	10.8	NO	0.39	0.61	0.06	0.47	0.03	0.07
OR4X2	rs7120775	11	48223312	91.1	NO	0.16	0.16	0.04	0.26	0.03	0.06
OR51I1	rs16930998	11	5419278	96.2	NO	0.13	0.13	0.15	0.22	0.17	0.25
OR5AK2	rs13343184	11	56512974	98.7	NO	0.00	0.00	0.00	0.01	0.00	
OR5D13	rs11230980	11	55297590	89.2	NO	0.00	0.00	0.00	0.00	0.00	0.01
OR7G3	rs17001893	19	9098263	61.0	NO	0.02	0.02	0.13	0.03	0.13	0.08
OVCH2	rs4509745	11	7669047	1.6	NO	0.49	0.51	0.12	0.50	0.13	0.23
PCDHB10	rs3733689	5	140552340	98.6	NO	0.00	0.00	0.01	0.00	0.00	0.00
PGAM2	rs10250779	7	44071421	69.3	YES	0.00	0.00	0.02	0.00	0.01	0.00
PKD1L3	rs4788587	16	70558637	54.4	YES	0.27	0.27	0.07	0.39	0.08	0.06
PKM2	rs11558352	15	70288149	79.8	YES	0.00	0.00	0.00	0.00	0.00	0.00
PLAT	rs1804182	8	42152676	0.5	NO	0.00	0.00	0.05	0.01	0.01	0.04
PML	rs11272	15	72122464	32.3	YES	0.00	0.00	0.01	0.00	0.00	0.01
PRL	rs6238	6	22398525	48.7	YES	0.00	0.00	0.01	0.00	0.00	0.00
PTPRE	rs3206183	10	129737879	89.3	YES	0.00	0.00	0.00	0.00	0.00	
Q2M2F3_HUMAN	rs7703216	5	177331574	32.3	NO	0.12	0.12	0.04	0.21	0.04	0.01
Q5R387_HUMAN	rs12139100	1	20374169	81.0	YES	0.21	0.21	0.05	0.33	0.05	0.04
Q5SVS6_HUMAN	rs9567547	13	44863210	80.6	NO	0.01	0.01	0.10	0.03	0.08	0.02
Q8IXR4_HUMAN	rs1001586	22	41000237	0.9	NO	0.17	0.17	0.07	0.28	0.09	0.00
Q8N7E8_HUMAN	rs16885508	5	55797209	6.5	NO	0.00	0.00	0.06	0.01	0.03	0.01
Q8N8G3_HUMAN	rs4723884	7	39615800	68.4	NO	0.22	0.22	0.23	0.35	0.26	0.18
Q8NH80_HUMAN	rs2512227	11	123561942	19.5	NO	0.50	0.50	0.07	0.50	0.08	
Q96NA9_HUMAN	rs2400941	14	100370320	96.6	YES	0.27	0.27	0.07	0.39	0.05	0.13
Q96NK0_HUMAN	rs13422553	2	201853604	40.0	NO	0.17	0.17	0.07	0.28	0.04	0.15
Q9H579-2	rs11539065	20	35241114	94.2	YES	0.00	0.00	0.01	0.00	0.00	
Q9UI72_HUMAN	rs642354	5	32185000	75.7	NO	0.01	0.01	0.10	0.03	0.11	0.09
RBPJ	rs5007634	4	26035389	71.6	YES	0.00	1.00	0.01	0.00	0.00	0.00

Gene ID	SNP ID	Chr	Position (B36)	Truncated (%)	NMD	Heterozygosity	MAF	DAF	F_{ST} (37 pops in study)	F_{ST} (5 pop division)	F_{ST} (HapMap,4 pops)
REG4	rs1052972	1	120138308	8.8	NO	0.49	0.49	0.21	0.50	0.27	0.24
ROBO1	rs1065217	3	78749727	20.1	YES	0.00	0.00	0.00	0.00	0.00	0.00
RORC	rs17582155	1	150070837	98.1	YES	0.00	0.00	0.00	0.00	0.00	0.00
RRM2	rs15516	2	10186932	2.6	NO	0.00	0.00	0.00	0.00	0.00	
SEMA4C	rs12471298	2	96890515	16.9	NO	0.04	0.04	0.47	0.08	0.52	
SEMG1	rs2233885	20	43270193	39.3	YES	0.00	0.00	0.00	0.00	0.00	0.00
SERPINA10	rs2232698	14	93826422	80.2	YES	0.02	0.02	0.03	0.04	0.03	0.00
SIGLEC12	rs16982743	19	56696715	95.1	YES	0.20	0.20	0.22	0.32	0.28	0.17
SLAIN1	rs17777179	13	77216390	89.3	NO	0.05	0.95	0.07	0.10	0.10	0.05
SLAMF8	rs10430458	1	158066432	77.3	YES	0.01	0.01	0.03	0.02	0.03	0.01
SLC17A4	rs2328894	6	25886161	13.1	YES	0.00	0.00	0.01	0.01	0.00	0.00
SLC22A10	rs1790218	11	62814501	82.3	YES	0.43	0.43	0.08	0.49	0.08	0.06
SLC25A5	rs11552294	X	118486534	96.7	YES	0.00	0.00	0.00	0.00	0.00	
SLC41A3	rs11543281	3	127269593	89.4	YES	0.00	0.00	0.00	0.00	0.00	
SLC7A8	rs17183863	14	22668816	41.1	YES	0.11	0.11	0.00	0.20	0.00	0.02
SOX13	rs3737659	1	202362310	15.8	YES	0.18	0.18	0.10	0.29	0.13	0.16
SPATA8	rs3812907	15	95128397	67.9	YES	0.12	0.12	0.10	0.22	0.11	0.05
SPG7	rs1057803	16	88143204	40.4	YES	0.00	0.00	0.00	0.00	0.00	0.00
SPTBN5	rs2271286	15	39972774	98.0	YES	0.01	0.01	0.03	0.02	0.04	0.03
SRD5A2	rs9332960	2	31659458	97.6	YES	0.00	0.00	0.00	0.00	0.00	
STARD6	rs17292725	18	50134887	91.4	YES	0.02	0.02	0.02	0.04	0.03	0.05
SURF4	rs2240173	9	135220580	1.1	NO	0.00	0.00	0.00	0.00	0.00	0.01
SYNE2	rs2781377	14	63629845	88.2	YES	0.09	0.09	0.00	0.16	0.00	0.00
TAAR2	rs8192646	6	132980535	59.9	NO	0.06	0.06	0.11	0.12	0.10	0.11
TALDO1	rs1804554	11	754350	11.2	YES	0.00	0.00	0.00	0.00	0.00	
TANC1	rs6755758	2	159794817	17.0	NO	0.00	0.00	0.00	0.00	0.00	
TBCA	rs1802165	5	77039855	60.6	YES	0.00	0.00	0.00	0.00	0.00	
TCP11L1	rs3758741	11	33063192	4.3	YES	0.27	0.27	0.09	0.40	0.08	0.11

Gene ID	SNP ID	Chr	Position (B36)	Truncated (%)	NMD	Heterozygosity	MAF	DAF	F_{ST} (37 pops in study)	F_{ST} (5 pop division)	F_{ST} (HapMap,4 pops)
THSD7B	rs12622896	2	137746704	52.6	YES	0.04	0.96	0.07	0.08	0.01	0.01
TLR4	rs5030720	9	119516006	29.4	NO	0.00	0.00	0.00	0.00	0.00	
TMEM143	rs16982007	19	53537872	62.8	YES	0.00	0.00	0.03	0.01	0.01	0.05
TMEM162	rs541169	19	40410860	49.2	YES	0.32	0.32	0.07	0.43	0.06	0.01
TMPRSS7	rs340142	3	113263361	68.8	YES	0.01	0.01	0.08	0.01	0.07	0.05
TREM2	rs2234258	6	41234407	13.2	NO	0.00	0.00	0.03	0.01	0.01	
TRPM1	rs3784589	15	29082006	14.9	NO	0.07	0.07	0.06	0.13	0.05	0.01
TSNARE1	rs11988455	8	143308760	0.4	NO	0.01	0.01	0.17	0.01	0.05	0.04
UNC93A	rs2235197	6	167629692	67.0	YES	0.12	0.12	0.04	0.21	0.03	0.04
USP29	rs9973206	19	62334594	1.1	NO	0.04	0.04	0.09	0.08	0.05	0.15
UTS2D	rs16866426	3	192475738	7.5	NO	0.00	0.00	0.01	0.00	0.01	0.01
WDR37	rs10794716	10	1132208	5.3	NO	0.01	0.01	0.07	0.02	0.06	0.06
WRN	rs11574410	8	31150077	1.9	NO	0.00	0.00	0.00	0.01	0.00	0.00
XR_017624.1	rs17107991	14	70005689	2.9	NO	0.06	0.06	0.03	0.12	0.02	0.04
ZAN	rs2293766	7	100209294	33.0	YES	0.26	0.26	0.40	0.39	0.39	0.50
ZNF544	rs3745136	19	63465654	12.8	NO	0.00	0.00	0.00	0.00	0.00	0.00
ZSWIM3	rs11557696	20	43940042	31.1	NO	0.00	0.00	0.00	0.00	0.00	0.00

Appendix G

SNP variation data for the resequenced genes *MAGEE2* is available in tab delimited text files on the accompanying CD.