# Molecular Evidence for Dietary Adaptation in Humans

**Elizabeth Caldwell**
**Ph.D. Genetics Thesis**

**The Centre for Genetic Anthropology**
**Department of Biology**
**University College London**

**Supervisor: Dr. Mark G. Thomas**

**Abstract**

Starch digestion begins in the mouth where it is hydrolysed into smaller polysaccharides by the enzyme salivary amylase. Three salivary amylase genes (AMY1A, B & C) and a pseudogene (AMYP1) have been described and are located in tandem on chromosome 1. Polymorphic variation has been demonstrated in Caucasians in the form of the number of repeats of the AMY1 genes, as follows: (1A-1B-P1)n-1C. This variation has been reported to result in differing levels salivary amylase enzyme production and, as a result, differences in the efficiency of starch digestion in the mouth. It is proposed in this thesis that an increase in salivary gene copy number may be an adaptation to high starch diets as a result of the adoption of agriculture. Reliable high-throughput multiplex PCR based methods have been designed to quantify AMY1 gene copy number and to also to type 6 microsatellite markers closely linked to the AMY gene cluster. Data have been collected for 14 human populations, with different histories of cereal agriculture and ancestral levels of starch in the diet. Data have also been collected on AMY1 gene copy number in 5 chimpanzees (*Pan troglodytes*).

The AMY1 allele frequency difference (measured using $F_{ST}$) between the two most extreme populations, the Mongolians and Saami, was not an outlier on a distribution of $F_{ST}$ based on presumed neutral 11,024 SNPs from the human genome. The chimpanzee data suggest that the most frequent allele (AMY1*H1) in humans may not be the ancestral allele, as all chimpanzee chromosomes tested carried the AMY1*H0 allele (containing only one copy of the AMY1 gene). A more sensitive selection test, the analysis of the intra-allelic variability of the AMY1 repeat alleles using closely linked microsatellites, showed no compelling evidence for recent positive selection at the AMY1 locus in humans. As a result, genetic drift could not be ruled out as an explanation for the observed AMY1 allele frequency differences among populations.

Alanine:glyoxylate aminotransferase (AGT) is an intermediary metabolic enzyme that is targeted to different organelles in different species. Previous studies have shown that there is a clear relationship between the organellar distribution of AGT and diet. Non-human primates show the herbivorous peroxisomal distribution of AGT. In humans a point mutation and insertion deletion polymorphism have been associated with peroxisome-to-mitochondria AGT mis-targeting. Data have been collected using a PCR/RFLP based method, in 11 human populations. In a comparison with $F_{ST}$ values from 11.024 SNP loci, 94.5% of SNPs had a lower $F_{ST}$ than a comparison of AGT allele frequencies for Saami and Chinese. This unusually high allele frequency difference between Chinese and Saami is consistent with the signature of recent positive selection driven by the unusually high meat content in the Saami diet.

## Contents

## List of Figures and Tables

**Chapter 4:**

**Acknowledgements**

Firstly, I would like to thank my supervisors Dr. Mark Thomas and Prof. Dallas Swallow. Mark has patiently coached me in molecular biological techniques since I first walked into his lab wondering how genetics could be used as a tool for studying human adaptation. He also lent his considerable expertise in protocol design to the technical challenges involved in designing PCR based assays posed by the amylase gene family. Most importantly, he has always provided me with the encouragement and inspiration to keep going just when I thought everything was hopeless. Dallas has also provided invaluable support all the way through the project and I would especially like to thank her for her help and advice when it came to writing this thesis.

The design and optimisation of the protocols for quantifying salivary amylase genes would not have been possible without the provision of DNA samples of known salivary amylase phenotype which were kindly provided by Prof Jan Pronk, Vrije Universiteit, Amsterdam.

I am indebted to Dr. Neil Bradman and The Centre for Genetic Anthropology (TCGA) at UCL, which have provided me with access to the extensive collection of DNA samples during this project. I would also like to thank all those who donated samples to the TCGA, as well as the tireless efforts of those who collected the samples in the field and extracted the samples in the lab. In particular I would like to thank: Dr. Ayele Tarekegn (Ethiopians), Prof. Levon Episkoposyan (Armenians), Ms Amanda Bradman (Ashkenazi Jews), Ms Noreen von Crammon-Taubadel (Germans and Irish), Prof. Pagbajavyn Nymadawa (Mongolians), Mr Mathew Sears (British), Ms Leila Leredj (Algerians) and Dr. Susanna Albustan (Kuwait). In addition I would like to thank Prof. David Goldstein for access to the Singapore Chinese samples and Prof. Dallas Swallow for access to the five chimpanzee DNA samples.

I would like to thank Dr. Mike Weale and Dr. Michael Stumpf for their assistance and advice with the statistical analysis of the data presented in this thesis. Mike Weale gave up many a lunch and coffee break to answer my endless questions about statistics. Michael Stumpf also devoted much of his energy to designing programs for the analysis of intra allelic variability in time to be utilised in this project.

Ms Noreen von Crammon-Taubadel was the first person other than myself to put the salivary amylase quantification protocols into practice. Supervising Noreen whilst she typed the Irish, German, Kuwait and Algerian samples finally gave me the confidence that I had designed a workable protocol that someone else could follow.

My thanks also to those long suffering members of the Thomas lab who had to endure my questions, debates, mood swings and music taste for the duration of my PhD: Dr. Ben Fletcher, Ms Abigail Jones, Dr. Kathy Dunn, Ms Charlotte Mulcare, Mr Krisna Veeramah and Dr Ian Barnes. I would also like to thank Charlotte and Ian for reading various chapters in their very rough draft state, before I dared show them to either of my supervisors. Ian also taught me how to clone PCR products, for which I am extremely grateful.

I would also like to thank Prof. Chris Danpure and Ms Lianne Mayor who consented to my desire to work the AGT project, on which they had made such a promising start.

This project was funded by The Wellcome Trust Bioarchaeology Studentship Scheme, without whose generous support the experience of completing this thesis would have been considerably more difficult.

My thanks in advance to my viva examiners – some of the lucky few that will read this thesis cover to cover.

And finally, I would like to thank my friends and family for their ongoing support and encouragement.
Elizabeth Caldwell
June 2004

## Abbreviations

| | |
|---|---|
| A | adenine |
| ADH | alcohol dehydrogenase |
| AGT | alanine:glyoxylate aminotransferase |
| AGXT | alanine:glyoxylate aminotransferase gene |
| AIDS | acquired immune deficiency syndrome |
| AMOVA | analysis of molecular variance |
| AMY1 | salivary amylase gene |
| AMY2 | pancreatic amylase gene |
| ASD | average squared distance |
| | |
| BAC | bacterial artificial chromosome |
| bp | base pairs |
| BP | before present |
| | |
| C | cytosine |
| cM | centiMorgan |
| | |
| DNA | deoxyribonucleic acid |
| $d_N$ | non-synonomous sites |
| $d_S$ | synonomous sites |
| | |
| EDNP | energy dense nutrient poor |
| EHH | extended haplotype homozygosity |
| EM | expectation-maximisation (algorithm) |
| ETPD | exact test of population differentiation |
| | |
| G | guanine |
| G6PD | gluscose-6-phosphate dehydrogenase |
| | |
| $Hb^S$ | haemoglobin sickle cell allele |
| HIV | human immunodeficiency virus |
| HKA | Hudson-Kreitman-Aguade |
| HLA | human leukocyte antigen |
| | |
| kb | kilobases / kilobase-pairs |
| kg | kilogram |
| | |
| LD | linkage disequilibrium |
| LGM | last glacial maxim |
| | |
| $\mu$ | mutation rate |
| Mb | megabases / megabase-pairs |

| | |
|---|---|
| MC1R | melanocortin-1 receptor |
| μg | microgram |
| MHC | major histocompatibility complex |
| ML | maximum likelihood |
| MRCA | most recent common ancestor |
| mtDNA | mitochondrial DNA |
| mya | million years ago |
| $N_e$ | effective population size |
| NEAP | net endogenous acid production |
| NIDDM | non-insulin dependent diabetes mellitus |
| NJ | neighbour joining |
| PAGE | polyacrylamide gel electrophoresis |
| PCR | polymerase chain reaction |
| RFLP | restriction fragment length polymorphism |
| SI | sucrase isomaltase |
| SMM | stepwise mutation model |
| SNP | single nucleotide polymorphism |
| STR | short tandem repeat (polymorphism) |
| T | thymine |
| Taq | *Thermus aquaticus* |
| TMRCA | time to most recent common ancestor |
| UEP | unique event polymorphism |
| VNTR | variable number of tandem repeats (polymorphism) |
| ybp | years before present |

# Chapter 1:Introduction

## 1.1 General overview

Diet is a major factor in the adaptation of an organism to its environment (Ulijaszek & Strickland 1993). An individual must be able to obtain and digest the food available in order to survive and reproduce. Diet has a major influence on many aspects of an organism's anatomy, reproductive strategies and behaviour (see Fig 1.1). Over the course of human evolution, subsistence patterns have undergone a number of dramatic changes that have had a major effect on the survival and success of the human species. These changes include a shift from a predominantly herbivorous primate heritage to a diet that includes a significant proportion of meat, the use of non-oral food preparation techniques and the intensive control of plant and animal resources through domestication and agriculture (Gordon 1987).

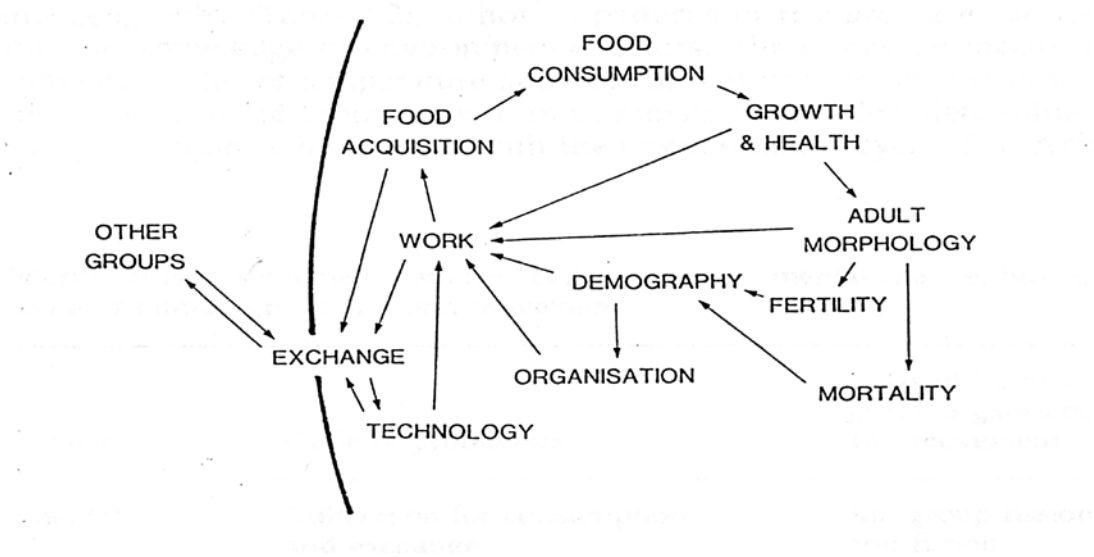



Fig 1.1: The ecology of nutritional adaptation in humans (after Ulijaszek & Strickland 1993) showing the complex inter-relationships between food and other aspects of the life cycle and adaptive niche.

Since Darwin's theory of natural selection was first published in 1859, questions have been raised about the biological processes behind human evolution. The study of genetics, starting with Mendel, has provided a framework to study adaptation and natural selection at the molecular level. Currently, one of the challenges in the study of human evolution is to pinpoint, in the language of

genetics, how humans evolved (see Enard et al. 2002*a*, Jobling et al. 2004). Given its importance in the survival of any species, diet is an obvious place to look in order to find molecular evidence for adaptation in humans.

This chapter gives an introduction to some of the issues involved in the search for molecular evidence of dietary adaptation in humans. Firstly it surveys existing examples of nutritional adaptation. Then it examines the methods available for detecting the signature of selection in genetic data. Following this is an outline the history of human subsistence including the origins and spread of agriculture, as well as an overview of the methods used in the reconstruction of pre-agricultural diets. Finally the chapter gives an introduction to the human alpha-amylase multigene family, which as will be explained, forms the basis of the hypothesis for dietary adaptation explored in this thesis.

## 1.2 Examples of Nutritional Adaptation

There are number of ways in which humans can adapt in order to exploit new foods. With their increased brain size and manual dexterity, humans have developed a number of cultural practices to exploit foodstuffs that would otherwise be inedible. In terms of biological adaptations, however, there are examples of changes in gross anatomy, such as shortened gut for digesting smaller amounts of high quality foods (Aiello & Wheeler 1995), as well as changes on a molecular level. These molecular adaptations often involve differences in enzymes involved in the metabolism of dietary components.

The most frequently cited example of dietary adaptation on a molecular level is lactase persistence in adulthood (see Swallow 2003 for review). Lactase (lactase phlorizin hydrolase) is the enzyme that catalyses the breakdown of lactose, the sugar in milk, to glucose and galactose. In humans and most mammals it is normally only expressed in infancy and early childhood. If milk is consumed with insufficient levels of the enzyme present, such as in non-persistent adults, unpleasant bloating, cramps and stomach upset occur (Hollox et al. 2001). However, individuals from many European, Middle Eastern and some African human populations continue to express the enzyme throughout adulthood, giving them the ability to digest milk and milk products (Mulcare et al. 2004). It has

been suggested (Cavalli-Sforza 1973, Hollox et al. 2001, Enattah et al. 2002, Poulter et al. 2003) that a selective advantage based on additional nutrition from dairy products explains these genetically determined population differences (see Enattah et al. 2002).

Sucrase-isomaltase (SI) is the enzyme responsible for sucrose digestion, as well as 80% of maltose digestion. Maltose is the main disaccharide produced by the digestion of starch by the salivary and pancreatic amylases and it is likely that the domestication of plants such as wheat led to an increase in starch in the diet of farmers compared with hunter-gatherers (Cavalli-Sforza 1981, Neel 1982, Turner 1979, Cordain et al. 2000*a*). Sucrose has probably only been consumed in large quantities in recent times. Congenital SI deficiency is a serious condition if large quantities of sucrose are consumed early in life. SI deficiency in the Greenland Inuit came to the attention of Danish researchers because it was associated with severe malnutrition when sucrose was fed to infants, and was shown to be common in this population. Sucrose-isomaltose malabsorption is found at 7-16% in Inuit populations contrasting with 0.2% in white North Americans (McNair et al. 1972). It is only recently have the Inuit been exposed to sucrose containing foods, as traditionally the majority of their dietary calories are from meat and fat (Shetty 2002). It has been suggested that in populations that consume little dietary starch and sucrose, the SI deficiency allele is not under selective constraints and so has reached polymorphic frequencies.

Type II diabetes (or non-insulin dependent diabetes mellitus, NIDDM) is another example where there is marked difference between certain populations in terms of disease incidence. In the last fifty years, a number of epidemics of type II diabetes have been noticed among populations that have recently adopted high calorie / low exercise western lifestyles, such as the Namura Indians of the tropical Pacific, Pima Indians in Arizona and urban Wanigela people in Papua New Guinea (Diamond 2003). Initially it was suggested that these populations had experienced extremely severe famines during their history and so selection for a thrifty genotype had occurred (Neel 1962). His hypothesis was that the rapid release of the hormone insulin in response to elevated blood-sugar levels was an advantage to our ancestors, allowing them to build up fat deposits in

times of plenty. However, in an environment where there is an abundance of food, this rapid response is detrimental – over production of insulin leads to insulin resistance, which in turn leads to diabetes. In 1982, Neel revised his theory and suggested that NIDDM was due to genetically based insulin resistance in response to hunter-gatherer diets low in carbohydrate and high in fat and meat. Such a diet would have selectively favoured the maintenance of blood glucose levels as the day-to-day body fuel, while synthesising and depositing fats as longer-term energy stores (McMichael 2002). Support for this idea came from the work of Lilloija and colleagues (1993) who studied individual differences in insulin secretion and insulin sensitivity the Pima Indians. They found that these differences are predictive of subsequent diabetes and act independently of obesity.

It is now considered that the state found in many of the Non- European populations studied is more likely to be the ancestral human condition (see Mc Michael 2002). The focus has now shifted to explaining why Europeans show such low incidence of type II diabetes considering their well-fed sedentary lifestyles. It is possible that selective constraints have been relaxed, or perhaps there has been selection in Europeans for increased insulin sensitivity since the establishment of agriculture and the resulting increase in carbohydrate in the diet.

One final example of inter-individual variation in a dietary enzyme is alcohol dehydrogenase (ADH). Many people are able to derive an appreciable proportion of their energy intake from alcohol (Roberts 1985). However 83% of Japanese have a variant of ADH that metabolises alcohol at a much higher rate (Osier 2002). This variant is only present at 6% in Europeans. If individuals who have the atypical form of ADH consume alcohol, acetaldehyde accumulation occurs causing the characteristic flushing syndrome consisting of nausea, tachycardia, dizziness, warmth and muscular weakness (Aebi et al. 1981). Osier (2002) found an ADH haplotype that was present at high frequencies in East Asians and rare or unobserved in other populations. They suggest that this haplotype is unlikely to have reached such high frequency because of random genetic drift alone (Osier 2002). The polymorphism could have attained such a high frequency through genetic drift or the effects of

selection. The authors argue that it would take a very strong population bottleneck and/or strong subsequent random genetic drift within eastern Asia for this rare haplotype to become frequent. However, more data are required to demonstrate that selection has indeed been operating on this ADH allele.

These examples illustrate the fact that there exist differences between populations in terms of the individual's ability to metabolise certain dietary components. Often, authors suggest adaptive scenarios to explain the frequencies of the variants found. However, observing an association between food practices and the ability (or inability) to digest certain foods, is not sufficient grounds to claim that natural selection has occurred to adapt us to our dietary environment. Rather, neutrality is the null hypothesis against which hypotheses of selection must be tested. There are a number of formal methods that have been developed to test for signatures of selection using genetic data, which will be outlined in the following section.

## 1.3 Natural Selection

In his seminal work Of the origin of species by means of natural selection Charles Darwin (1859) defines natural selection as the preservation of favourable individual differences and variations, and the destruction of those which are injurious. In the language of modern genetics this can be described as the differing reproductive success of genotypes in succeeding generations. Genotype variation produces individuals with different capacities to survive and reproduce in different environments (Hartl 1987). Natural selection acts on the phenotype of an individual, which may be determined by a number of genes as well as environmental factors. However natural selection can have no evolutionary effect unless phenotypic variation has a genetic component. That is, that some of the variation that created the phenotype must be able to be inherited by the next generation (Bamshad & Wooding 2003).

### *1.3.1 Genetic Variation*

In most populations of animals and plants inter-individual genetic variability can be found. In humans, variation at the protein level has been recognised since the 1960s. Less is known, however, about the functional significance of this variation. There are a number of types of molecular variation, which can be grouped into three classes: Single nucleotide polymorphisms (SNPs), insertion/ deletion polymorphisms (indels) and variable number of tandem repeat polymorphisms (VNTRs), which include mini and microsatellites. Many examples of these polymorphisms have no known function, but there are instances from all three classes that have phenotypic effects. Even if a polymorphism does not appear to have a function, it may still prove to be a useful marker for genetic analysis if it is closely linked to a locus of interest, as will be seen in the following sections.

*1.3.2 Types of selection*

The study of natural selection has led to a proliferation of terms to describe different ways that changes in allele frequencies occur (see Fig 1.2). Selection can operate at any stage in an organism's progress from fertilisation until the production of viable offspring, including survival into reproductive age, success in attracting a mate (sexual selection), ability to fertilise (gamete selection) and in the number of offspring produced.

In terms of allele frequencies, natural selection may lead to genetic uniformity or to diversity (Gale 1980). If selection favours phenotypes at one extreme of the range of variation present in the population then it is known as directional selection. Mutations that increase fitness are said to be subject to positive selection. In contrast, mutations that reduce the fitness of the individual will be subject to purifying selection (also known as negative selection). If a new deleterious

| Type of selection | Description | Leading to uniformity or diversity? | Directional or balanced? |
|---|---|---|---|
| Positive (diversifying)[a] | Mutations that increase fitness will be selected for | Uniformity | Directional |
| Negative (Purifying)[a] | Mutations that decrease fitness will be selected against | Uniformity | Directional |
| Codominant[a] | Mutations that reduce fitness of both heterozygote and homozygote will be selected against | Uniformity | Directional |
| Overdominant[a] (heterozygote advantage) | Mutations that increase fitness of heterozygote relative to both homozygotes | Diversity | Balanced |
| Frequency dependent[a] | Frequency of allele determines its fitness | Diversity | Balanced |
| Stabilising (normalising)[b] | Intermediate phenotype is advantageous | Diversity | Balanced |
| Underdominant (heterozygote disadvantage)[a] | Mutations that decrease fitness of heterozygote relative to both homozygotes | Diversity | Balanced |
| Disruptive[b] | Two or more phenotypes are fitter than intermediates between them | Diveristy | Balanced |
| Background[c] | The elimination of neutral polymorphisms as a result of negative selection of deleterious mutations at linked sites | Uniformity | Directional |

Fig 1.2: A summary of terms used in discussing natural selection. Notes: [a] See Jobling et al 2004, [b] See Futuyma 1998, [c] See Bamshad & Wooding 2003.

mutation arises that reduces the fitness of the heterozygote, as well as the homozygote, then it will be eliminated more rapidly from the population. This is known as co-dominant selection (see Futuyma 1998).

Alleles may increase the fitness of the heterozygote relative to both homozygotes. This is known as over-dominant selection (or heterozygote advantage) and this type of selection creates a balanced polymorphism. If an intermediate phenotype is fittest then stabilising (or normalising) selection is said to be operating. An example of heterozygote advantage in humans is sickle cell anaemia. The $Hb^S$ allele causes the debilitating sickle cell anaemia when homozygous, but also confers malarial resistance when heterozygous (Haldane 1949).

Balanced polymorphisms, as are maintained by over-dominant selection, can be generated by a number of processes, which are collectively described as balancing selection. Frequency dependent selection is an example of balancing selection. Here, the frequency of a genotype is inversely related to its fitness. The major histocompatibility complex (MHC) is suggested to have been under both frequency dependent and over-dominant selection (Hughes & Nei 1988). If pathogens have evolved to evade immune detection in individuals carrying the higher frequency alleles, frequency dependent selection will cause low frequency alleles to be favoured. In the case of heterozygote advantage, individuals with heterozygous MHC are better able to resist infectious disease as a result of having a broader spectrum of antigen binding specificities.

Another type of selection, under-dominant selection (also called heterozygote disadvantage), reduces the fitness of only the heterozygotes. This is an example of disruptive selection, where two or more of the phenotypes are fitter than the intermediates between them (Futuyma 1998).

*1.3.3 The neutral theory of molecular evolution*

Before the 1960s it was assumed that most of the polymorphisms in a populations were maintained by balancing selection (Bamshad & Wooding 2003). During the 1960s however, protein sequencing and electrophoresis of allozymes started to provide data on the extensive amount of amino acid polymorphisms both within and between species. Mooto Kimura (1968) estimated the rate of amino acid substitution in $\alpha$ and $\beta$ haemoglobin sequences. He argued that the genetic load, or the proportion of a population's maximum fitness that would be lost as a result of selection against the deleterious genotypes is contains, would be too great if selection was that only driving force in protein evolution (Kimura 1968). Instead he proposed that most polymorphisms, and fixed differences between species, are selectively neutral. This idea is known as the neutral theory of molecular evolution. Kimura's neutral theory has provided the framework for evolutionary analysis of DNA sequence variation and change since the 1960s. Selective neutrality is an appropriate null hypothesis against which to test for evidence of selection (Kreitman 2000). Selection tests that compare observed diversity with that expected under neutral evolution and are known as neutrality tests (see Wayne & Simonsen 1998 for a review).

*1.3.4 The confounding effect of human demographic history*

It should be noted here that a significant difference from the neutral expectation might not always be the result of selection. The neutral model assumes that the population is in a mutation – drift equilibrium, which is the case in a large constant sized population. However in humans it is known that the species population size has expanded dramatically in the past 10,000 – 100,000 years, from a few thousands of individuals to over 6 billion (Yang 2002). The human population is not, therefore at a stationary equilibrium for neutral variants. This can cause problems for testing for evidence of selection, as some genetic signatures of positive selection can be similar to signatures of population's expansion (Kreitman 2000). In addition the human species has a history of major migrations and population subdivision, which can give rise to patterns of

variation that depart from the neutral expectation under simple models of evolution.

According to Kreitman (2000), the only current safeguard against gross misinterpretation of test results in terms of distinguishing between selection and historical demography, is to have an a priori hypothesis about the type and direction of selection that is expected for the locus under investigation. Population history, however, affects all nuclear genes equally, where as signatures of selection should only be detectable at the particular locus of interest (Payseur & Nachman 2002, Bamshad & Woodman 2003). This idea, which has formed the basis of tests for neutrality for decades (see Lewontin & Krakauer 1973, Cavalli-Sforza 1966) has recently been given a new lease of life. The availability of large genome-wide data sets consisting of thousands of SNPs and microsatellite markers from the human genome, typed in a number of global human populations, has opened up the possibilities for identifying regions of the genome that have been influenced by local natural selection (see Akey et al. 2002, Kayser et al. 2003). Following on from this, departures from neutrality can be detected in loci that have been hypothesised to have been under recent local selection, by means of a simple comparison of frequency distribution from the candidate locus with the genome-wide pattern estimated from large numbers of markers that have been typed in the same individuals or populations.

### *1.3.5 Testing for selection*

There are many formal methods for formally testing for the signature of past selection. As can be seen in Fig 1.3, the different methods available are appropriate for different types of data. In addition the different approaches have different abilities to detect different modes of selection, such as directional, balancing etc. When considering dietary adaptations that have occurred in different human populations, we are concerned with looking for evidence for selection using within-species polymorphisms. Some tests apply equally well to between-species and within-species comparisons whereas some tests require both types of data. In addition, some tests focus specifically on within-species

| Test | Type of data required | Designed to detect | Best Use | Reference |
|---|---|---|---|---|
| Tajima's "D" | Within sp. | Skew in frequency spectrum | General purpose test of frequency spectrum skew | Tajima 1989 |
| Fu & Li's "D" | Within sp. | Recent vs ancient mutations | General purpose test of frequency spectrum skew | Fu & Li 1993 |
| Fu "W" | Within sp. | Departures in frequency spectrum | Population subdivision | Fu 1996 |
| Fu "Gη" | Within sp. | Departures in frequency spectrum | Population subdivision, shrinkage & overdominance | Fu 1996 |
| Fu "Gξ" | Within sp. | Departures in frequency spectrum | Population subdivision, shrinkage & overdominance | Fu 1996 |
| Fu "$F_s$" | Within sp. | Excess or rare alleles (one sided) | Pop growth, hitchhiking, background selection | Fu 1997 |
| Hudson | Within sp. and allele | Unexpectedly low variation in allele class | Directional selection | Hudson et al 1994 |
| Wall B ans Q | Within sp. | Linkage disequil. between adjacent segregating sites | Pop. Subdivision, balancing selection | Wall 1999 |
| Andolfatto's "$S_k$" | Within sp. (sliding window) | Non-neutral haplotype structure | Balancing and directional pop. subdivision | Andolfatto et al 1999 |
| HKA | Within vs between sp. (2 loci) | Differences in variation not accountable by constraints | Balancing selection, recent selection sweeps | Hudson et al 1987 |
| McDonald (run test) | Within vs between sp. (Contigous region) | Regions with non-neutral patterns of polymorph and diversity | Eqm. balancing selection | McDonald 1996, 1998 |
| McDonald Kreitman"G" | Within vs between sp. (synon. vs nonsynon.) | Adaptive evolution | Adaptive protein evolution; mutation / selection | McDonald & Kreitman 1991 |
| Intra-allelic variabitilty | Within sp. | Discrepancies between allele frequency and variability at linked loci | Directional selection | Slatikin & Bertorelle (2001) |

Fig 1.3 A summary of the different tests of selection (adapted from Kreitman 2000). Sp. is an abbreviation for species.

polymorphisms. The following review will focus on tests that can be applied to data from within-species comparisons only.

*1.3.5.1 Codon-based selection tests*

Nucleotides within coding sequence can be divided into synonymous mutations which do not result in a change in amino acids, and non-synonymous mutations, which result in that codon specifying a different amino acid. Synonymous sites are assumed to be selectively neutral. Under diversifying selection, the proportion of non-synonomous sites ($d_n$) that are variable will be greater than synonymous sites ($d_s$). Under purifying selection the reverse is true. This can be expressed as follows:

$$\omega = d_n / d_s$$

If $\omega > 1$, this indicates that positive selection has occurred, if $\omega = 1$ then no selection has occurred and if $\omega < 1$ then there is evidence that purifying selection has been operating. By testing if $\omega$ is significantly different from 1, $d_n$ is tested against the neutral expectation $d_s$ (see Yang & Bielawski 2000 and Yang 2001 for a review). Rooney & Zhang (1999) examined the ratio of synonomous to non-synonomous substitutions in the protoamine gene in primates. Protoamines are proteins that bind sperm DNA during spermatogenesis in vertebrates. They found that the nucleotide substitution rate at non-synonomous sites is significantly higher than the rate at synonomous and intron sites in protoamine P1 of hominoids and Old World Monkeys. This result suggests that positive selection has been operating on protoamine P1 in these species.

*1.3.5.2 Tests based on the frequencies of variant sites*

Under the neutral model in a constant sized population, the level of diversity in a population is assumed to have reached equilibrium, where the generation of new alleles by mutation is equal to the elimination of alleles by genetic drift. Therefore it is possible to define an expected level of diversity in a population ($\theta$) in terms of the mutation rate ($\mu$ per site per generation) and drift. Since drift is inversely proportional to the effective population size ($N_e$), or the size of the ideal population in which the effects of drift would be the same at those seen in

the actual population, then this is used as a proxy for drift. This can be expressed as follows:

$$\theta = 2n\ N_e\ \mu$$

Where n is the number of heritable copies of the locus per individual.
There are a number of different ways of estimating θ from sequence data, which take into account different parameters derived from the observed diversity. Under the neutral model the various different methods of estimating θ should give the same result. This forms the basis of a number of different tests for selection (see Fig 1.3) (see Wayne & Simonsen 1998 & Kreitman 2000 for a review). The best known of these is Tajima's D (Tajima 1989). This compares two estimates of θ, one based on the number of segregating sites, and the other based on the number pair-wise differences. Under neutrality Tajima's D is expected to be zero. Negative values indicate positive selection, where as positive values indicate balancing selection.

FOXP2 is a transcription factor involved in speech and language development. J. Zhang et al. (2002) found that the level of polymorphism in the introns of FOXP2 is lower than other neutral non-coding regions they examined. They obtained a Tajima's D value of –1.36 for the FOXP2 intron data, which indicates positive selection. Events such as background selection on deleterious mutations in tightly linked exons or quick fixation of advantageous mutation in these exons (also known as a selective sweep) could have lead to a reduced present-day polymorphism in the introns of FOXP2.

*1.3.5.3 Intra-allelic variability*

Intra-allelic variability is the joint distribution of the frequency of alleles and the extent of variability at closely linked marker loci. Slatkin & Bertorelle (2001) developed a method for finding this joint distribution of allele frequencies and diversity. If the population growth rate is known, then the joint distribution provides the basis of a test for neutrality by testing whether the observed level of intra-allelic variability is consistent with the observed allele frequency (Slatkin 2001). Under the neutral model the frequency of an allele will be related to its age. This is because it takes a long time for rare (including new) alleles to drift

to high frequencies in populations. If an allele is young but at high frequency in a population, then this reflects a departure from neutrality.

Intra-allelic variability can be modelled in a number of ways: i) number of recombinants at a linked binary marker, ii) the length of a conserved haplotype and iii) the number of mutations at linked markers (including microsatellite markers).

Slatkin & Bertorelle (2001) demonstrate an intra-allelic variability model using allele frequency data from the CCR5 locus, and two closely linked microsatellites from data presented by Stephens et al. (1998). The Δ32 deletion at the CCR5 gene causes an absence of the CCR5 chemokine receptor on lymphoid cells. This receptor serves and an entry port for a number of pathogens including the human immunodeficiency virus (HIV)-1. The absence of the receptor, as caused by the Δ32 deletion, is associated with a strong resistance against HIV infection and AIDS. The frequency of the Δ32 deletion allele is more than 10% in European populations, although diversity at the closely linked microsatellites suggests that it is quite young. Slatkin & Bertorelle (2001) provide evidence with their analysis that the Δ32 deletion allele is not neutral.

Sabeti et al. (2002) developed a framework for detecting recent positive selection in humans by analysing the conservation of long-range haplotypes. They examined two loci: Gluscose-6-phosphate dehydrogenase (G6PD) and the CD40 ligand (TNFSF5). Both these genes have alleles that are thought to provide protection against malaria, and so it has been suggested that these alleles may well be under positive selection. Sabeti et al. (2002) identified the local haplotypes of these alleles and then examined extended haplotypes up to 500 kilobases either side of the genes. Under neutral evolution new variants will require a long time to reach high frequency in the population and the linkage disequilibrium (LD) around them will decay over time, due to recombination. Therefore common alleles will typically be old and have only short range LD. A signature of positive selection is that an allele will have unusually long-range LD

given its populations frequency. Sabeti et al. (2002) found evidence of positive selection at both loci they examined using this method. Interestingly they also explored whether positive selection could have been detected using traditional methods. They performed Tajima's D, Fu & Li's D, Fay and Wu's H. The $d_s/d_n$ test, the McDonald Kreitman test and the HKA test. None of these tests showed a significant deviation from neutrality for either G6PD or TNFSF5. This indicates that intra-allelic variability approaches are more sensitive than traditional methods.

The method developed by Sabeti et al. (2002) was used by Bersaglieri and colleagues (2004) to investigate the signatures of selection at the lactase gene. As mentioned earlier (see section 1.2) the ability to digest lactose contained in milk usually disappears in childhood but in European populations, lactase activity frequently persists into adulthood. It has been suggested that this ability to digest milk in adulthood is an adaptation to pastoralism. Individuals who could digest lactose as adults would have an additional source of nutrition in the form of dairy products from their herds of cows, and would therefore have a selective advantage over individuals who could not. However, this intriguing theory has only recently been provided with formal population-genetics evidence for selection. Bersaglieri et al. (2004) typed 101 SNPs covering 3.2Mb around the lactase gene. They showed that in northern Europeans, a common (~77%) haplotype is unusually long given it high frequency. They estimated that strong selection occurred with the past 5,000-10,000 years, consistent with an advantage to lactase persistence in the setting of dairy farming. In addition they remark that the signals of selection they observed are among the strongest yet seen for any gene in the genome.

Humans have traditionally represented a far from ideal species on which population geneticists can test their models. Not only do humans have low levels of nucleotide polymorphism, the inability to control matings, overlapping generations, geographic subdivision, but also there are also ethical issues with sampling from native populations (Kreitman 2000). However, the more sensitive haplotype-based methods for identifying selection, increasing amounts of data on neutral loci and a greater understanding of the effect on variation of human

demographic history, should all increase our ability to detect the signature of natural selection in the human genome.

## 1.4 A brief chronology of dietary change in human evolution

Gordon (1987) has recognised three general phases in the evolution of hominid subsistence behaviour. The first phase involves a shift from unprocessed primarily vegetarian foods eaten by the common ancestor of the hominid line and chimpanzees, to a diet with a significant proportion of meat as well as substantial non-oral food processing. The second phase is characterised by the development of specialised hunting and gathering strategies, and the final phase marks the transition to food production and the domestication of plant and animal species.

### *1.4.1 The diets of Pleistocene hominids*

The earliest hominids likely derived most of their food from plants (Wrangham et al. 1999). Eaton et al. (1988) suggested that plant foods composed >90% of australopithecine diets. Stone tools appear in the archaeological record around 2.5 million years ago (Toth 1985, Susman 1994) and around this time there is evidence from that meat was beginning to become an increasingly important component to the human diet (Aiello & Wheeler 1995). There has been much debate about how the early hominids obtained their meat – whether through hunting or scavenging from the meals of other carnivores (see Chase 1989). Deliberate hunting behaviour has often been thought to be associated with the development of group co-operation and communication (see Binford 1985), and has therefore traditionally been an area of archaeology that has generated much interest. It must however be noted here that groups of chimpanzees have been observed in the wild to successfully hunt and eat small mammals including monkeys in organised groups (Boesch & Boesch 1989, Stanford et al. 1994).
It has been suggested that the robust australopithecines (*Paranthropus sp*.), often interpreted as an evolutionary offshoot of the hominid family tree (see Fig 1.4), were

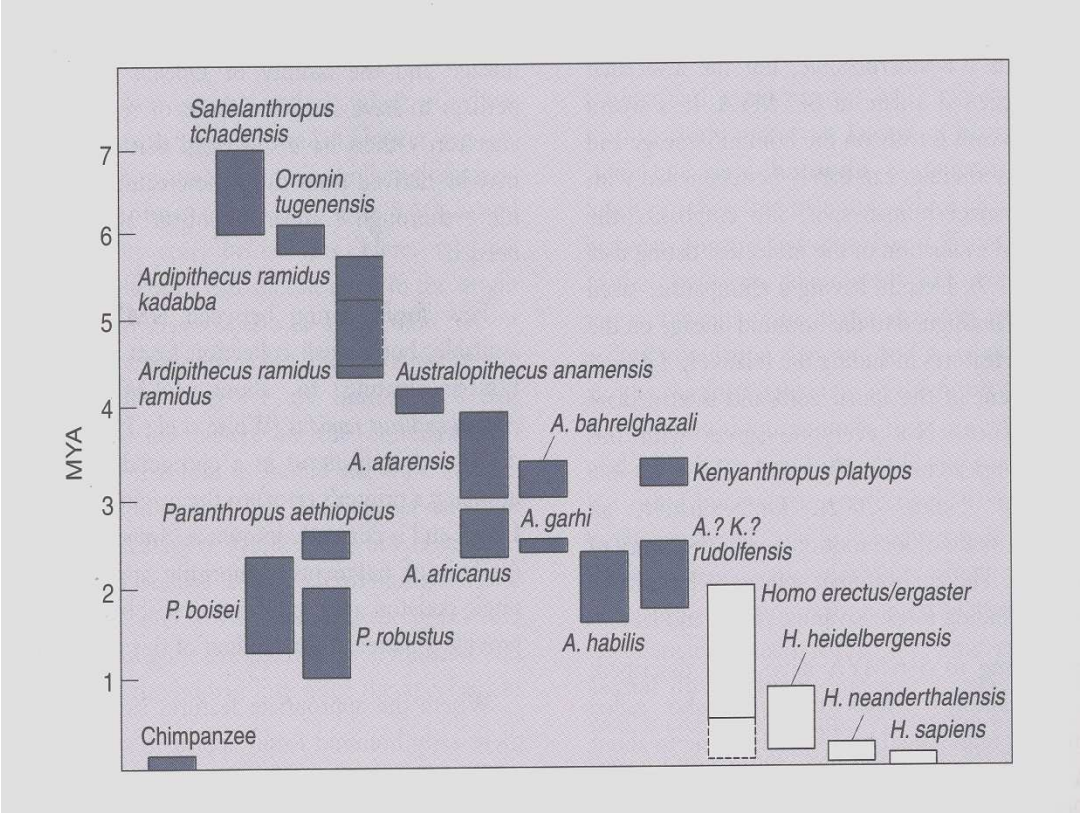



Fig 1.4: Fossil hominids, after Jobling et al (2004). The time span of each species indicates either the uncertainty in dating or the times of the earliest and latest fossils, whichever is larger. Dotted lines indicate either a lack of intermediate species *Ardipithecus ramidus* fossils, or particular uncertainty about the later dates for *Homo erectus*. Dark blue: found only in Africa. White: found in Africa and elsewhere, or only outside Africa. Many aspects of the classification of these fossil are still debated and likely to be revised.

more vegetarian than early *Homo.* However analysis of strontium calcium ratios and strontium isotopic ratios suggest that the robust australopithecines from the site of Swartkrans, South Africa were omnivores (Sillen & Hall 1994). The massive jaws and teeth of the robust australopithecines have been interpreted as an adaptation to a substantial proportion of the diet stemming from very tough food, perhaps seeds or nuts (see Toth & Shick 1986). The identity of these foods however remains unknown.

The earliest use of fire in the archaeological record is still the subject of much debate (see James 1989 for a review). A number of claims have been made for the use of fire as far back as 1.7 mya. However most archaeologists accept that *Homo erectus* was using fire approx 0.5 mya. Fire is necessary for cooking, which makes food more available digestible by cracking open skins and husks, bursting cells, breaking down complex molecules and denaturing toxins. Wrangham et al. (1999) suggest that signals of cooking can be detected in the fossil record from 1.9mya, in the form of the smaller teeth (reduced digestive effort) and larger female body size (increased supply of food energy) of *Homo erectus.*

Around 700,000 years ago there is indisputable evidence in the archaeological record for hunting (Chase 1989). Brain size was comparable to that of modern humans, and stone tool technology shows substantial advances and diversification from previous tool making traditions. During this time there was also a shift toward larger prey species. Faunal assemblages from the Upper Palaeolithic period at around 40-11,000 years ago in Europe, contain remains from enormous numbers of larger herbivore species such as reindeer, woolly mammoth, bison, and horse (Olsen 1988). Marine resources (shellfish, fish, marine mammals &birds) began to be exploited during this phase (Richards 1999). Plant foods are almost invisible in the archaeological record, although the pollen record for this period shows the presence of nut tree species such as walnut, hazelnut and pine nuts (Dumayne-Preaty 2001). There is also evidence that pre-agricultural populations collected wild grasses and grains in Western Asia and in North East Africa (Harlan 1989).

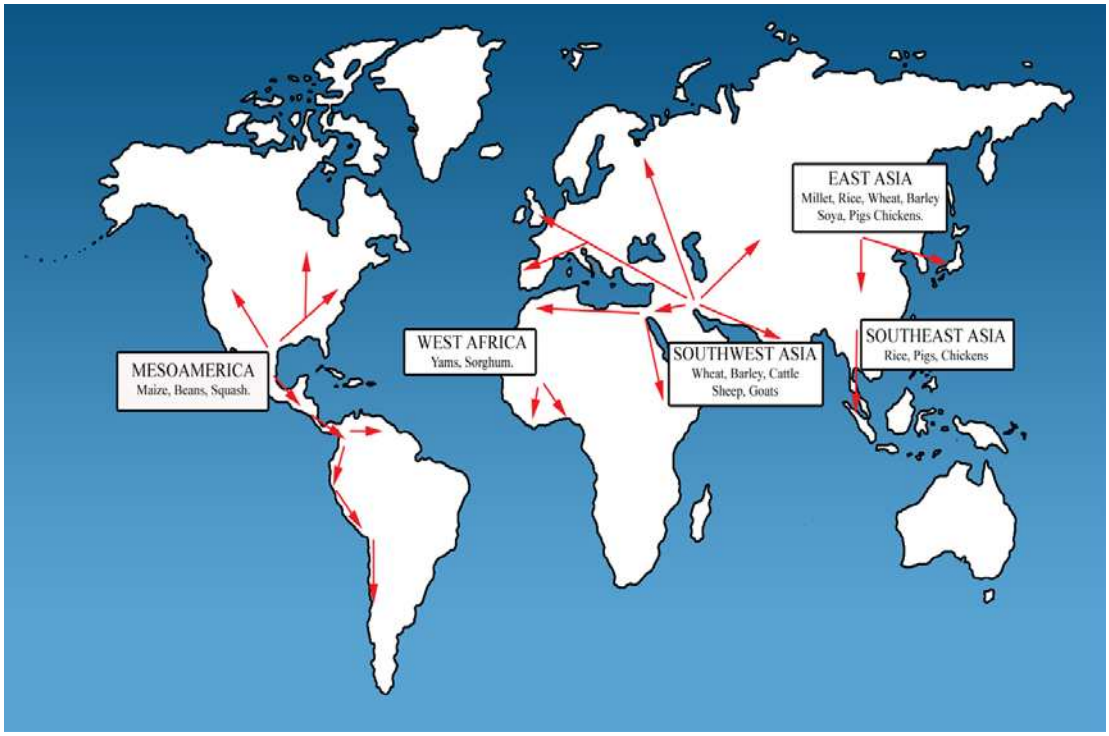



Fig 1.5: The origins and spread of agriculture

### *1.4.2 The origins and spread of agriculture*

The final phase in the history of human diet and subsistence began at the end of the Pleistocene and the beginning of the Holocene, at around 12-10,000 years ago. The most fundamental change was the shift from food collection to food production. The beginnings of agriculture were made possible by the warming after the last glacial maximum (c. 18,000 years), which gave rise to the beginnings of the modern patterns of climate, vegetation and fauna (Harris 1996). The domestication of both plant and animal species, and the development of farming economies, is thought to have happened independently at a number of locations around the world (See fig 1.5).

#### *1.4.2.1 The fertile crescent*

The warmer climate at the end of the last glacial period favoured the growth of rich stands of wild cereals such as wheat and barley in an area in Western Asia known as the Fertile Crescent (see fig 1.6). It is likely that the domestication of certain plant and animal species began by sedentary foragers, who gradually developed more intensive techniques of plant exploitation, including storage and possibly small scale cultivation (Harris 1989). An example of this mode of living are the settlements from the Natufian period (12500-10000 ybp) in the Levant, where stone houses in small villages have been found, along with sickle blades and mortars for grinding seeds (Hillman 1989). The period of cold and dry conditions, which occurred between 11500-10,600 ybp, would have reduced the wild plant food resources. This in turn would have increased the dependence on the small-scale cultivation of large-seeded grasses and herbaceous legumes.

Cultivation caused substantial genetic changes in the wild cereals. For example, deliberate sowing and the use of sickles may have favoured the retention of non-shattering variants (Bar-Yosef & Kislev 1989). Fully domesticated specimens of emmer wheat as well as other crops have been found at sites all around the Fertile

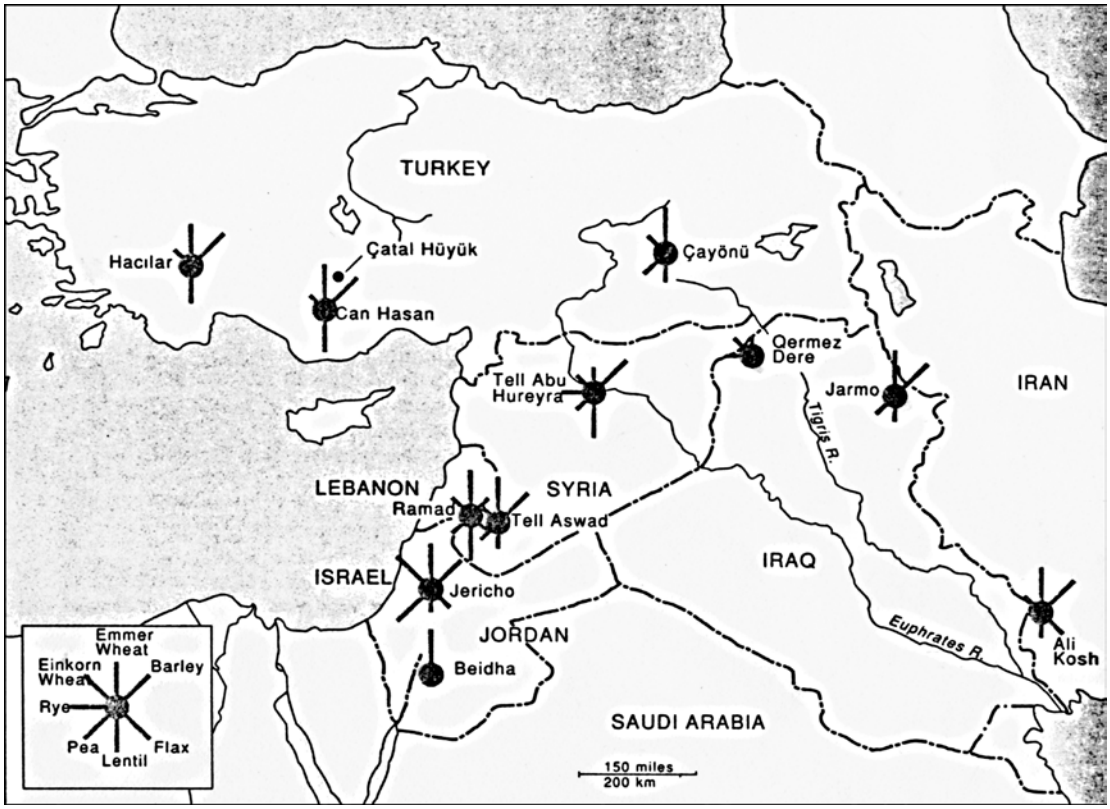



Fig 1.6 The fertile crescent, showing major Neolithic archaeological sites (after Renfrew & Bahn 1997)

Crescent dating from around 9,500 ybp (Zohary 1989) (see fig 1.6). Also in this period, there is evidence for the domestication of animals, such as sheep from the site of Cayonu in South East Turkey, goats in Shanidar, Iraq and cattle at the site of Bourqras in Syria (Harris 1981). By 8000 ybp agriculture was being developed and extended, with new tools such as hoes, larger trade networks, irrigation, and town sized settlements such as Catal Huyuk in Turkey. These developments opened up the possibility of colonising new territories for cultivation and as a result, farming expanded from the fertile crescent east to Iran and west to Europe and Egypt (Miller & Wetterstrom 2000).

*1.4.2.2 Europe*

Farming expanded into Europe from the south-eastern lobe of the Fertile Crescent (Harris 1996). The farming economy comprised of wheat and barley as well as sheep, goats, cattle and pigs. Farming spread north to the Balkans and from there in two directions; further northwards to along the Danube and Rhine towards the north sea, as well as westwards along the Mediterranean to Italy, Southern France and Eastern Spain. Lastly, the British Isles and Scandinavia show evidence of farming, although most of Scandinavia was too cold for growing cereals until some time after the initial farming expansion (Cavalli-Sforza et al.1994).

One of the mostly hotly debated topics in European Neolithic archaeology is whether farming spread across the continent as a result of the movement of people or through cultural transmission (Ammerman & Cavalli-Sforza 1973). It is likely that the initial arrival of agriculture in Europe was brought about by pioneering faming groups (Sherrat 1994). They brought with them a range of crops, technology and new styles of living that were very different to the hunter-gatherers that populated the continent. After 6,500 ybp, in the west of the continent it is likely that features of the farming economy were adopted by the indigenous inhabitants (*ibid*).

There is still much interest, however, in the extent to which the modern European gene pool is contributed to by Neolithic farmers from the near east, and the European Palaeolithic populations. Early work using data from classical

markers (blood groups etc) suggested that Europe-wide gradients of allele frequencies were a result of the admixture between low density local hunter-gatherers and large numbers of new-coming farmers from the Near East (Ammerman & Cavalli-Sforza 1973) Richards et al. (2000*a*) examined maternally inherited mitochondrial DNA (mtDNA) from a range of modern European and Near-Eastern populations to shed light on the colonisation of Europe. They estimated that the immigrant Neolithic component comprises less than a quarter of the mtDNA pool of modern Europeans. This estimate contrasts with the data from classical markers and suggests that the influx of farmers from the Near East was much smaller than previously assumed. Richards et al. (2002) also found evidence for substantial back-migration into the middle East, and that the majority of extant mtDNA lineages in Europe entered in several waves during the Upper Palaeolithic.

Semino et al. (2000) sampled Y chromosomes from males and identified two Palaeolithic and one Neolithic migratory episode that contributed to the modern European male gene pool. From a subset of their data they estimated the contribution of Near-Eastern farmers to the European gene pool to be approximately 22%. This estimate agrees with the data from the mtDNA. However, Chikhi et al. (2002) analysed the entire Y chromosome dataset of Semino et al. (2000) using a likelihood based method to estimate the change from place to place in Europe of admixture proportions of Neolithic and Palaeolithic genes. They found an average Neolithic contribution of 50% across all samples, 56% for the Mediterranean subset and 44% in non-Mediterranean samples. These estimates of a large Neolithic contribution are in agreement with the data from classical markers and at least twice as large as the estimate of Semino et al. (2000) as well as the estimates from mtDNA. It is important to note, however, that the estimates for the average Near Eastern contribution to the European gene pool do not represent the relative proportions of farmers and hunter-gatherers during the initial formation of settlements, but rather the proportion of genes that can be traced back to ancestors in the Near East.

*1.4.2.3 East Asia*

In the Yellow river area of present day China, millet was domesticated and became a major crop by around 8500 ybp (Sabban 2000). There is also evidence that the pig and dog were domesticated about this time in the same region. Further south in the Yangtze area, rice was cultivated and is present in large amounts in the archaeological record from around 7000 ybp. Rice and pigs also formed the basis of the Asian South Coastal cultures, which developed around 6000 ybp. By 3000 ybp, important techniques such as crop rotation, irrigation and the use of organic fertilizers had been developed (Sabban 2000)

In the steppes of Central Asia, pastoral nomadism started as a secondary development of the farming economy. The steppe was a difficult environment for agriculture but the open grasslands enabled animal husbandry. Goat, sheep and cattle remains are found from 6000 ybp. There is evidence for the use of horse and camel from 5000 ybp, which helped to bring about the dominance of pastoralism over agriculture. Pure pastoral nomadism with no agriculture is rare (Cavalli-Sforza et al. 1994). Most societies practiced semi-nomadic pastoralism where agriculture supplements the diet and is usually practiced for some of the year, or by specific groups within the society, such as women. Many nomads also engage in trade with their agricultural neighbours (Morgan 1990).

*1.4.2.4 Africa*

Africa is home to the earliest evidence for wild-grass collection in the world (Harlan 1989). The site of Wadi Kubbaniya, in modern day Egypt, which has been dated to 17,000-16,000 years, has charred remains of the tubers from a variety of wild grass species (Hillman 1989). Wild grains were collected for food from 13,000 years ago by the peoples of the Cataract tradition, who inhabited a region that stretched from the Nile, east to the Red sea and south to the Ethiopian highlands (Ehret 2002). With the dramatic changes in climate towards a wetter and warmer phase at the end

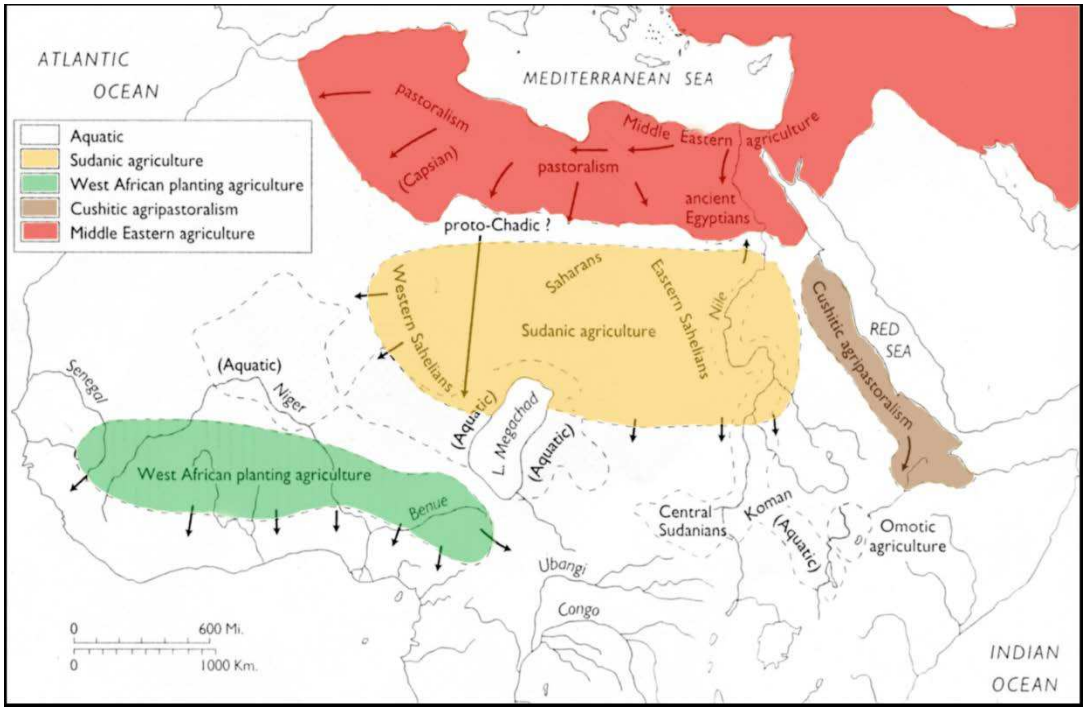



Fig 1.7 The centres and spread of African agriculture c 7500 ybp, according to Ehret (2000).

of the last glacial age, a number of local developments of farming took place in Africa (see fig 1.7). Between 9500-7000 ybp sorghum and millet were cultivated in the South-Eastern Sahara region. Meanwhile, in the woodland savannas of West Africa planting agriculture was developed with yam as the most important crop around 8000-5000 years ago (Fage 1988). The Bantu-speaking peoples who originated from the area of modern day Nigeria and Cameroon spread westwards and southwards through Africa (Curtain et al. 1995). The Bantu people brought with them iron-smelting technology, crops (such as millet and squash) and farming techniques which enabled them to displace and marginalise the hunter-gatherers that they encountered (Needham et al. 1984)

*1.4.2.5 The Americas*

The earliest evidence of cultivation in the Americas comes from the Guila Naquitez in central Mexico (Smith 1995). Excavations there have shown evidence of small scale cultivation of both squash and beans and dates to approx 9000 ybp. The earliest known maize cobs date to about 4700 ybp and were found in the Tehuacan valley in southern Mexico (Smith 1995). However maize pollen has been found in Oaxaca and dated to 7100 ybp (Pope et al. 2001). The wild ancestor of maize is thought to be Teosinte, a wild grass that grows over much of Central America. From there cultivation techniques and crops, such as maize and beans, spread northwards to the North American South West region, where they formed the basis of the Pueblo Indian cultures from 3500 ybp, and south through central America (Smith 1995).

In the Andean highlands, there is evidence for cultivation of potatoes, beans, quinoa and maize that dates to approx 4,500 ybp. There is also evidence for the domestication of the llama from around this time. By 2800 ybp agriculture has spread to the coastal region where there were large scale irrigation schemes and intensive cultivation (Smith 1995).

*1.4.3 Consequences of the development of agriculture for health and nutrition*

The advent of agriculture and the domestication of crops such as wheat, barley, rice and maize brought about a major change in the diets of the earliest farmers. Pre-agricultural hunter-gatherer diets were typically low in carbohydrate but high in fat and protein (Neel 1982, Cordain et al. 2000*b*). In contrast with this, the adoption of agriculture brought about a diet that is low in fat and protein but high in carbohydrate (Cavalli-Sforza 1981, Turner 1979). Analysis of ratios of strontium to calcium in bone from the Levant indicates that between 15,000 and 10,000 years ybp there is a marked increase in plant food consumption at this time (Schoeniger 1982). Plant foods are considered to be lower quality components of the diet than animal foods as they have a lower calorific value and lower concentrations of many essential nutrients such as vitamin B12, vitamin D, calcium and iron, as well as essential amino and fatty acids (Sullivan 1998)

Although the vast majority of essential nutrients can be obtained from different plant species, a wide variety of plants must be included in the diet so that all required nutrients are consumed in sufficient quantities. Archaeological evidence as well as observations of living peasant cultivators indicates that diets of agriculturalists tend to be dominated by a single staple: rice in Asia, wheat in western Asia and Europe, millet or sorghum in Africa and maize in the new world (Larsen 2000). (See Fig 1.8) An over reliance on a single staple, rather than consuming the broad spectrum of plant species eaten by hunter-gatherers, can lead to dietary deficiencies and malnutrition (Cassidy 1980).

Analysis of skeletal and dental material from early agricultural populations demonstrates the consequences of the major changes in diet and subsistence brought about by the development of farming. Turner (1979) analysed 64 archaeological and living populations from around the world and found that hunter-gatherers exhibited 1.7% carious teeth where as agriculturalists exhibited 8.6% carious teeth. Dental

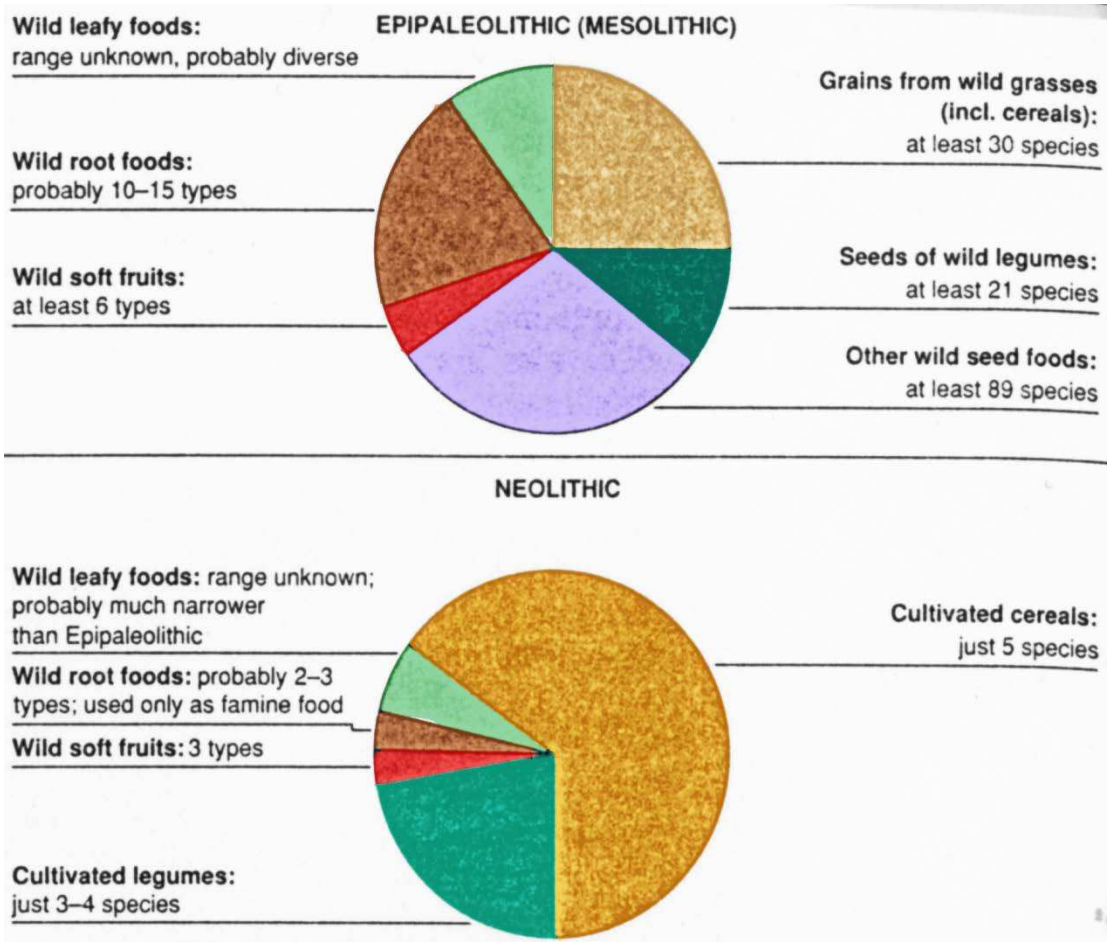



Fig 1.8 Pie charts comparing Epipaleolithic and Neolithic plant remains from Tell Abu Hureyra (after Renfrew & Bahn 1997). The charts illustrate the dramatic increase in grains and cereals, and reduction in wild roots and leaves during the Neolithic period at this site.

caries is a disease involving demineralisation of dental tissue and is caused by the fermentation of dietary carbohydrate by oral bacteria (Larsen 2000). Signs of chronic malnutrition in skeletal material as a result of lower overall quality of nutrition, such as shorted adult stature and reduced skeletal robusticity, can be seen as well as a marked increase in the incidence of infection. The signature of episodic stress and reduced life expectancy during the transition to agriculture is also present (see Ulijaszek & Strickland 1993 for review, Cassidy 1980). Despite a flourishing of material culture, general growth of human populations and the beginnings of civilisation that were eventually brought about as a consequence of the development of agriculture, patterns of poor nutrition can clearly be seen in the archaeological record.

### *1.4.4 Methods used in the reconstruction of Pre-agricultural Diets*

There have been a number of different methods used in reconstructing the diets of pre-agricultural humans. These attempts rely on two main approaches – examining the archaeological and fossil record for clues to the food of past populations, and studying the diet of modern hunter-gatherers as a proxy for Palaeolithic diets.

#### *1.4.4.1 Archaeological approaches*

Archaeologists can gain insight into the diets of our ancestors from both the analysis of human remains as well as examination of human habitation sites. Analysis of stomach contents (see Holden 2001) and coprolites (see Poinar et al. 2001) is perhaps the most direct method of collecting data on single meals that were eaten. Trace element and stable isotope analysis of skeletal remains can give a picture of the individual's diet over their lifetime in terms of the relative amounts of animal, marine and plant resources consumed (see Richards et al. 2000*b*). Carbon (13C/12C) isotope analysis can also distinguish consumers of plants with different photosynthetic pathways; C4 plants (tropical grasses such as maize, sugarcane, sorghum and millet) as opposed to C3 plants (most leafy plants as well as wheat, barley, oats and rye) (Sealey 2001). A study by Richards et al. (2003) uses carbon isotope analysis to investigate the dietary habits of the inhabitants of the British Isles during the Neolithic period. They found that before the introduction of domesticated plants and animals into

Britain, there was a moderate to large proportion of marine foods in the diets of Mesolithic Britons. However, during the Neolithic period diets are largely based on terrestrial species.

Inferences about the types of food eaten can also be made from micro and macroscopic analysis of teeth as well as the number of dental caries (Cassidy 1980, Larsen 1998). Assessments of an individual's nutritional status over their lifetime can also be made from skeletal and dental remains by examining the growth and development of the individual as well looking for any signs of nutritional deficiency diseases (see Yesner 1980, Brothwell 1969).

More indirect methods of reconstructing the diets of past people include analysing faunal assemblages at butchery sites, as well as looking at the remains of hearths, middens, and storage pits. Material culture can also give us insight into the food behaviours of archaeological populations such as assessing the tool marks on animal bones for evidence of scavenging versus hunting behaviours, micro-wear analysis of tools and presence or absence of specific food preparation equipment e.g. mill stones. It should be noted here that plants are relatively poorly preserved in the archaeological record compared to animal remains, so our knowledge the role of Palaeolithic plant foods is virtually non-existent (Larsen 2000). However, the human impact on vegetation, as seen through the pollen record can give indications of woodland clearance for agriculture (Dumayne-Peaty 2001).

Despite these varied approaches, knowledge of the diets of our ancestors through human evolution and prehistory remains limited. Whilst we can gather information on some elements of archaic feeding behaviour, it is clear that the details of the broad range of (especially plant) species and the relative contribution of those foods to the Palaeolithic diet are still unknown.

*1.4.4.2 Dietary reconstruction using modern populations*

Humans have survived through hunting, fishing and gathering wild food resources for the large majority (99.6%) of the 2 million years of their existence (Harris 1981, Sebastien et al. 2002). It is only in the last 10,000 years that some humans have domesticated plant and animal species for intensive exploitation for food. However, many anthropologists are of the opinion that few if any hunter-gatherer societies exist today that have not had at least some contact with agricultural societies (Cordain et al. 2000*b*). It is necessary to ask how good an analogue the diets of present day hunter-gatherer societies are for our ancestral diet, considering many have been marginalised to environments that are impossible to cultivate, such as deserts and polar regions. Despite these difficulties there have been a number of studies that have attempted to estimate the relative contributions of plant versus animal foods as well as the macronutrient composition of Palaeolithic diets.

Eaton & Konner (1985) used Lee's (1968) analysis of the Ethnographic Atlas (Murdock 1967), a collection of ethnographic data on 862 of the world's societies, and estimated that hunter-gatherer societies have an average of 65% calorific intake from plant foods and 35% derived from animal sources. From this average ratio, Eaton et al. (1997) estimated that the average dietary macronutrient composition of Palaeolithic humans was 22% fat, 37%protein and 41% carbohydrate. Since Eaton & Konner's (1985) original estimate of 65:35 ratio of plant: animal energy intake, many researchers now argue that the average hunter-gatherer subsistence pattern would have included much higher amounts of animal food (45-60%) (Cordain et al. 2000*b*). Cordain et al. (2000*b*) analysed the 229 hunter-gatherer societies in the ethnographic atlas and found that 73% of societies derived >50% of their subsistence from animal sources. The result of this is that the contribution of protein to overall energy intake is elevated at the expense of carbohydrates.

| Study | Data Source | Plant to animal ratio | Estimate of dietary macronutrients | | |
|---|---|---|---|---|---|
| | | | Protein | Carbohydrate | Fat |
| Eaton & Konner (1985) | 862 societies from Ethnographic Atlas | 65:35 | 37% | 41% | 22% |
| Cordain et al. (2000)*b* | 229 societies from Ethnographic Atlas | 50:50 | 20-31% | 31% | 38-49% |
| Third US National Health & Nutrition Survey (1994) | 29,105 adults >20years | No data | 15.5% | 50.5% | 34% |

Fig1.9: A summary of dietary macronutrient estimates for hunter-gatherer and contemporary American diets. The table shows the increase in carbohydrate, and reduction in protein in the American diet compared to the two estimates for hunter-gatherer diets.

This contrasts greatly to typical western diets today. In the United States, the third National Health and Nutrition Survey (1994) showed that among adults aged >20y protein contributed 15.5%, fat 34%, carbohydrate 49% and alcohol 3.4% of energy intake. Not only have the relative contributions of macronutrients changed, but the foods from which these are obtained have also altered. The 1987-1988 National Food consumption survey indicated that cereal grains on average contributed 31% of the total energy intake of an individual, dairy products 14%, beverages 8% and sugar 4%. In short, the amount of protein consumed in western diets today has decreased and the proportion of carbohydrate has increased. This is largely due to the high reliance of diets in western societies on cereal grains, dairy products and refined sugars, none of which would have been available to Palaeolithic populations.

In another study, Sebastian et al. (2002) compared the net systemic load of acid (or net endogenous acid production, NEAP) of pre-agricultural diets with the

diets of modern western societies. Their analysis of the components of hunter-gatherer diets suggested that the transition to modern agricultural diets involved a switch from net base production to net acid production. They explain this shift as the result in a replacement of base-rich plant foods such as roots, tubers and leafy green vegetables by cereal grains, which are net acid producing, in addition to energy dense nutrient poor (EDNP) foods such as refined sugars and separated fats. The potential dangers of a chronic net acid producing diet are conditions such as osteoporosis, age-related muscle wasting, calcium nephrolithiasis, sodium chloride-sensitive hypertension, infertility and renal insufficiency (Sebastian et al. 2002).

Much of the research in this field has pointed out the potential health hazards of the chronic consumption of a diet to which our bodies are not sufficiently adapted. The dramatic changes in the diets of agricultural societies that occurred with the Neolithic transition have left little time for genetic adaptations to respond. Sebastian et al. (2002) comment that natural selection has had <1% of hominid evolutionary time to eliminate the inevitable maladaptations to the dramatic changes in diet that have occurred as a result of the development of agriculture. Modern diets with heavy dependence on agricultural products and high in EDNP foods and sodium chloride as well as low in fibre have been implicated in the 'diseases of civilisation': non insulin-dependant (type II) diabetes mellitus, atherosclerosis, hypertension, osteoporosis and certain types of cancer. As Eaton & Eaton (2000) put it genetically, humans remain Stone Agers – adapted for a Palaeolithic dietary regimen

Cordain et al. (2000)*a* draw attention to the insight that the study of hunter-gatherer macronutrient composition may have into therapeutic dietary recommendations for contemporary populations. In fact, Milton (1999 & 2000) argues that we should look to the diets of non-human primates to ascertain which foods are compatible with our digestive system. However, the human digestive system is reduced compared to that of non-human primates, possibly as an adaptation to the energetic cost of a large brain (see Aiello & Wheeler 1995). There is also much popular literature advocating various diets that claim in various ways to be more suited to our biological make-up (Graham 1998,

Cordain 2001). A number of questions remain, however: Firstly, exactly what was the ancestral human diet before agriculture, and to what extent have we adapted biologically to our agricultural foods? It is unlikely that we will ever have a satisfactory answer to the first question. It is the latter question that is the topic to be addressed in this thesis.

### 1.5 Starch, agriculture & amylase

We have already seen how the development of agriculture has had a profound impact on the diet and nutritional status of the populations that adopted it. But what biological adaptations have there been in the human digestive system to this agricultural diet? Cavalli-Sforza (1981) points out that it thus seems reasonable to suggest that the adaptation to agriculture may have involved an adaptation to low levels of protein and fat intake or high levels of carbohydrate, or their joint effects.

The main carbohydrate that is found in agricultural staples such as wheat, rice and maize, is starch. It is important to note that pre-agricultural human diets (see Wrangham et al 1999, Pennisi 1999, Harlan 1989) as well as non-human primate diets(see Wrangham et al 1991) would have contained some starch. However, after the development of agriculture the amount of starch in the diet is likely to have increased dramatically (see Cordain 2000b, Neel 1982, Schoeniger 1982, Turner 1979, Cavalli-Sforza 1981).

Amylase is an enzyme that is ubiquitous among animals that metabolize starch as part of their diet. Amylases break down glucose-polymers such as starch, glycogen and dextrines. The enzymes hydrolyse $\alpha$-1,4 glucosidic bonds between the glucose and maltose units that make up the starch molecule. In humans, amylase is produced by both the salivary glands and the pancreas. Starch digestion begins in the mouth where is broken down by the enzyme salivary amylase. After mastication the food is swallowed and enters the stomach. Amylase exhibits maximum activity at neutral pH. When the bolus enters the acid environment of the stomach, however, starch digestion can only continue as long as the acid has not penetrated into the bolus. As the food reaches the small

intestine, starch digestion is continued by the action of pancreatic amylase, secreted by the pancreas and delivered to the small intestine via the pancreatic duct. Townes et al. (1976) demonstrated that salivary amylase is present in the duodenum in patients who have lost pancreatic function and cannot secrete pancreatic amylase. This suggests that the enzyme is able to survive the passage through the acidic stomach.

There has been much debate over the relative roles of pancreatic and salivary amylase in starch digestion in humans. Merrit & Karn (1977) estimated that 60% of amylase activity is derived from pancreatic amylase where as salivary amylase contributes to around 40% of overall starch digestion. There are however difficulties in estimating this figure accurately, such as ascertaining which tissue a certain enzyme has been produced in, as well as problems with the sensitivity of protein detection assays and difficulties in distinguishing salivary and pancreatic isozymes.

### *1.5.1 Determination of the Structure & Evolution Human amylase multigene family*

Since the discovery of amylase in 1831 (Leuchs 1831) much has been published about the enzyme. The majority of these studies have focussed on characterisation of the biochemistry and genetics of amylase protein variants (Karmynt & Laxova 1966, de Soyza 1978, Pronk & Frants 1979, Merritt & Karn 1977). In 1965 Karmayt & Laxova produced the first evidence for the existence of two amylase loci in the human genome, one coding for salivary amylase and the other for pancreatic amylase. These two loci were later mapped to band p21 of the short arm of chromosome one by in situ hybridisation (Zabel et al. 1983, Tricoli & Shows 1984). In 1982 Pronk et al. found evidence for duplication of the human salivary amylase gene in humans, through studying amylase protein variants in a family, which contained an individual with three different salivary amylase gene products.Sequences of cDNAs for human salivary and pancreatic alpha-amylases were first published by Nakamura et al. (1984). The nucleotide sequences of the two cDNAs were 96% identical in the coding region, with predicted amino acid sequences of 94% identity. The sequence and structure of the exons of human salivary amylase was obtained by Nishide et al. (1986) by using human salivary

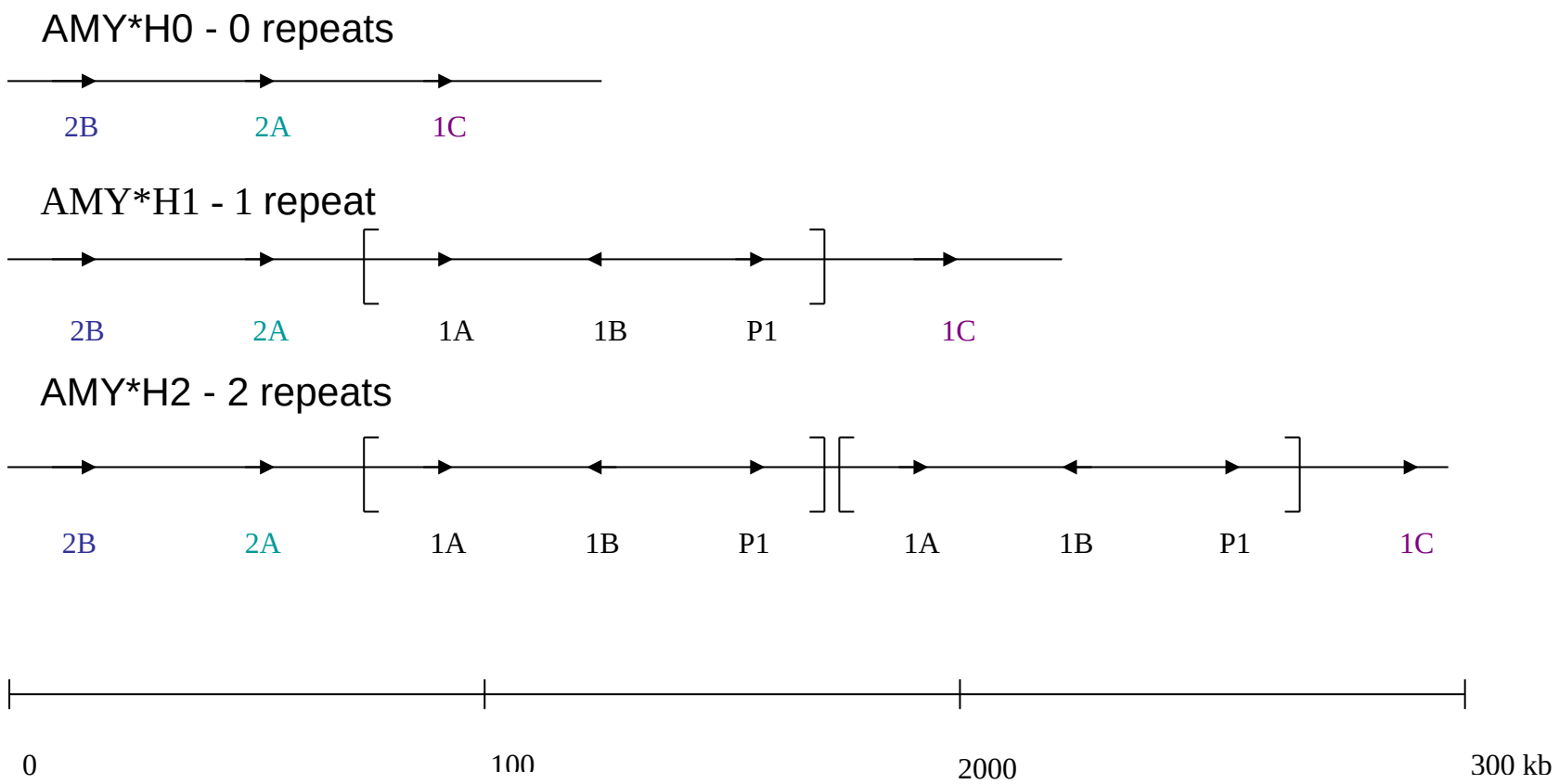



Fig 1.10: The Amylase gene cluster in humans according to Groot et al (1989): On each chromosome there are two pancreatic amylase genes: AMY2B & AMY2A as well as variable numbers of identical salivary amylase genes: AMY1A, AMY1B, AMY1C. The salivary genes are arranged into 100kb repeat units, which contain AMY1A, AMY1B and a pseudogene AMYP1.

amylase cDNA as a probe, followed by restriction mapping and sequencing, on a recombinant phage that was found to contain the whole human salivary amylase gene in a single insert. Horii et al. (1987) followed this with determining the exon structure and sequences for human pancreatic amylase. They found that the major difference between AMY1 and AMY2 lies in the fact that AMY1 has one extra exon on the 5' side. A third type of human amylase gene was identified from mRNA in a long carcinoid tissue, by Youchouchi et al. (1990), which they named AMY2B. This gene is highly homologous to AMY1 and AMY2, except that it has two untranslated exons in the 5' region so that the promoter lies far upstream relative to the other two AMY genes. The pancreatic genes (AMY2B and AMY2A) are closely related with approx 93% identity in the 5' flanking region (Groot et al. 1988, Groot et al. 1989*b*).

Gumuchio et al. (1988) reported finding seven distinct amylase genes in cosmid clones of 250 kilobases (kb) of genomic DNA. They found 2 pancreatic amylase genes, three salivary amylase genes and two truncated pseudogenes. Finally Groot et al. (1989*a*) demonstrated that the human amylase multigene family consists of haplotypes with variable numbers of AMY1 gene copies. Using a cosmid library and restriction maps from the same individual that led Pronk et al. (1982) to suggest the existence of duplicated salivary amylase genes, Groot et al. (1989*a*) identified two haplotypes consisting of different numbers of salivary amylase genes. The short haplotype contains two pancreatic amylase genes (AMY2B & AMY2A) and 1 salivary gene( AMY1C) arranged in the order 2B-2A-1C. In addition to this, haplotypes exist with repeated regions containing additional salivary amylase genes. The approx 100kb repeated region consists of two salivary amylase genes (AMY1A & AMY1B) and a truncated pseudogene (AMYP1). A general designation: 2B-2A-$(1A\text{-}1B\text{-}P1)_n$-1C can describe the different haplotypes, which range from n=0 (as in the short haplotype AMY1*H0) to n=4 copies of the repeated section, which produces a haplotype containing 9 functional copies of the salivary amylase gene (See fig 1.10). Groot et al. (1990) proposed that the AMY1 repeat haplotypes were formed through a series of unequal homologous crossover events (See fig 1.11).

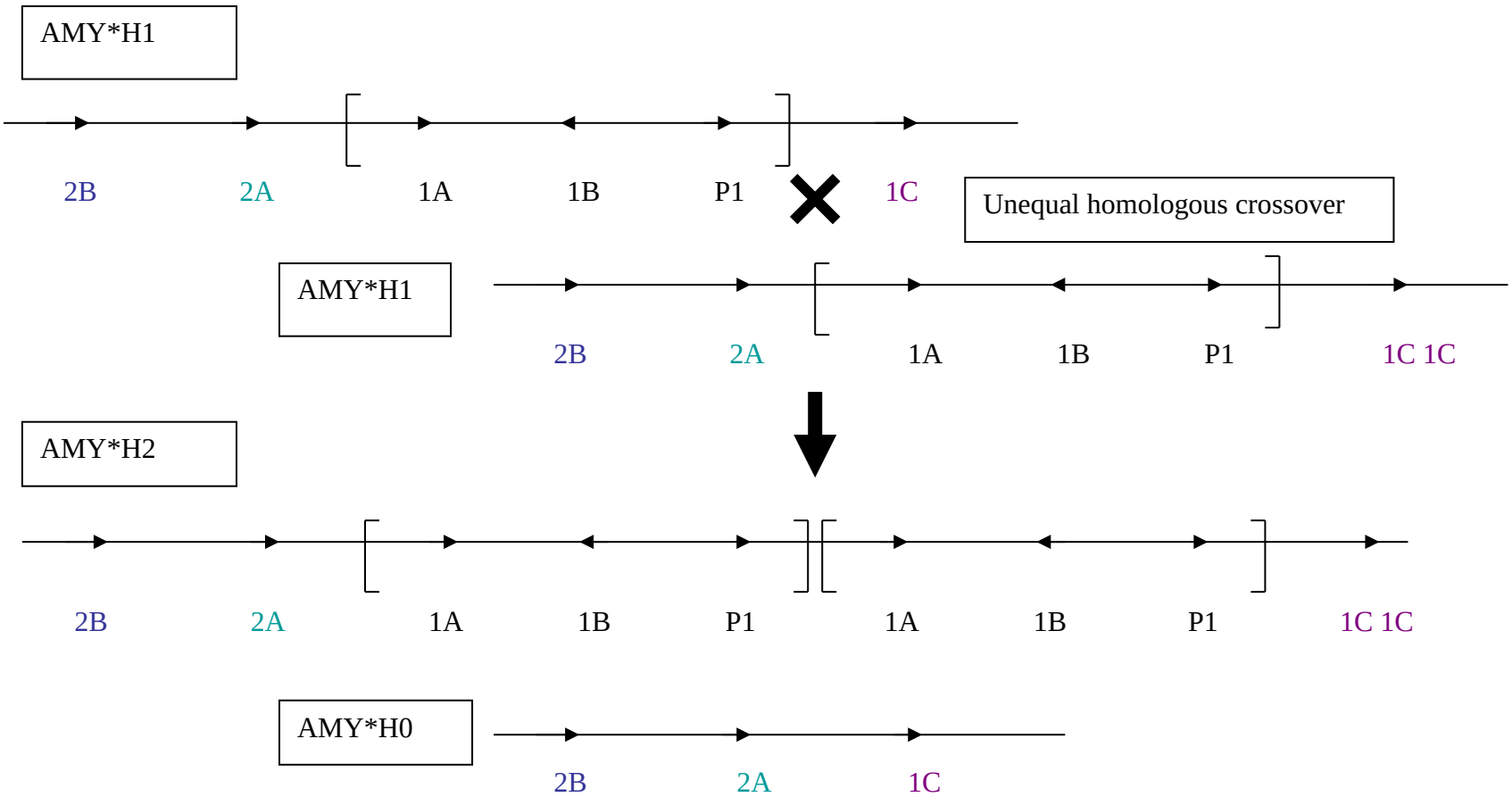



Fig 1.11: The expansion of the human amylase multi-gene family by unequal, but homologous crossovers, according to Groot et al (1989). It can be seen from the diagram how AMY*H2 and AMY*H0 can be created from two AMY*H1 chromosomes.

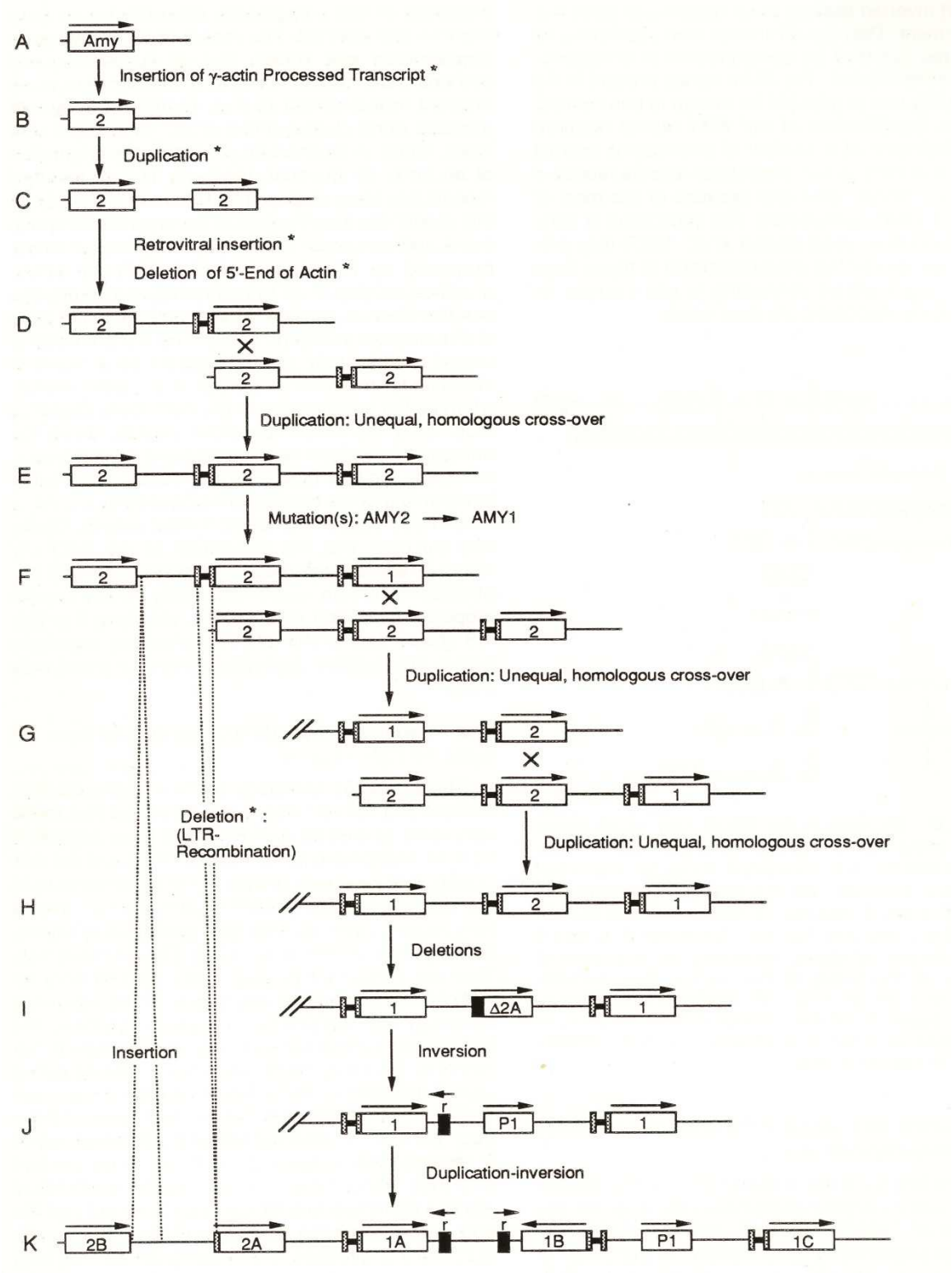
The three salivary genes (AMY1A, AMY1B, AMY1C) are almost identical to one another. The only published differences are that AMY1A & AMY1B have a small fragment (529bp, designated r) located approx. 3.5kb downstream of the genes (Groot et al. 1990). This r fragment contains sequence from exon three of the amylase genes and is completely absent in AMY1C. In addition, AMY1B is in reverse orientation to the other amylase genes. The pseudogene (AMYP1) is derived from exons 4-10 of AMY2A. Groot et al. (1990) suggested that the r fragments and the pseudogene are the remnants of the same ancestral pancreatic gene. It has also been suggested by Groot et al. (1990) and Gumucio (1988) that the evolution of the human amylase multigene family can be explained by a number of consecutive events involving duplications, insertions, deletions and inversions, gene conversions and unequal crossovers (See fig 1.12).

Samuelson et al. (1990, 1996) investigated amylase transcription in New-World monkeys, Old-World monkeys and apes. They studied two inserted elements, a γ-actin pseudogene and an endogenous retrovirus, in the salivary amylase promoter region. They found that the γ–actin peudogene was integrated after the divergence of the New-World monkeys from the primate ancestral tree and the retrovirus was integrated later after the divergence of the Old-World monkeys. They found that all human amylase (pancreatic and salivary) genes contain the γ-actin insert and therefore conclude that all the human amylase genes diverged from each other after this insertion event approximately 40 million years ago (Samuelson et al. 1990).

*1.5.2 Phenotypes & Methods employed to detect AMY1 haplotypes*

As the structure of the amylase gene family and the AMY1 haplotypes became clear, it was suggested that since there is inter-individual variation in number of salivary amylase genes, extensive variation in salivary amylase expression would also be expected. In the mouse strain YBR, Meisler et al. (1986) showed that salivary amylase synthesis was double that of wild type mice. They also found evidence in this strain of mice, from quantitative analysis of genomic DNA by Southern blotting, for duplication of the *Amy-1* locus. Bank et al. (1992) reported extensive

Fig.1.12: Hypothesis of the evolution of the human amylase multigene family (after Groot et al 1990). * = From the data of Samuelson et al (1990).



quantitative variation in salivary amylase enzyme in a sample of 369 individuals of Caucasian origin. They then went on to explore the variation at the DNA level using the polymerase chain reaction (PCR). They designed an assay to amplify an area surrounding a 22bp poly-A insertion in AMY1 genes that is not present in AMY2. This gave rise to PCR products of 604bp (all AMY1 genes) and 582bp (AMY2B & AMY2A). They then measured the difference in intensity between the two PCR products. As the number of pancreatic genes is always constant in humans, the 582bp AMY2 fragment acted as an internal standard. They found that in the majority of cases observed relative intensities of the PCR products fit well with the expected values derived from the protein quantification phenotyping. They concluded that quantitatively different salivary amylase enzyme phenotypes are encoded by haplotypes with different AMY1 gene copy number.

## 1.6 Experimental Rationale

Individuals with increased expression of the salivary amylase enzyme would be at an advantage in populations with high starch diets, such as the early agriculturalists. If the forces of natural selection for this locus were significant, we would therefore expect to find a high mean number of AMY1 gene copies per individual chromosome in populations with a long history of high starch diets. Conversely, we would expect to find lower numbers of salivary amylase genes in populations such as hunter-gatherers, who have low starch diets. Since the small scale study done by Bank et al. (1992) there has been no work published on the distribution of salivary amylase gene copy number variation in human populations.

Earlier in this chapter a number of different tests for selection were discussed. The analysis of Intra allelic variability (Slatkin & Bertorelle 2001) is currently the most powerful method for detecting the signature of natural selection within species (Sabeti et al. 2002). In order to conduct an analysis of intra allelic variability data must be collected on the allelic state at the locus in question as well as from a number of closely linked markers such as SNPs or microsatellites. With the recent advances in genotyping technology it is now possible to design fast, reliable and cost-effective methods to type large numbers of DNA samples from multiple populations.

If data on salivary amylase gene copy number variation, as well variation at a number of closely linked microsatellites, was available in a range of human populations with different dietary histories, it would be possible to employ tests of intra allelic variability to assess the level of evidence for selection at the AMY1 locus. Evidence that natural selection had been operating on the salivary amylase gene would provide us with a novel example of human dietary adaptation at the molecular level.

## Specific Aims of Thesis

1) Construct PCR based assays employing GeneScan technology for the high throughput typing of AMY1 polygenic repeat alleles and closely linked microsatellite markers.

2) Establish the frequency of AMY1 repeat alleles in a large number of human populations with different histories of agriculture and high starch diets,

3) Establish whether differences in AMY1 allele frequencies between populations were unusual compared to the rest of the genome.

4) Construct a PCR based assay employing GeneScan technology for the typing of AMY1 repeat alleles in chimpanzees.

5) Investigate whether variation in AMY1 gene copy number is present in chimpanzees.

6) Combine the microsatellite data with AMY1 repeat allele data for use in powerful haplotype based selection tests to test whether selection has operated on any of the AMY1 polygenic repeat alleles.

7) Extend the skills and experience gathered in the salivary amylase project to investigate other loci that may have had a role in dietary adaptation in humans.

# Chapter 2: Materials and methods

## Introduction

The first aim of this thesis was to construct PCR based assays for typing large numbers of individuals for the salivary amylase polygenic repeat alleles and a number of closely linked microsatellites. This chapter details the materials and methods that were employed in order to type human DNA samples for these markers. These protocols required a lengthy development and optimisation phase, which is described in Chapter 3. The data collected using these protocols are described in chapters 4 and 5.

In addition to collecting data on human DNA samples, this thesis also investigated AMY1 gene copy number in chimpanzees. The materials and methods for typing chimpanzee DNA samples are also detailed in this chapter in section 2.6. The development of this protocol and the resulting data are described in Chapter 6.

Note: Buffer compositions are given in section 2.10.3

## 2.1 DNA sample collection

Human DNA samples were obtained from The Centre for Genetic Anthropology, University College London, in the form of either buccal cells or extracted DNA. The exceptions were the Singapore Chinese family samples, which were obtained with kind permission from Prof David Goldstein, Dept. Biology University College London. The Irish and German family samples were collected for this project by Ms Noreen von Crammon-Taubadel between January and April 2003.

DNA samples from six Dutch individuals of known salivary amylase genotype (see Groot et al. 1989*a*) were kindly provided by Prof. Jan Pronk, Vrije Universiteit, Amsterdam. Chimpanzee DNA samples (see Ruano et al. 1992) were obtained with kind permission from Prof Dallas Swallow, Dept. Biology University College London.

Buccal cells were collected from subjects by rubbing a sterile applicator, in the form of a tube with a swab integral to the lid (Sarstedt, Numbrecht, Germany), gently along the inner surface of both cheeks for approximately 30 seconds. One ml of 0.05M EDTA/0.5% SDS preservative was then added. The samples were stored at room temperature during transit and then at 4°C or –20°C until extraction. Samples were collected from unrelated families (2 parents and at 1+ children) as well as unrelated adult males.

Informed consent was obtained from all donors. Ethical approval was obtained from University College Hospitals and University College London Joint Committee on Ethics of Human Research (ref. 99/0196). Appropriate permissions were obtained in each of the collection countries. All donors provided details of self-defined ethnic identity, first and second language and place of birth with similar information on their mother, father, maternal grandmother and paternal grandfather.

**2.2 DNA extraction**

Firstly 40 µl of 10 $mgml^{-1}$ protinase K was added to 40 ml of sterile distilled water. Once the buccal swab tubes had been defrosted, 0.8ml of the water/protinase K mix was added to each swab tube and then incubated at 56°C for a minimum of two hours. After incubation 0.8ml of the solution was taken and added to 0.6ml phenol/chloroform (1:1) in a 1.5ml tube. The mixture was vortexed thoroughly and then centrifuged for 10 min at maximum speed in a microfuge. The remainder of sample was stored at –20°C as a backup. The upper aqueous layer was transferred to a fresh 1.5ml tube containing 0.6ml chloroform and 30µl of 5 M NaCl, and then mixed and centrifuged for 10 min. The upper aqueous layer was then transferred to a fresh 1.5ml tube containing 0.7ml chloroform. The samples were then mixed and centrifuged for 10 min

DNA was precipitated by adding the aqueous layer to a screw top microfuge tube containing 0.7ml isopropanol, mixing and then leaving it at –20°C for a

minimum of 2 hours. It was then centrifuged at 13,000 RPM for 15 min to pellet the DNA. Supernatant was poured off, the tube inverted and the pellet allowed to dry for 1 min. The pellet was then washed by adding 0.8ml 70% ethanol and centrifuged at 13,000 RPM for 15 min. The ethanol was then poured off and the pellet was left to dry for 20 min. 400 µl TE buffer (pH 8.0) was added and the sample was incubated for 10 min at 56°C whilst being mixed occasionally. The extracted DNA samples were stored at –20°C. Agarose gel electrophoresis was used to assess DNA yield. Ten µl of extracted DNA sample was run on a 0.8% agarose gel in an Advanced Biotechnologies (Columbia, MD) gel tray at 25v for 10 min, followed by 100v for 30 min. DNA was visualised with UV/ Ethidium Bromide staining.

## 2.3 Polymorphism detection

### *2.3.1 AMY1 gene copy number quantification*

The arrangement of the salivary amylase genes in the amylase gene family cluster on chromosome one, presented a number of challenges for designing assays to type individuals for AMY1 polygenic repeat alleles. A number of different approaches were explored and the process involved in the development and optimisation of the resulting protocols is described in Chapter 3.

The aim was to design a method to determine the AMY1 polygenic repeat alleles present in individuals. As was outlined in Chapter 1, salivary amylase gene (AMY1) copy number varies between individuals. However, the number of pancreatic amylase genes (AMY2) remains constant. All the methods that were explored were aimed to quantify the number of AMY1 genes in an individual by using the AMY2 genes as an internal control. By using the amount of AMY1 product relative to the amount of AMY2 product, the number of AMY1 genes present in the individual could be determined.

The differences between the protocols that were developed, was in the method used to distinguish the AMY1 and AMY2 specific products from each other prior to quantification. Two main approaches were investigated: The first used restriction endonucleases to distinguish PCR products resulting from AMY2 and

AMY1 genes. The second method was based on PCR assays that amplified areas around small deletions in either AMY2 or AMY1 genes resulting in different length AMY1 and AMY2 PCR products. The following section details the final materials and methods used for each of the two approaches.

*a) Restriction endonuclease protocol*

This protocol involved a PCR with two primer pairs. Each primer pair amplified a region in all the amylase genes that contained a restriction endonuclease recognition site. However, the PCR products resulting from the first primer pair had a restriction enzyme recognition site was present only in the AMY1 genes. Thus, after lysis with the restriction endonuclease, the AMY1 products were cut where as the AMY2 products remained uncut. In the case of the PCR product resulting from the other primer pair, the restriction enzyme recognition site was present in only the AMY2 genes, so that the AMY2 products were cut and the AMY1 products were not (See Table 2.1).

PCR reactions were performed in 10 µl volumes containing 200µM dNTPs, 10 mM Tris HCl (pH 9.0), 0.1% Triton-X-100, 0.01% gelatin, 50mM KCl, 1.5 mM $MgCl_2$, 0.13 units Taq polymerase enzyme (HT Biotech), 2.4 µM TaqStart Monclonal Antibody (MAb) (BD Biosciences Clontech, San Jose CA) and primers to the concentrations given in Table 2.1. Cycling parameters were: pre-incubation for 5 min at 95°C, followed by 37 cycles of 93°C for 1 min, 59°C for 1 min, 72°C for 1 min and then a final incubation step of 72°C for 20 min.

All PCR reagents except the Taq polymerase and TaqStart MAb were premixed in batches sufficient for 96 reactions and stored at –20°C. DNA samples were typed in batches of 95 with the one remaining tube acting as a negative PCR control. The use of TaqStart Mab increases the specificity of the PCR (Thomas et al. 1999), and so was used in all reactions. The Taq and TaqStart Mab were mixed as 2 volumes of 5 units/µl Taq : 1 volume of 7 µM TaqStart Mab and stored at –20°C in 20µl aliquots. Primers were also mixed and stored as a 10x stock to save time and reduce the errors associated with pipetting small volumes. To minimise the time that the Taq enzyme was in contact with primers and other

| Primer name | Primer sequence (5'-3') | ABI dye label | Tm | Primer pair final conc (μM) | Genes Cut | Genes not cut | Product lengths | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | cut | uncut |
| AMY 02-U-HEX | 5′- AAA GGC AAT TTT GGA CAA ACT G –3′ | HEX | 52.9 | 0.2 | AMY 2B, 2A, P1 | AMY1 | 52 bp | 82 bp |
| AMY-02-L | 5′- TAC CTC CTG GTA AAT GAA AGG TTT A –3′ | - | 52.7 | | | | | |
| AMY-04-U-FAM | 5′- GTC TTC CTG CTG GCA CAT ACT –3′ | FAM | 51.0 | 0.08 | AMY1 | AMY 2B, 2A, P1 | 60 bp | 83 bp |
| AMY-04-MML-B | 5′- AGA AAC GTA GAT TTT AAT GCC TCT –3′ | - | 50.7 | | | | | |

Table 2.1: Primers used for AMY1 quantification using the restriction enzyme *Pst1*.

reagents, 1 μl of DNA template was added to the bottom of each sample's 0.2 ml PCR tube. Only then was the PCR premix, containing all the primers and buffer reagents, thawed out. The Taq/TaqStart mix was added to the other PCR components in the PCR premix, just prior to thermal cycling and the mixture vortexed thoroughly. Nine μl of the PCR mix was pipetted into the lid of each PCR tube, and the plates centrifuged to collect all PCR components at the bottom of the tubes. All amplifications were performed in a BioMetra (Whatman Biometra, Goettingen, Germany) Uno II thermal cycler using a 10-μl reaction volume.

The PCR products were precipitated before lysis with the restriction endonucleases to remove excess PCR reagents and to minimise inhibition of restriction enzymes. One μl of 3M sodium acetate pH 5.2 and 24.2μl of 100% ethanol was added to the PCR products which were then mixed before being placed at –20°C for at least 1 hour. The sample was centrifuged at 13,000 RPM for 12 mins, after which the supernatant was poured off. One ml of 70% ethanol was added and the tube was inverted gently. The sample was then centrifuged again at13,000 RPM for 10 mins, and the supernatant pored off. One ml of 70% ethanol was again added and the sample centrifuge at 13,000 RPM for 10 mins. Finally the supernatant was poured off and the pellet allowed to dry for 30 mins before the DNA was re-suspended in water. Digestions were performed in 384-well microtiter plates in a final volume of 15 μl. Each reaction contained 3 μl of PCR product, NEB buffer 2 (New England Biolabs, Beverly, MA) to a 1x concentration, 0.01 μg/μl acetylated BSA and 5 units *Pst1*. Plates were incubated at 37°C overnight.

*b) The QAMY protocol*

This protocol involved two separate PCRs for the two markers (QAMY02 & QAMY03). The PCRs were designed to amplify regions around small deletions

present either in AMY1 or AMY2 genes. The resulting PCR products could be distinguished by differences in length depending on the presence or absence of the deletion.

PCR reaction conditions were 200 μM dNTPs, 10mM Tris-HCl (pH 9.0), 0.1% Triton X-100, 0.01% gelatin, 50 mM KCl, 1.2 mM $MgCl_2$, 0.13 units *Taq* polymerase (HT Biotech, Cambridge, UK), 9.3 nM TaqStart monoclonal antibody (Mab) (BD Biosciences Clontech, San Jose, CA), and primers to the concentrations given in table 2.2. Cycling parameters were 95°C for 5 min followed by 30 cycles of 94°C for1 min, 60°C for 2 min and 72°C for 3 min; and then a final incubation step of 72°C for 10 min. After the PCR reactions were completed Products from both PCRs (QAMY02 & QAMY03) were mixed together and diluted (1/2) with water.

Table 2.2: Primers and final primer concentrations for AMY1 quantification protocol. Tm was calculated using Oligo v4.0 software (see Table 2.5)

| Primer Name and Sequence (5'-3') | Primer Conc. (μM) | TM | Product size (bp) |
|---|---|---|---|
| QAMY02-U<br>5′-ATG TGC TGT TAA TAT TTT CAA GAG AT-3′ | 0.0375 | 50.1 | 1872B / 1912A+1 |
| QAMY02-L-TET<br>5′- CCA AGG TCT GAA AGG GTT GT -3′ | 0.0375 | 50.6 | |
| QAMY03-U<br>5′-TCA CAG TTG ATT TTT GAT CTT GTA G-3′ | 0.0375 | 50.2 | 263 AMY1 / 267 AMY2 |
| QAMY03-L-TET<br>5′-GAC TGC TGG AAA GTC CCT ACT T-3′ | 0.0375 | 51.1 | |

### 2.3.2 Microsatellite multiplex PCR protocol

This protocol was designed to amplify regions around 6 microsatellites closely linked to the amylase gene cluster. A six primer pairs were combined into a single PCR reaction known as a multiplex PCR.

Reaction conditions were 200 μM dNTPs, 10mM Tris-HCl (pH 9.0), 0.1% Triton X-100, 0.01% gelatin, 50 mM KCl, 2.2 mM $MgCl_2$, 0.13 units *Taq* polymerase (HT Biotech, Cambridge, UK), 9.3 nM TaqStart monoclonal antibody (Mab) (BD Biosciences Clontech, San Jose, CA), and primers to the concentrations given in table 2.3. Cycling parameters were 95°C for 5 min followed by 37 cycles of 94°C for 1 min, at 58°C for 1 min, and 72°C for 1 min; and then a final incubation step of 72°C for 10 min.

*2.3.3 Electrophoresis and GeneScan Analysis*

The ABI377 / GeneScan™ (PE-Applied Biosystems, Foster City, CA) is an integrated genotype technology platform based on polyacrylamide gel electrophoresis with a laser detection system and a laboratory information management system. DNA molecules are fluorescently labelled, and both the wavelength and intensity of the fluorescence and the time taken for the molecule to migrate toward the laser from the start of the run is measured. The time taken for a DNA fragments to migrate is proportional to its size. The fluorescence detected by the laser system is displayed as peaks in the GeneScan™ Analysis v3.1 software (PE-Applied Biosystems, Foster City, CA). Relative quantities of two 5'- end labelled fragments can be determined by comparing the corresponding peak areas or peak heights on the resulting GeneScan™ electropherogram.

The size of fragments was determined using the Local Southern method, which uses the reciprocal relationship between fragment length and mobility. Each sample was loaded with an internal size standard (such as TAMRA-350) with a range of fragment lengths of known size. The software uses the four fragments closest in size to the unknown fragment to determine a best-fit line value for the

| Primer Name | Primer sequence (5' – 3') | ABI dye label | Final concentration (μM) | Repeat Motif | Tm °C |
|---|---|---|---|---|---|
| D1S2896-U | 5′- CAT AGT TTC AAC CAC TGG CTA AT -3′ | / | 0.025 | CA | 50.0 |
| D1S2896-L-TET | 5′- GTG CCC AAT CCA ATT TAA TTC -3′ | TET | 0.025 | | 50.6 |
| D1S2888-U-FAM | 5′- GGC AAT ACA AAA TTC AAG TTA TAG AC -3′ | FAM | 0.05 | CA | 50.2 |
| D1S2888-L | 5′- GTA AGT TAG GCA ACA ATT AAC ACA TAG -3′ | / | 0.05 | | 50.5 |
| D1S2759-U | 5′- CAT CTC ACC TTC ACA ACC TCC -3′ | / | 0.05 | CA | 50.9 |
| D1S2759-L-HEX | 5′- CCC CTT TCA GTG ATA TAA AAT TAA A -3′ | HEX | 0.05 | | 50.5 |
| D1S2626-U | 5′- ACA GGA TGT AGG GAA GAA TTG TAT A -3′ | / | 0.15 | CA | 50.3 |
| D1S2626-L-FAM | 5′- CCT CCC TGA CAG ATT TTG AAC -3′ | FAM | 0.15 | | 50.4 |
| D1S535-U-TET | 5′- GTG GGA ATT ATG GGG GTT AC -3′ | TET | 0.1 | GATA | 50.0 |
| D1S535-L | 5′- TGC TAA GTG AGA AAA CAC ATT GTT A -3′ | / | 0.1 | | 50.9 |
| MS-AMY02U | 5′- ACT GTC CTT ATT TAT GTG GGT TTG T -3′ | / | 0.05 | CA | 52.2 |
| MS-AMY02L-FAM | 5′- TCT CTT CTT CCA TTG CGA CTG -3′ | FAM | 0.05 | | 52.1 |

Table 2.3: Primers for use in the microsatellite multiplex PCR protocol. Tm was calculated using Oligo v4.0 (see Table 2.5)

unknown fragment. Fragment sizes for TAMRA-350 size standard are as follows:
35, 50, 75, 100 139, 150, 160, 200, 300, 340 & 350 bp.

All PCR products were run on an ABI-377 automated sequencer; 1.1-µl aliquots of PCR product was mixed with 2.0 µl of loading buffer (formamide: dextran blue: TAMRA -labelled 350bp size standard in the ration 12:2:1). Electrophoresis was performed on a 5% polyacrylamide 36cm gel (National Diagnostics, Atlanta, Georgia). For both the QAMY and microsatellite protocol, the electrophoresis conditions were 2.5h at 3000 volts. For the restriction enzyme protocol the run time was 1.6h. ABI PRISM$^{TM}$ collection software (PE-Applied Biosystems, Foster City, CA) saves scan profiles as gel files, from which the raw data is then extracted. GeneScan Analysis v3.1 software (PE-Applied Biosystems, Foster City, CA) was then used to analyse the data.

**2.4 DNA Sequencing**

DNA sequencing was used at a number of different stages in the development of protocols:
1) Confirming the presence of restriction endonuclease recognition sites and deletions in PCR products whilst developing the AMY1 quantification protocols
2) Determining the number of microsatellite repeat motifs contained in PCR products from the microsatellite protocol
3) Confirming the presence of deletions, and to aid in the design of primers for the chimpanzee AMY1 quantification protocol

The PCR product was purified using an equal volume of MicroCLEAN (Microzone Ltd, Haywards Heath, UK) to PCR product and mixing. After this mixture had been left at room temperature for 10 min, the PCR product/MicroCLEAN was centrifuged at between 2000 and 4000 g for 40 min in a plate centrifuge. The supernatant was then removed by inverting the plate and centrifuging at 50 g, and then 150µl of 70% ethanol was added. This was

then centrifuged at 4000 g for 10 min. The supernatant was removed and the sample was allowed to air dry for 15 min at room temperature. 15µl of water was then added to re-suspend each sample and 5.5µl of this was used for the sequencing reaction. Sequencing reaction conditions were 5µl Better Buffer (Microzone Ltd, Haywards Heath, UK) 1µl Termination mix from the ABI Prism BigDye Terminator Kit (PE-Applied Biosystems, Foster City, CA), and primers at 1.6 pm/µl. The sequencing reactions were performed in a GeneAmp PCR system 9700 thermal cycler (PE-Applied Biosystems, Foster City, CA) with 25 cycles of 96°C for 10 seconds, 60°C for 5 seconds and 60°C for 4 minutes. To purify the sequencing reaction products, 80µl of 80% isopropanol was added to each reaction and mixed thoroughly and left at room temperature for 10 min. The samples were then centrifuged at between 2000 and 4000 g for 40 min. The supernatant was removed and then 150µl 70% isopropanol was added to each sample. The reactions were then spun at between 2000 and 4000 g for 10 min. Supernatant was once again removed and the samples were allowed to air dry at room temperature for 15 min. The samples were run on an ABI 3100 genetic analyser. Prior to electrophoresis of the samples, 10µl of HiDi formamide was added to each sample. They were mixed and heated to 65° for 5 min to dissolve the sequencing products fully in the formamide. The samples were then transferred to a 96 well plate suitable for use on the ABI3100 machine. The samples were denatured at 96°C for 4 min and then cooled on ice for 5 min. Samples were run on an ABI3100 machine (PE-Applied Biosystems, Foster City, CA), and aligned and checked for read quality using Sequencher software (Gene Codes, Ann Arbor, Michigan).

**2.5 Cloning PCR products**

During the development of the chimpanzee AMY1 quantification protocol it was necessary to clone the PCR fragments into plasmid vectors prior to sequencing in order to separate the AMY1, AMY2A and AMY2B products.

PCR reactions were performed in 10 µl volumes containing 200µM dNTPs, 10 mM Tris HCl (pH 9.0), ).1% Triton-X-100, ), 0.01% gelatin, 50mM KCl, 1.2 mM $MgCl_2$, 0.13 units *Taq* polymerase enzyme (HT Biotech), 2.4 µM TaqStart

Monclonal Antibody (MAb) (Clontech) and primers at 0.3μM. Cycling parameters were: pre-incubation at 95°C for 5 min, followed by 37 cycles of 93°C for 1 min, 56°C for 1 min, 72°C for 1 min and then a final incubation step of 72°C for 20 min. Amplification reactions were performed in a GRI DYAD™ DNA Engine Thermal cycler. Two μl of PCR product were run on 2% agarose gel to confirm the presence of DNA of the expected size.

PCR products were purified by adding 3 times the volume of the samples, of 4/3 MicroClean (Microzone Ltd), mixing and then leaving to stand at room temperature for 10 min. The samples were then centrifuged at 13000 RPM for 15 min. The supernatant was removed and 200 μl 70% ethanol was added. The samples were centrifuged at 13000 RPM for 5 min. The supernatant was removed and the samples were air dried at room temperature for 15 min. The samples were finally re-suspended in their original volume with water.

Cloning was performed using TOPO TA Cloning Kit for Sequencing (Invitrogen, Carlsbad, CA). Ligation of the PCR product into the pCR®4-TOPO® plasmid vector was carried out using 1 μl PCR product, 1μl water, 0.5μl salt solution (200mM NaCl, 10mM $MgCl_2$), 0.5μl vector. The resulting components were mixed and left at room temperature for 5 min and then cooled on ice. 2μl cloning reaction was transferred to a vial of One Shot® TOP10 Chemically Competent *E. coli* and mixed gently. The mixture was incubated on ice for 5 min and then heat shocked for 30 seconds at 42°C, and then put on ice. 250 μl of room temperature SOC medium was added and incubated, whilst shaking, at 37°C for one hour. 100μl of sample was spread onto a pre-warmed LB plates containing 50 μg/ml ampicillin and the plates were incubated overnight at 37°C.

24 colonies were picked from the plates and a PCR reaction was performed using M13 Forward (-20) (5'-GTAAAACGACGGCCAG-3') and M13 reverse (5'-CAGGAAACAGCTATGAC-3') primers. PCR reactions were performed in 10 μl volumes containing 200μM dNTPs, 10 mM Tris HCl (pH 9.0), ).1% Triton-X-100, ), 0.01% gelatin, 50mM KCl, 1.2 mM $MgCl_2$, 0.13 units *Taq* polymerase

enzyme (HT Biotech), 2.4 μM TaqStart Monclonal Antibody (MAb) (Clontech) and primers at 0.3μM. Cycling parameters were: pre-incubation at 94°C for 5 min, followed by 30 cycles of 94°C for 1 min, 55°C for 1 min, 72°C for 1 min and then a final incubation step of 72°C for 10 min. Amplifications were performed in a GRI DYAD™ DNA Engine Thermal cycler. Sequencing of the resulting PCR products was then carried according to the protocol described in section 2.4.

**2.6 Chimpanzee QAMY protocol**

The chimpanzee QAMY protocol was developed in order to investigate whether the variation in AMY1 gene copy number that is found in humans is also present in chimpanzees. Details of the development of this protocol can be found in Chapter 6.

PCR reactions were performed in 10 μl volumes containing 200μM dNTPs, 10 mM Tris HCl (pH 9.0), ).1% Triton-X-100, ), 0.01% gelatin, 50mM KCl, 1.2 mM $MgCl_2$, 0.13 units *Taq* polymerase enzyme (HT Biotech), 2.4 μM TaqStart Monclonal Antibody (MAb) (Clontech) and primers QAMY02-CH-U (5′-GAA TGG CGA TGG GTT GAT AT-3′) and QAMY02-LTET (5′-CCA AGG TCT GAA AGG GTT GT-3′) at 0.2μM. Cycling parameters were: pre-incubation for 5 min at 95°C, followed by 30 cycles of 93°C for 1 min, 56°C for 2 min, 72°C for 3 min and then a final incubation step of 72°C for 20 min. Amplification reactions were performed in a GRI DYAD™ DNA Engine Thermal cycler. 2μl PCR product was run on 2% agarose gel to confirm the presence of PCR products of the expected size.

The PCR products were diluted 1 in 5 with water. 1.1μl of the diluted PCR product was then added to 2μl of 1:2:12 mixture of TAMRA-500 size standard: dextran blue loading buffer: deionised formamide. Samples were denatured at 96°C for 3 min and loaded onto the ABI377. Electropheresis was conducted in a 5% polyacrylamide gel for 3.5 hours.

Two different sized PCR products were obtained from the PCR reaction. The 462bp PCR product was interpreted as originating from AMY2B genes, where are the 466bp product was interpreted as originating from AMY1A and AMY2A genes.

### 2.7 Establishing phase

The methods for quantifying the number of AMY1 genes present in individuals as described in section 2.2, require and additional step to determine the apportionment of the genes, arranged into polygenic repeat alleles, between the maternal and paternal chromosomes. This process if known as establishing phase, or haplotyping. In order to establish the phase of the AMY1 repeat alleles, a set of functions, called EMamy, incorporating an expectation maximisation (EM) algorithm was written for the MATLAB programming environment by M. Weale (see Table 2.6). The aim of the program was to resolve the haplotypes of both the father and the mother, given their resulting children's genotypes. The functions were designed to analyse data in the form of the total number of AMY1 repeat units present in an individual, from families consisting of one father, one mother and two children. However, the genotypes for either or both children were allowed to be missing, so that non-family samples could be analysed. The functions return EM estimates for allele frequencies and counts for each allele type given by families that can be fully resolved. All the possible parental genotypes are also reported, given the children, together with the relative probability for each genotype, using all the data as well as the allele frequencies from the EM estimates. Further details of the development of the *EMamy* functions can be found in section 3.3.5.

Compound haplotypes consisting of both microsatellite alleles and EM estimates of AMY1 repeat number alleles were established using DNA samples from families (2 parents and at least one biological child). Haplotypes were assigned by following the pattern of co-inheritance of the alleles from the parents to the children, (see Nehati-Javeremi & Smith 1996) (See Fig 2.1).

Fig 2.1: A summary of haplotype assignment (after Nejati-Javaremi & Smith 1996)

| **Father** | Haplotypes | | | **Mother** | Haplotypes | | |
|---|---|---|---|---|---|---|---|
| Genotype | Locus | a | b | Genotype | Locus | c | d |
| 18-19 | D1S2896 | 18 | 19 | 19 | D1S2896 | 19 | **19** |
| 19 | D1S2626 | 19 | 19 | 25 | D1S2626 | 25 | **25** |
| 18-22 | D1S2888 | 18 | 22 | 23 | D1S2888 | 23 | **23** |
| 21-27 | D1S535 | 21 | 27 | 20-21 | D1S535 | 20 | **21** |
| 17-18 | D1S2759 | 17 | 18 | 18-21 | D1S2759 | 21 | **18** |

| **Child 1** | Haplotypes | | | **Child 2** | Haplotypes | | |
|---|---|---|---|---|---|---|---|
| **Genotype** | **Locus a** | **c** | | **Genotype** | **Locus** | **b** | **d** |
| 18-19 | D1S2896 | 18 | 19 | 19 | D1S2896 | 19 | **19** |
| 19-25 | D1S2626 | 19 | 25 | 19-25 | D1S2626 | 19 | **25** |
| 18-23 | D1S2888 | 18 | 23 | 22-23 | D1S2888 | 22 | **23** |
| 20-21 | D1S535 | 21 | 20 | 21-27 | D1S535 | 27 | **21** |
| 17-21 | D1S2759 | 17 | 21 | 18 | D1S2759 | 18 | **18** |

Numbers indicate the repeats count of each microsatellite motif. Given that the genotypes of both parents and progeny are known, haplotypes over several linked loci can be assigned by listing the allele type at each locus along the haplotype known to be inherited from each parent. Thus in this example the assignment procedure occurs as follows:

At locus D1S2896 the 18 allele present in Child 1 can only be transmitted paternally, along with the 19 (D1S2626), 18 (D1S2888) and 17 (D1S2759) alleles. Although the 21 allele (D1S535) in Child 1could be maternally derived, the presence of the 20 (D1S535) which could only come from the mother confirms that the 21 allele is paternally derived. These observations are independently confirmed by assigning Child 2's haplotypes.

## 2.8 Statistical analysis

*2.8.1 Analysis of AMY1 gene copy number data to test for significant differences between populations*

Non parametric multivariate analysis of AMY1 gene copy number in human populations was carried out using the Kruskal Wallis test, calculated in the statistical analysis software Instat (GraphPad, San Diego, CA). Variances were compared using F statistics calculated using Microsoft Excel. The Dunn-Sidak correction was used on all multiple comparisons of populations to take care of type 1 error in pair-wise comparisons:

$\alpha' = 1-(1-\alpha)^{1/k}$

Where k= no objects (populations)

$\alpha$ = significance level applied to any one test (0.05)

*2.8.2 Analysis of AMY1 repeat allele frequency data from human populations to test for significant differences between populations*

Estimates of AMY1 repeat allele frequencies were obtained from the *EMamy* functions implemented in MATLAB (Mathworks, Natick, MA) as described in sections 2.8 & 3.3.5. These frequency estimates were used to estimate the genetic distance between populations using an analysis of molecular variance (AMOVA) (see Wier & Cockerham 1984, Excoffier et al. 1992, Wier 1996) based on the statistical measure of population difference - $F_{ST}$, implemented using the Arlequin program (Schneider et al. 2000). The $F_{ST}$ measure, first suggested by Wright (1951) is defined in many ways. One formulation appropriate for data on AMY1 gene copy number is :

$$F_{ST} = \frac{V_T - V_W}{V_T}$$

where $V_t$ = Total variance of AMY1 repeat alleles of a set of n populations

and $V_W$= mean variance of AMY1 repeat alleles within populations. In practice the above method is further modified by bias correction factors (see Wier 1996).

Expected heterozygosity (*h*) (equivalent to genetic diversity, see Nei 1987) was calculated using the formula:

$$1 - h = \sum_{i=1}^{m} x_i^2$$

where m- number of alleles

and $x_i$ = EM estimate of frequency of *i*th allele

*2.8.3 Displaying allele frequency differences between populations*

The pair-wise $F_{ST}$ comparisons for the populations, using the unbiased 'random populations' formula for haploid data given by Weir (1996), were compiled as a matrix and subjected to a principal co-ordinate analysis using the Genstat v3.2 software (VSN, Hemel Hempstead, UK). Similar to Principal Component Analysis, this procedure explains the principal vectors of variance between population groups and extracts as many vectors as required to account for these differences. The first and second vectors were plotted against each other to visualise trends in variation between groups using the MATLAB programming environment (Mathworks, Natick, MA).

*2.8.4 Analysis of microsatellite data*

Microsatellite haplotype data was analysed by an AMOVA, as well as an exact test of population differentiation based on haplotype frequencies, and genetic distances measured using $R_{ST}$ implemented using the Arlequin program (Schneider et al. 2000). $R_{ST}$ is analogous to $F_{ST}$ but incorporates into the model the step-wise mutation process (see Slatkin 1995). (see Michalakis & Excoffier 1996 and Rousset 1996 for details on the relationship between $F_{ST}$ and $R_{ST}$).

*2.8.5 Comparisons with polymorphism data from other loci in the genome*

Data on SNPs typed in 42 African Americans, 42 East Asians and 42 European Americans (Sachidanandam et al. 2001) was taken from a dataset of 33,487 SNPs typed by the Orchid Laboratory, publicly available at the SNP Consortium web site

(http://snp.cshl.org/allele_frequency_project/panels.shtml). Statistical analysis of the SNP data set was performed using the statistics package 'R' (URL: http://www.R-project.org/). All $F_{ST}$ values were calculated using the unbiased 'random populations' formula for haploid data given by Weir (1996).

$R_{ST}$ data for STRs typed in 48 Europeans (blood donors from Leipzig, Germany) and 23 East Africans (from Gondar, Ethiopia) as well as 24 Southern Africans (from the Nguni, Sotho-Tswanga and Tsonga groups of South Africa) was taken from a dataset of 332 STRs typed by Kayser et al. (2003). Statistical analysis of the STR data was performed using the statistics package 'R' (URL: http://www.R-project.org/).

*2.8.6 Analysis of Intra allelic variability*

Analysis of intra-allelic variability was carried out on compound haplotypes consisting of the AMY1 repeat allele as well as the 6 microsatellites using the program SYSSIPHOS written by Dr Michael Stumpf, Imperial College London. SYSSIPHOS is an updated version of the programs *NeutraliyTest* (Slatkin & Bertorelle 2001), *EstimateGrowth* (Slatkin & Bertorelle 2001) and *EstimateS* (Slatkin 2001) available from Prof. Montgomery Slatkin, University of California, Berkeley. In contrast to earlier programs, SYSSIPHOS was designed to analyse data from multiple microsatellites simultaneously, as well as take into consideration recombination between the microsatellite loci. In addition, departures from the stepwise mutation model (Slatkin 1995) such that a length dependent microsatellite mutation rate (see Stumpf and Goldstein 2001) is taken into account. For a given allele at the candidate locus, the likelihood of the data is estimated over a range of selection coefficients ($s$) and exponential population growth ($r$) parameter values. Post-processing of the SYSSIPHOS output files was carried using the statistics package 'R' (URL: http://www.R-project.org/). (See Appendix B)

Support intervals were calculated by taking a reduction in log-likelihood of 1.92 from the maximum (Edwards 1992). In addition, a likelihood ratio test was used to test between the hypothesis that significantly greater selection has been operating on one AMY1 repeat allele compared to another, and the null hypothesis that the difference is not

significant. Likelihood ratio tests compare the ability of alternative models to explain the data, by considering the significance of the following statistic:

$$2\log \frac{\text{maximum likelihood under alternative hypothesis}}{\text{maximum likelihood under null hypothesis}}$$

This test statistic approximates to the $\chi^2$ distribution with one degree of freedom.

*2.8.7 Estimating the age of alleles*

The average ages of AMY1 repeat alleles were estimated using both the intra-allelic variability method (implemented in SYSSIPHOS) and an Average Squared Distance (ASD) method using YTIME written by M. Weale, University College London (see Thomas et al. 2002) (See section 5.3).

## 2.9 Bioinformatics and population genetics analysis tools

Table 2.4 Human genome databases

| Name | Purpose | URL |
|---|---|---|
| GDB | Sequences from AMY exons | http://www.gdb.org/ |
| OMIM | Sequences from AMY exons | http://www.ncbi.nlm.nih.gov/Omim/ |
| GenBank | Sequences from AMY exons | http://www.ncbi.nlm.nih.gov/Web/GenBank/ |
| UniGene | Sequences from AMY exons | http://www.ncbi.nlm.nih.gov/UniGene/ |
| Sanger Centre | Sequence for 1p21 | http://www.sanger.ac.uk |
| Draft Genome Browser | Sequence for 1p21 | http://genome.cse.ucsc.edu/ |
| BAC & Accession maps | Sequence for 1p21 | http://genome.wustl.edu/gsc/human/mapping/ |

Table 2.5 Sequence handling software

| Name | Purpose | Reference |
|---|---|---|
| Sequencher v.4 | Sequence alignments, identifying restriction enzyme sites | Gene Codes, Ann Arbor, Michigan |
| ClustalX | Multiple sequence alignments | Thompson et al. (1997) |
| Blast | Finding sequence matches in genome databases | http://www.ncbi.nlm.nih.gov/BLAST/ |
| etandem | Searching for repeat sequences | HGMP EMBOSS telnet://tin.hgmp/mrc.ac.uk |
| Oligo v4 | Primer design | MBI, Cascade, CO |
| PAUP*v4 β10 | Phylogenetic analysis | Sinauer Associates, Sunderland, Mass. |

Table 2.6 Software used in population genetics analysis

| Name | Purpose | Reference / Supplier |
|---|---|---|
| Instat | Summary statistics, ANOVA | GraphPad, San Diego, CA |
| Arlequin | $F_{ST}$, AMOVA, Exact test of population differentiation | Schneider et al. (2000) |
| Genstat v3.2 | Principle Co-ordinate vectors | VSN, Hemel Hempstead UK |
| MATLAB Programming environment | Implement functions written by M. Weale for EMamy | Mathworks, Natick, MA (see http://www.ucl.ac.uk/tcga for functions) |
| 'R' | Statistics package | http://www.R-project.org/ |
| Generic Mapping Tools | Contour maps | Wessel & Smith (1998) http;//gmt.soest.Hawaii.edu |
| SYSSIPHOS | Analysis of Intra-allelic variability | M. Stumpf (unpublished) |
| YTIME | Estimating the age of alleles | Behar et al (2003) http://www.ucl.ac.uk/tcga |

## 2.10 Miscellaneous

### *2.10.1 Suppliers*

Unless stated in the text the following companies were suppliers of laboratory consumables for this thesis:

Sigma-Aldrich, St Louis, Missouri (General)

Fisher Scientific, Loughborough, Leicestershire UK (General)

Fissons Scientific Equipment, Loughborough, Leicestershire UK (General)

Merck BDH Chemicals, Poole, Dorset, UK (General)

Sartstedt, Numbrecht, Germany (101x16.5mm transport swab tubes)

New England Biolabs, Beverly, MA (Restriction enzymes)

MWG Biotech Ebersberg, Germany (Oligonucleotides)

HT Biotech, Cambridge, UK (Taq polymerase)

BD Biosciences Clontech, San Jose, CA (Mab)

Whatman BioMetra, Goettingen, Germany (PCR machine)

National Diagnostics, Atlanta, Georgia, (Acrylamide solution)

Microzone Ltd, Haywards Heath, UK (Sequencing reagents)

Invitrogen, Carlsbad, CA (Cloning kit)

PE-Applied Biosystems, Foster City, CA (PCR machines, DNA sequencing & GeneScan™ equipment)

*2.10.2 Units*

All values measured in this thesis use SI units.

*2.10.3 Buffers and reagents*

All pH values at 25°

| Buffer Name | Abbreviation | Contents |
|---|---|---|
| Tris-EDTA Buffer | 1xTE | 1mM EDTA, 10mM Tris-HCl ph 8.0 |
| Tris-borate-EDTA Buffer | 1xTBE | 0.09M Tris-borate pH8.0, 2mM EDTA, pH 8.3 |
| NEB Buffer No.2 for restriction endonucleases | NEB 2 | 10 mM Tris-HCl, 10mM $MgCl_2$, 50mM NaCl, 1mM dithiothreitol, pH 7.9 |
| pCR®4-TOPO® plasmid vector | pCR®4-TOPO® | 10ng/µl plasmid DNA in 50% glycerol, 50mM Tris-HCl pH 7.4, 1mM EDTA, 2mM DTT, 0.1% Triton X-100, 100 µg/ml BSA, 30µM phenol red. |
| SOC Medium | SOC | 2% Tryptone, 0.5% Yeast extract, 10mMNaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10mM $MgSO_4$, 20 mM glucose. |

# Chapter 3: Developing protocols for typing human DNA samples for AMY1 repeat alleles and six closely linked microsatellites

## 3.1 Introduction

The amylase multi-gene gene family consists of 2 pancreatic genes (AMY2A and AMY2B) and variable number of salivary genes (Groot et al. 1989*a*) (See Table 3.1). Three salivary amylase genes (AMY1A,B & C) and a pseudogene (AMYP1) have been described and are located in tandem on the short arm of chromosome 1 (Tricoli & Shows 1984). Extensive quantitative variation has been demonstrated in Caucasian populations in the form of polygenic repeats of the AMY1 genes as follows: 2B-2A-$(1A\text{-}1B\text{-}P1)^n$-1C (Bank et al. 1992) (see Fig 1.9). The first aim of this thesis was to design PCR based protocols to genotype and, haplotype large numbers of individuals for the salivary amylase polygenic repeat alleles as well as a number of closely linked microsatellites. This chapter describes the development of these protocols.

As was outlined in chapter 2, two different approaches were explored whilst developing the method for AMY1 quantification. Both methods were based on quantifying the variable number of salivary amylase genes, through a comparison with the pancreatic amylase genes, which are constant in number. Thus both schemes were based around a semi-quantitative PCR that involved co-amplification of target sequences from both the AMY2 and AMY1 genes. In addition, both approaches used the ABI 377 / GeneScan™ genotyping system. This technology not only provides accurate sizing of DNA fragments, it can also be used to quantify the relative quantities of fragments.

In order to ensure equal amplification efficiency of the AMY1 and AMY2 target sequences it was important the PCR products were the same or very similar length (See Hirano 2002, Arezi 2003). However if the PCR products from both the AMY1 and AMY2 genes are the same length then an additional method must be used to distinguish them. The first approach used PCR primers that amplified a region in

Table 3.1 Nomenclature of Human alpha-amylase genes, haplotypes and genotypes after Groot et al (1989). The gene designations for the amylase genes were introduced by Gumucio et al (1988) and are in agreement with the guidelines for human gene nomenclature (Shows 1987). AMY1 repeat alleles here refers to the number of 100kb repeat units containing AMY1A, AMY1B and AMYP1 that are present on each chromosome.

| Item | Abreviation |
|---|---|
| Salivary amylase genes | AMY1A; AMY1B; AMY1C |
| Pancreatic amylase genes | AMY2A; AMY2B |
| Amylase pseudogenes | AMYP1 |
| Salivary amylase repeat alleles | AMY1*H0; AMY1*H1; AMY1*H2;<br>AMY1*H3;<br>AMY1*H4; |
| Salivary amylase genotypes | AMY*H0/H0<br>AMY*H0/H2 etc |

both AMY1 and AMY2, which also contained a restriction endonuclease recognition site in either the AMY1 or the AMY2 genes. If the restriction endonuclease recognition sequence was contained in the AMY1 product, the AMY1 product would be cut into two shorter fragments, thus distinguishing them from the AMY2 product.

The second approach was based on an update of Bank and colleagues (1992) protocol. This method used primers that amplify an area around an insertion in the AMY1 sequence producing a longer AMY1 specific product and a shorter AMY2 specific product. The protocol described here used very small deletions (maximum 4bp) to minimise the risk of unequal amplification efficiency between the AMY1 and AMY2 products. Experiments comparing the accuracy of the two approaches in AMY1 quantification found that the second approach was the most reliable in assigning genotype.

## 3.2 Designing an assay for AMY1 Quantification

### *3.2.1 Previous methods of AMY1 repeat number quantification: Bank et al. (1992)*

Bank et al. (1992) used a PCR based method to quantify the relative amounts of salivary amylase (AMY1) PCR product to pancreatic (AMY2) PCR products. Primers were designed to amplify a region around a 22bp insertion present in AMY1, which is absent in AMY2. It was therefore possible to use the ratio of AMY1 PCR products to AMY2 products to quantify the number of AMY1 genes present in the individual (See Table 3.2). The PCR primers were designed to anneal equally well to both the AMY1 genes and the AMY2 genes, by having a sequence complementary to regions that are identical in all AMY genes. The PCR results in an AMY1 specific product of 604bp and an AMY2 specific product of 582bp. The PCR reaction used radioactively labelled dNTPs for detection, and PCR products were separated by polyacrylamide gel electrophoresis (PAGE). Quantification was performed using a LKB2202 Ultrascan densometer to analyse the bands on autoradiograms. To ensure accuracy and reproducibility, PCR products were run twice and an average taken.

Table 3.2: Expected ratios of AMY2:AMY1 PCR products for Bank et al (1992) protocol. The ratio of AMY2:AMY1 PCR products can be used to determine the number of AMY1 genes in an individual. However it must be noted that in some cases different genotypes contain the same number of AMY1 genes and so will give the same AMY2:AMY1 ratio.

| Ratio of PCR products AMY2B+AMY2A: AMY1 | Total number of AMY1 genes in an individual | GENOTYPE(S) |
|---|---|---|
| 4:2 | 2 | H0/H0 |
| 4:4 | 4 | H0/H1 |
| 4:6 | 6 | H0/H2 or H1/H1 |
| 4:8 | 8 | H0/H3 or H1/H2 |
| 4:10 | 10 | H0/H4 or H1/H3 or H2/H2 |
| 4:12 | 12 | H0/H5 or H1/H4 or H2/H3 |

*3.2.2 Potential Modifications to Bank et al. (1992) method for AMY1 quantification*

Bank et al. (1992) use a PCR based assay that amplifies a region around a 22bp insertion in AMY genes, which is not present in AMY2 genes, resulting in two PCR fragments of different length. There is usually a negative relationship between the amplification efficiency and the length of the fragments that are amplified (Arezi et al. 2003). Thus the AMY1 product would have a lower amplification efficiency than the AMY2 product. In order to minimise the risk of unequal amplification efficiency the two PCR products should be more similar or the same in length. As outlined in the introduction two solutions to this problem were explored. Firstly, amplifying regions from both AMY1 and AMY2 of the same length that are distinguished by the presence or absence of a restriction enzyme site. Secondly amplifying regions around much smaller insertion / deletions in either AMY1 or AMY2.

Another improvement to Bank et al.'s (1992) protocol would be to develop more than one PCR based assay so that a correction could be applied if any differences in the efficiency of the PCR were identified. Two assays could be developed where the gene most efficiently amplified in the first is least efficiently amplified in the second. An example of this would be to design a PCR protocol that amplified a region around an insertion in AMY1. This would produce an AMY1 specific product that is longer than the AMY2 specific product. In addition to this protocol, a PCR could also be used that amplified a region around an insertion in AMY2 that is not present in AMY1. This would produce a longer AMY2 specific product. The combination of both assays would provide a means of identifying and correcting any PCR based inefficiencies.

Table 3.2 shows that in some cases different AMY1 genotypes contain the same number of AMY1 genes and so will give the same AMY2:AMY1 ratio. Bank et al.'s (1992) method can only determine the total number of AMY1 genes in an individual and not the way that these genes are apportioned between the maternal and paternal chromosomes. This information is important in constructing compound haplotypes

Table 3.3: Possible combinations of AMY1 repeat alleles. Numbers in the table indicate total number of AMY1 genes present in the individual. It is important to note that in some cases several AMY1 repeat allele combinations produce the same total number of AMY1 genes.

| Paternal Chromosome / Maternal Chromosome | AMY1 *H0 | AMY1 *H1 | AMY1 *H2 | AMY1 *H3 | AMY1 *H4 | AMY1 *H5 |
|---|---|---|---|---|---|---|
| AMY1*H0 | 2 | | | | | |
| AMY1*H1 | 4 | 6 | | | | |
| AMY1*H2 | 6 | 8 | 10 | | | |
| AMY1*H3 | 8 | 10 | 12 | 14 | | |
| AMY1*H4 | 10 | 12 | 14 | 16 | 18 | |
| AMY1*H5 | 12 | 14 | 16 | 18 | 20 | 22 |

of AMY1 repeat alleles and closely linked microsatellite alleles for an analysis of intra-allelic variability. Consequently an additional method would need to be designed to resolve the structural arrangement of the AMY1 genes along the chromosome. For example, a diagnostic assay for AMY1*HO would distinguish between AMY1*HO/H2 and AMY1*H1/H1.

A final modification to Bank et al.'s (1992) assay would be the replacement of radioactively labelled dNTPs visualised with autoradiography, with fluorescent based detection methods.

After having reviewed the improvements that could be made to Bank et al. (1992) method, two separate approaches were designed and tested experimentally to determine the best method for AMY1 quantification. The first of these methods amplified AMY1 and AMY2 specific fragments that are of the same length, but with a small number of base changes, and used restriction endonucleases to distinguish AMY1 products from AMY2 products. The second method was an updated version of Bank et al. (1992) protocol and amplifies and areas around small insertion/deletions producing AMY1 & AMY2 specific products that differ in length by a maximum of 4 base pairs.

### 3.3 Obtaining sequence information for 1p21

At the outset of this work, the draft human genome sequence had not been completed and there were still some major gaps in the assembly for 1p21. Sequence information for the amylase gene family was required in order to design the PCR based assays for AMY1 quantification and also search for microsatellite markers closely inked to the amylase gene cluster. Once suitable microsatellite markers had been located, the sequence alignments were used design a multiplex PCR/ GeneScan based protocol to type for microsatellite repeat length variation (See Section 3. 6).

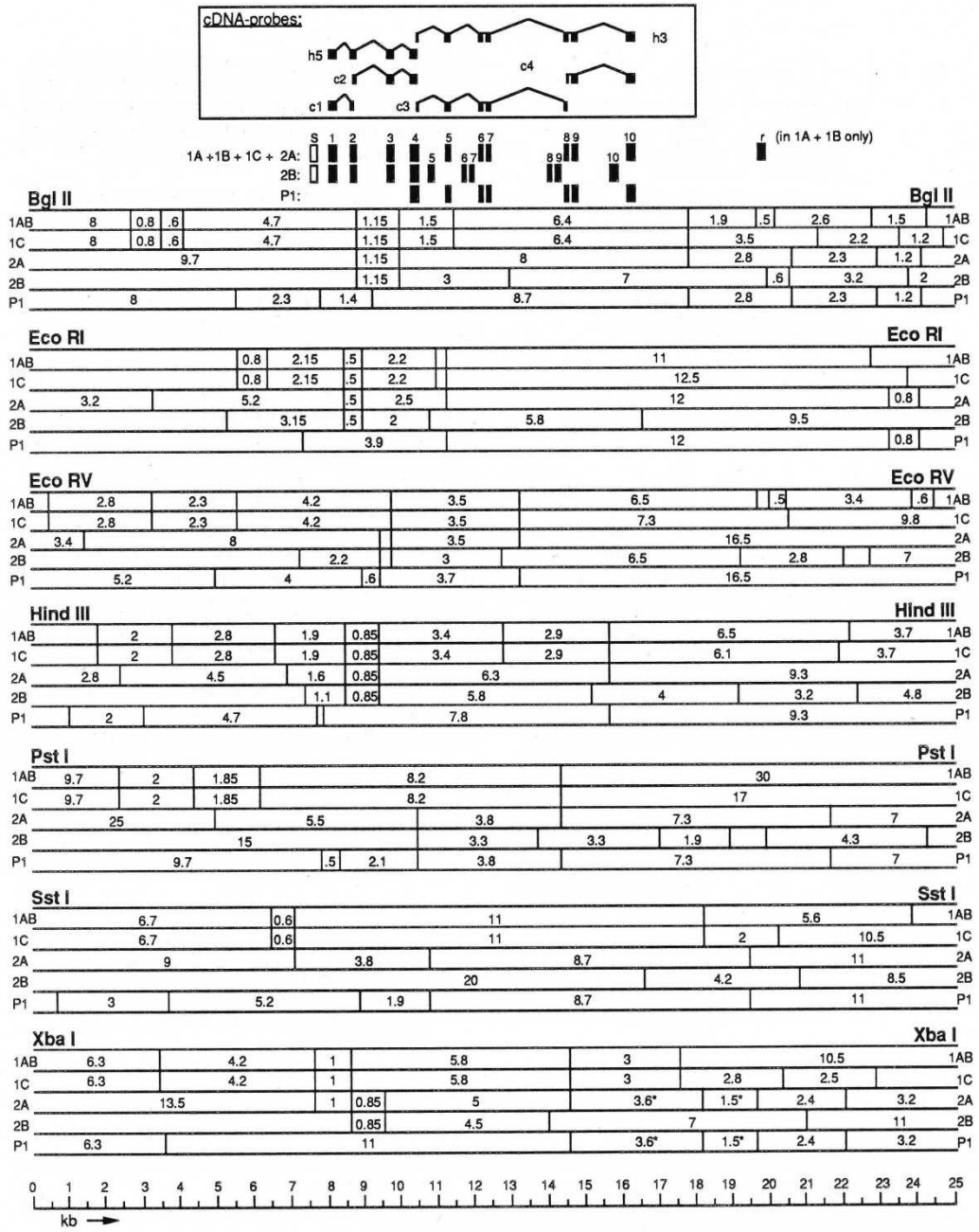



Fig 3.2: Restriction enzyme maps of human amylase genes showing differences in genomic clusters, after Groot et al (1989). 1AB = AMY1A or AMY1B; 1C = AMY1C; 2A = AMY2A; 2B = AMY2B; P1 = AMYP1; R = DNA region in AMY1A and AMY2B hybridising with probe c2.

In the absence of a reliable sequence assembly for this region on the human genome working draft, an attempt was made to construct an assembly using the sequences that were available (see Table 3.4). Firstly, exon sequences from the pancreatic (AMY2A and AMY2B) and salivary (AMY1A, 1B, 1C) genes, as well as from the pseudogene (AMYP1), were downloaded from GenBank (www.ncbi.nlm.nih.gov/Web/GenBank/). Following this, the sequence from four unfinished BAC clones that span the region were obtained from the UCSC Human Genome Working Draft (Nov 2000 assembly) and aligned, along with the amylase exon sequences, using Sequencher v.4 (Gene Codes, Ann Arbor, Michigan). This served to identify the positions of amylase genes within the clones, as well as to build contigs of the clone fragments (see Table 3.4). Larger pieces (>2kb) of BAC clone were compared to the published restriction maps (see Groot et al. 1989*a*) (Fig3.1**)** so that they could be plotted onto a map of the region, and contigs constructed (see Fig 3.2). The sequence from the contigs were used to design primers for AMY1 quantification and microsatellite protocols.

The extremely high degree of similarity between all the salivary amylase genes has caused problems for sequencing and mapping the 1p21 region (S. Gregory, The Wellcome Trust Sanger Institute, Cambs., *pers. comm.*) Alignments of this region produce assemblies that superimpose the sequences from the three genes AMY1A, AMY1B and AMY1C, and interpret them as multiple sequences from a single salivary amylase gene. Thus 1p21 was severely truncated in many of the early assemblies of the region (See Fig 3.3). Despite the recent advances that have been made in the finishing of the 1p21 region, it is still not known if the AMY1 100kb repeats are absolutely contiguous to one another, and indeed if there are any distinguishing features between the sequence of the intergenic regions between AMY2A and AMY1C genes in the AMY1*H0 allele and AMY2A and AMY1A in the AMY*H1 allele (see fig 1.9). This information would be useful for designing

Table 3.4**:** Name and Accession number of BAC clones from Nov 2000 UCSC Genome browser assembly.

| Clone Name | Accession Number | Length (bp) | Number of pieces | Amylase genes covered |
|---|---|---|---|---|
| RP11-727M5 | AC0255933 | 189,612 | 65 | 2B, 2A, 1A, 1B, P1, 1C |
| RP11-9N17 | AC013599 | 128,792 | 27 | 2B, 2A, 1A, 1B, P1 |
| RP11-259N12 | AC026662 | 82,294 | 7 | 2A, 1A, 1B, P1, 1C |
| RP5-1108M17 | AL356363 | 114,392 | 12 | 2B, 2A, 1C |

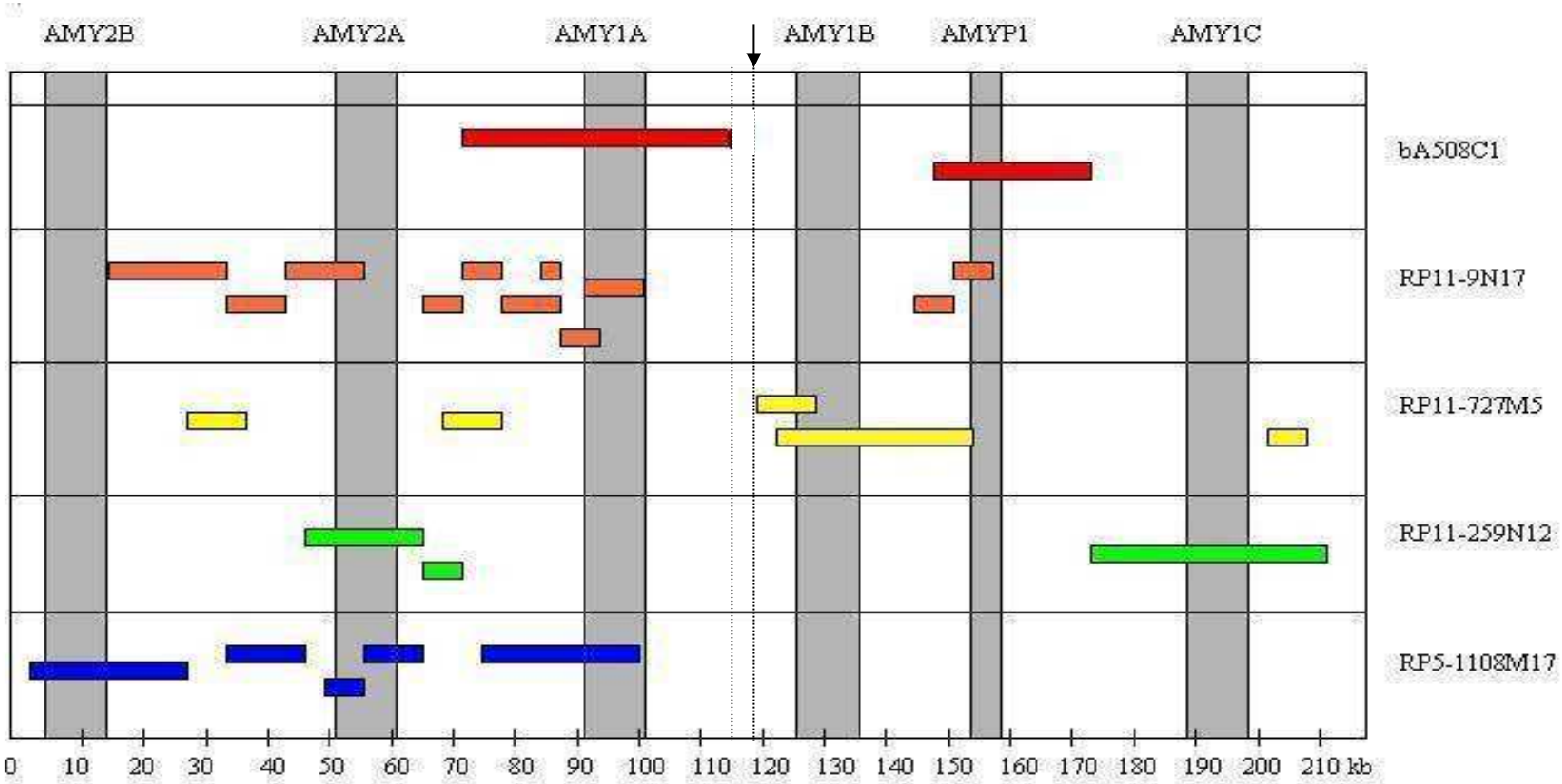



Fig 3.2: A grid showing the relative positions of the amylase genes and pieces of 5 BAC clones that span the 1p21 region as plotted by comparison to restriction maps of the amylase gene cluster published by Groot et al (1989). Clone fragments are taken from the Nov 2000 assembly of the UCSC human genome draft. bA508C1 was added from the April 2001 assembly. Note the absence of contiguous coverage between AMY1A and AMY1B (shown by the black arrow and dotted lines).

PCR based assay to detect the apportionment of the AMY1 genes between the maternal and paternal chromosomes.

## 3.4 Using restriction enzymes in AMY1 quantification

### *3.4.1 Principles of the protocol*

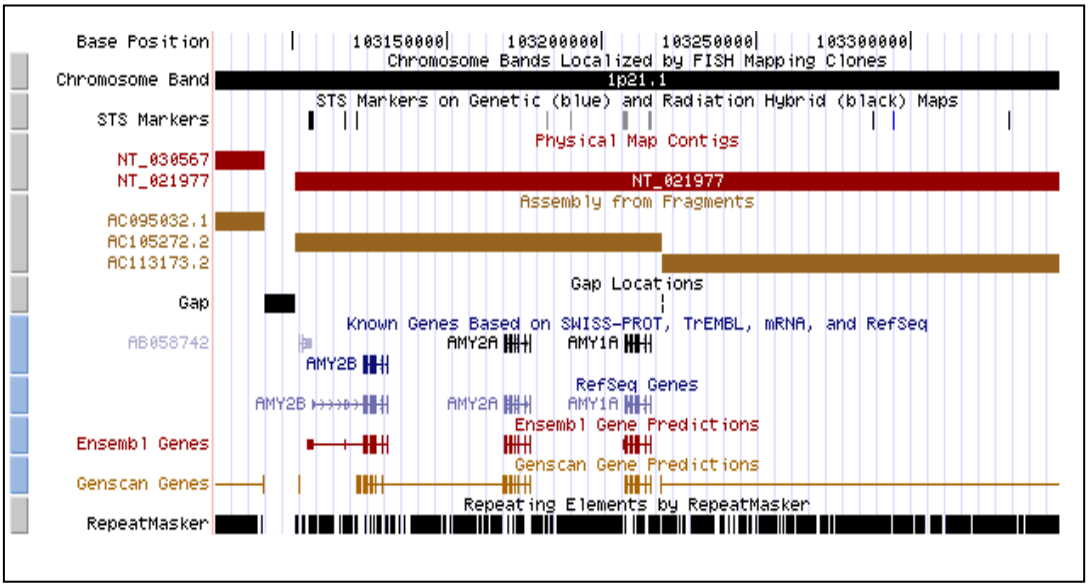
In this approach, PCR protocols were designed to amplify regions of the same length from both AMY1 and AMY2 genes but have small difference in their sequence that can be distinguished by the presence (or absence) of a restriction enzyme recognition site. Two different systems were designed: the first primer pair (AMY02) amplified a region around a restriction site present in AMY1, but absent in AMY2 (See Fig 3.4). The restriction enzyme would cut the AMY1 products but not the AMY2 products, resulting in fragments of different sizes. The longer uncut fragments would be specific to AMY2 products, where as the shorter cut fragments would have resulted from AMY1 PCR products. The ratio of cut to uncut products could then be determined. The second primer pair (AMY04) amplified an area around a restriction site of the same restriction enzyme as used with the first primer pair, present in AMY2 but not in AMY1. The results obtained from both systems should give identical results. However, if the restriction enzyme does not cut to completion then the combination of the two systems would provide a way of correcting for the inefficiency of the enzyme.

### *3.4.2 Assay design and optimisation*

Exon sequences from all the amylase genes were aligned against each other using ClustalX (Thompson 1997). Base changes between AMY1 and AMY2 genes were identified and tested to see whether they formed a recognition site for a common restriction enzyme in either the AMY1 or AMY2 genes. As there were a shortage of suitable restriction sites mismatch primers were designed to force a mutation in the PCR product, which would result in an enzyme recognition site.

Primers were designed to anneal to regions that are identical in both AMY1 and AMY2 so as to ensure equally efficient amplification of both genes. Suitable

UCSC Human Genome Browser November 2002




UCSC Human Genome Browser July 2003

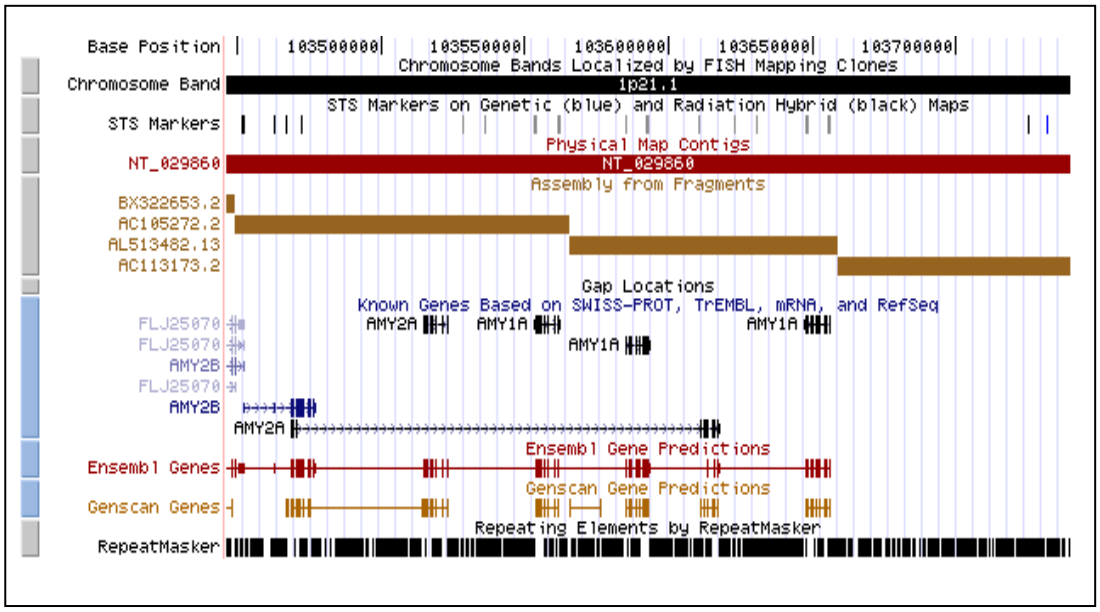



Fig 3.3: Two freizes from the UCSC Human Genome Working draft browser. The November 2002 freeze shows both pancreatic amylase genes (AMY2B & AMY2A) but only one salivary amylase gene (AMY1A). The July 2003 freeze has two additional salivary genes, which correspond to AMY1B and AMY1C. The right hand end of the additional elongated AMY2A gene corresponds to AMYP1.

primers were chosen to optimise for compatibility of annealing temperature, elimination of false priming sites, primer dimers, high 5' stability and low 3' stability using the primer design software Oligo v4.01 (MBI Cascade, CO). Low 3' stability is required to reduce the probability of false priming in non-target areas of the genome and in PCR products.

It was intended that both the primers pairs (AMY02 & AMY04) should be combined into a multiplex PCR reaction. Multiplex PCR systems have been widely used to increase the throughput and decrease the cost of typing large numbers of both SNPs and microsatellite loci (Thomas et al. 1999, Fletcher et al. 2003). Multiple pairs of PCR primers are added to a common reaction mixture so that multiple regions of the genome can be amplified at the same time. Multiplex PCR protocols however often require considerable optimisation. PCR primers must however be designed to remove the possibility of false priming elsewhere in the genome, as well as to reduce the probability of primers pairs forming 3' dimers. All the primer pairs must also have similar annealing temperatures, and the resulting PCR products should have different enough lengths so that the products from the various primer pairs will be easy to distinguish when using fluorescent based detection systems. Primers can also be distinguished with different colour fluorescent dye labels (HEX, TET, and FAM). The resulting PCR products must all within a limited size range (75-2,500 bp) to be suitable for GeneScan™ analysis.

The PCR reactions were tested experimentally to find the optimal annealing temperature, primer concentrations, $MgCl_2$ concentration and for use with DNA extracted from buccal swabs. Once primers were optimised in single PCRs, primers were tested in multiplex, initially at equal concentrations (0.2 μM). Following this, the concentrations of individual primer pairs were tested at a range of concentrations from 0.06μl to 0.5μl to achieve the optimal amplification in terms of minimising the amount of primer used to produce relatively equal amounts of PCR product for both markers. Final reaction conditions for PCR, restriction enzyme digestion and electrophoresis are described in sections 2.5.2 & 2.5.6.

Fig 3.4: Using restriction enzymes in AMY1 quantification: AMY02 and AMY04 systems

**AMY02**

Primers in bold. Labelled primer (AMY 02-U-HEX) highlighted in yellow. Restriction enzyme recognition site underlined and highlighted in green. Differences in AMY1 and AMY2 sequence highlighted in red.

AMY1 PCR Product:
82bp: **AAAGGCAATTTTGGACAAACTG**CATAATCTAAACAGTAACTGGTTCCCGGAAGGTAG**TAAACCTTTCATTTACCAGGAGGTA**
AMY2 PCR Product:
82bp: **AAAGGCAATTTTGGACAAACTG**CATAATCTAAACAGTAACTGGTTCCCTGCAGGAAG**TAAACCTTTCATTTACCAGGAGGTA**

AMY2 PCR Products after lysis with *Pst1* (AMY1 PCR product not cut):
52bp: **AAAGGCAATTTTGGACAAACTG**CATAATCTAAACAGTAACTGGTTCCCTGCA

30bp: GGAAG**TAAACCTTTCATTTACCAGGAGGTA** (not detected by GeneScan™ as primer not fluorescently labelled)

**AMY04**

Primers in bold. Labelled primer (AMY-04-U-FAM) highlighted in blue. Restriction enzyme recognition site underlined and highlighted in green. Differences in AMY1 and AMY2 sequence highlighted in red. Note that penultimate base (highlighted purple) of the lower primer (AMY-04-MML-B) is mismatched (there is a C in both the AMY1 and AMY2 genomic sequences) and so creates the restriction enzyme recognition site in the AMY1 PCR product.

AMY2 PCR Product:
83bp: **GTCTTCCTGCTGGCACATACT**GTGATGTCATTTCTGGAGATAAAATTAATGGCAACTGC**AGAGGCATTAAAATCTACGTTTCT**

AMY1 PCR Product:
83bp: **GTCTTCCTGCTGGCACATACT**GTGATGTCATTTCTGGAGATAAAATTAATGGCAACTGC**AGAGGCATTAAAATCTACGTTTCT**

AMY1 PCR Products after lysis with *Pst1* (AMY2 PCR product not cut):
60bp: **GTCTTCCTGCTGGCACATACT**GTGATGTCATTTCTGGAGATAAAATTAATGGCAACTGCA

23bp: **GAGGCATTAAAATCTACGTTTCT** (not detected by GeneScan™ as primer not fluorescently labelled)

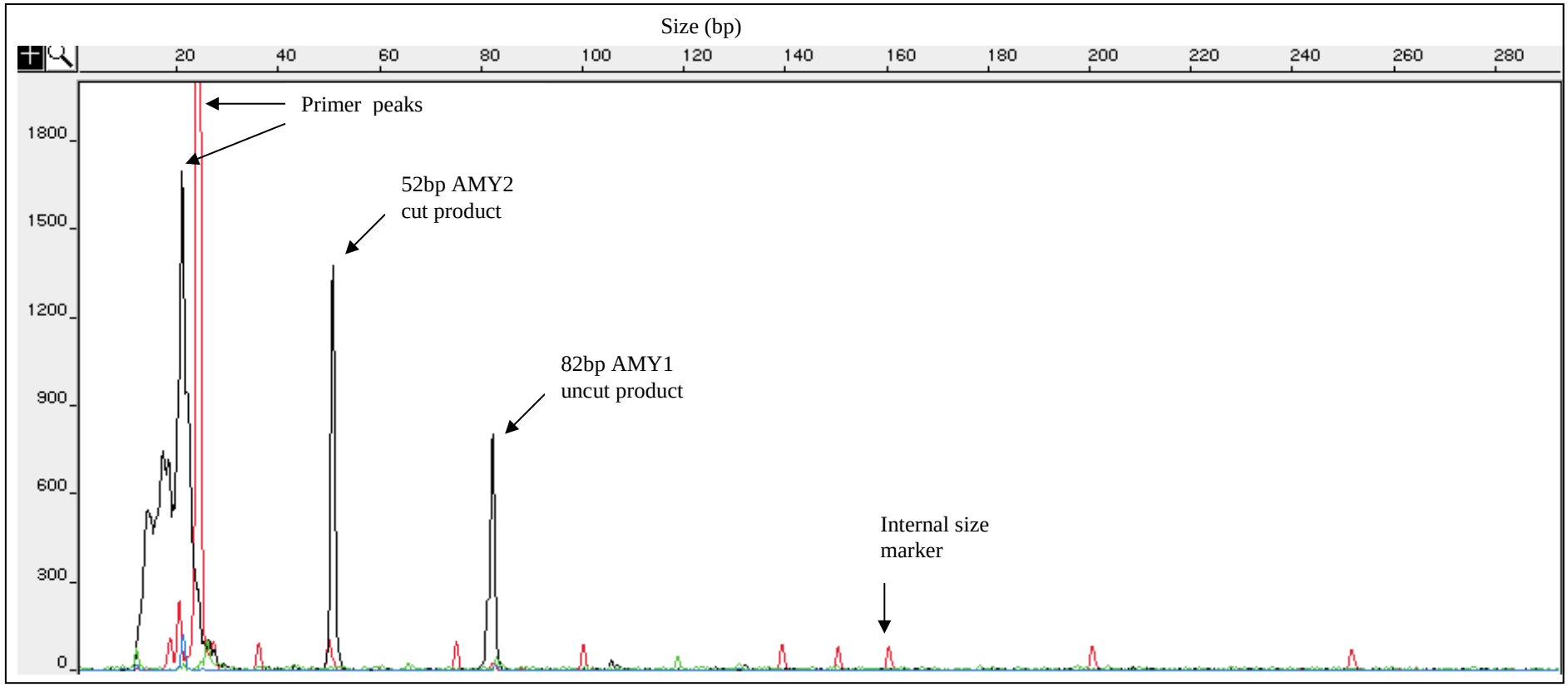



Fig 3.4b: A GeneScan™ output for the AMY02 protocol. The peaks shown correspond to fluorescence detected by the laser system in the ABI377. This display shows two black peaks which correspond to the cut (52bp, AMY2) and uncut (82bp, AMY1) AMY02 protocol PCR products. In addition, the smaller red peaks represent the internal size marker.

GeneScan™ software produces graphical representations of size and fluorescent label as peaks of fluorescence, which are relative to the amount of PCR product present (See Fig 3.4b). In order to estimate the relative number of molecules of two different sized PCR products, the ratio of respective peak areas or heights must be calculated. Experiments to find the most reliable measure of the amount of PCR product were performed. It was found that recording peak height not only gave a value closer to the expected value for the known genotype samples, but also had a lower variance across runs (see fig 3.5) As a result of these experiments, the heights of peaks corresponding to the PCR products were recorded and the ratio of peak heights from the AMY1 and AMY2 products were calculated.

DNA samples from six Dutch individuals of known AMY1 repeat allele genotype (see Groot et al. 1989*a*) were obtained from Prof. Jan Pronk, Vrije Universiteit, Amsterdam. These samples were used to test the accuracy of the protocol at determining genotype.
The ratio of AMY2:AMY1 products obtained experimentally, was compared to the expected ratios (see Table 3.5) and genotype was assigned. However, there was still an excess of uncut PCR product, compared to expected results. This was interpreted as a failure of the restriction enzymes to completely digest all the PCR products which contained cut sites. Condition for the restriction enzyme lysis conditions were optimised for buffer composition, amount of PCR product, and units of enzyme. Following this, a mathematical correction was applied to remove the effects of incomplete digestion as follows:

The system comprises of two primer pairs which each amplify a regions in both AMY1 and AMY2 around a recognition sites for the same restriction enzyme. The first primer pair amplifies a region around a recognition site that is present in AMY1 but not in AMY2. Thus the restriction enzyme cuts the AMY1 PCR product but not the AMY2 product.

Let P1 = amount AMY1, P2 = amount of AMY2A+B, and E = Enzyme efficiency,

So that only a proportion E of AMY1 cuts:

Observed fraction of cut/(cut+uncut) products,

$F1 = EP1 / (EP1 + P2 + (1-E)P1$

$= EP1 / (P1+P2)$

This enzyme also cuts at another site in another PCR product, from a different location in the gene. However, in this case the restriction site is present in AMY2 but not AMY1. Assuming that this site is cut with the same efficiency as the site described above, then:

Observed fraction of cut/(cut+uncut) products,

$F2 = EP2 / ( EP2 + P1 + (1-E)P2)$

$= EP2 / (P1 + P2)$

E can be removed by dividing F1 by F2.

$F1/F2 = P1/P2$.

Despite the corrections for enzyme efficiency, the method still failed to produce reliable genotype assignments across multiple runs on the same sample of known AMY genotype.

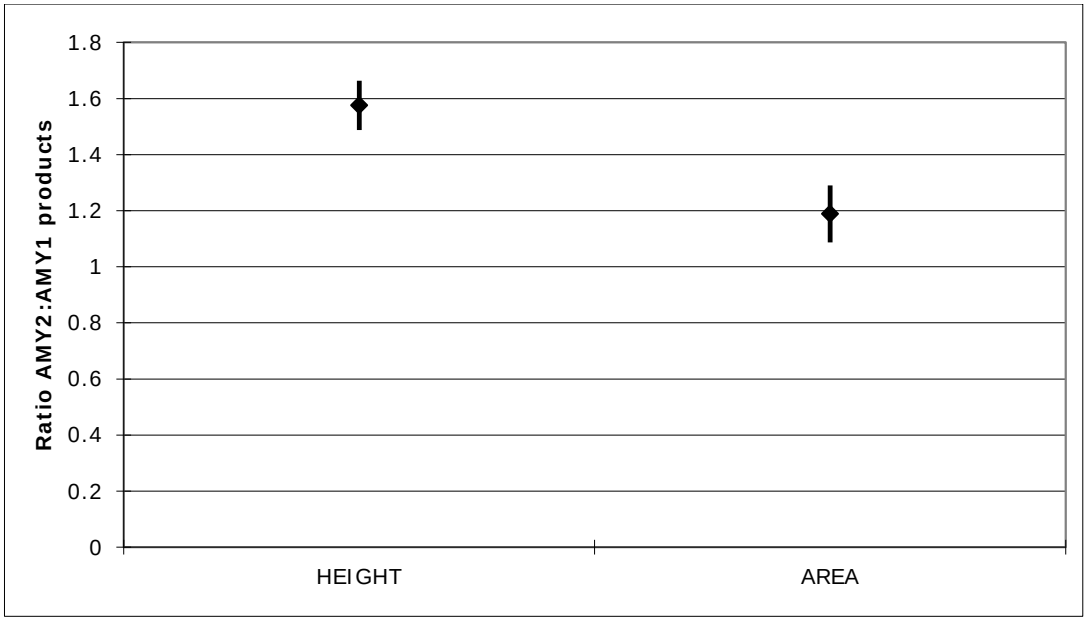



Fig 3.5: A comparison of the mean ratios of AMY2: AMY1 PCR products' fluorescence between the height and area of peaks of fluorescence measured using an ABI377/ GeneScan™ system. The mean values are black dots and error bars represent plus or minus one standard error of the mean. The data are from the restriction enzyme protocol for AMY1 quantification. 73 measurements of peak height were taken, and 47 of peak area, from separate electrophoresis runs from the same PCR reaction. The expected ratio of AMY2:AMY1 products for this individual is 2.0. As can be seen, the mean value for peak height is closer to the expected value for the DNA sample used. However it must be noted that both peak height and peak area produce lower than expected values using the restriction enzyme protocol for AMY1 quantification.

Table 3.5: Expected ratios of cut and uncut PCR products for AMY02 and AMY04. Numbers in table are given as number of genes present in individual with genotype shown.

| GENOTYPES | AMY02 CUT:UNCUT AMY2B+AMY2A+P1: AMY1 | AMY04 CUT:UNCUT AMY1: AMY2B+AMY2A+P1 |
|---|---|---|
| AMY*H0/H0 | 4:2 | 2:4 |
| AMY*H0/H1 | 5:4 | 4:5 |
| AMY*H0/H2 or AMY*H1/H1 | 6:6 | 6:6 |
| AMY*H0/H3 or AMY*H1/H2 | 7:8 | 8:7 |
| AMY*H0/H4 or AMY*H1/H3 or AMY*H2/H2 | 8:10 | 10:8 |
| AMY*H0/H5 or AMY*H1/H4 or AMY*H2/H3 | 9:12 | 12:9 |

*3.4.3 Possible sources of error in the restriction enzyme protocol*

There are a number of possible explanations for the failure of this protocol to produce accurate genotype assignments. Firstly, it is possible that the PCR may not have been amplifying the AMY1 and AMY2 products with equal efficiency thus giving inaccurate ratios of AMY1:AMY2 products. However, as the PCR products are distinguished through whether they are cut by the restriction enzyme, it was difficult to isolate the nature and degree of error at the PCR stage.

It is also possible that a high degree of heteroduplex formation during the PCR was reducing the efficiency of the restriction enzymes. Ruano & Kidd (1992) modelled heteroduplex formation during PCR from mixtures of human and chimpanzee DNA templates. They found that the degree of heteroduplex formation depends on the ratio of starting templates. When two templates are in equal concentrations a high degree of heteroduplex formation was found. To improve the AMY quantification protocol, the use of a nuclease enzyme such as *T7 Endonuclease I*, which cleaves non perfectly matched DNA, cruciform DNA structures, Holliday junctions and heteroduplex DNA, prior to the restriction enzyme digest was considered.

However, rather than adding another costly and time consuming step to the protocol, an alternative approach to AMY1 quantification was developed. This method was based on an update Bank et al.'s (1992) protocol that used PCR based method for AMY1 quantification that amplifies AMY1 and AMY2 specific fragments of different lengths, removing the need to use restriction enzymes for distinguishing the AMY1 and AMY2 fragments.

## 3.5 Updating Bank et al.'s (1992) method: the QAMY protocol.

*3.5.1 Principles of the protocol*

Initial experiments involved reproducing Bank et al.'s method, using the published primers and PCR conditions adapted for use with a fluorescent based detection system such as ABI377 / GeneScan™. However, as explained in section 3.2.2 there are two additional modifications to Bank et al.'s protocol that should improve the accuracy of the method. Firstly, in order to minimise

unequal amplification efficiencies between the AMY1 and AMY2 PCR products, it was decided to design protocols that amplified areas around small (4bp) insertion/deletions. Thus the AMY1 and AMY2 specific products could still be distinguished by length, but because the length difference is small, the potential for unequal amplification efficiency would be reduced.

Secondly, two different PCR systems were developed: The first system (QAMY02) amplified a region surrounding a 4bp deletion present in AMY2B, but not present in AMY1 and AMY2A. This gave rise to two fragments of different size, the longer one originating from AMY1 and AMY2A, and the shorter one originating from AMY2B. . The second system (QAMY03) was designed to amplify a region surrounding a 4bp deletion present in AMY1, not present in AMY2B or AMY2A. This system produced a longer fragment from AMY2 genes and a shorter fragment from AMY1 genes. Both protocols were expected give the same result from any one DNA sample. However one would expect that the longer fragment would have a higher chance of being amplified less efficiently than the shorter fragment (see Arezi 2003). As in one system the longer fragment was produced from AMY1+AMY2A, and in the other system it originated from AMY2 then it would be possible to correct for the unequal amplification of the fragments.

*3.5.2 Assay design and optimisation*

Sequences from all the amylase genes were aligned against each other, using ClustalX software (Thompson et al. 1997), and searched for insertion/deletions that would provide a means of distinguishing between the AMY1 and AMY2 genes. Small (<6 bp) insertion/deletions that are present in either AMY1 or AMY2 were identified from the aligned amylase gene sequences.

Primers were designed to amplify regions around these insertion/deletion sites the criteria outlined in section 3.4.2. (See Fig 3.7a). Primers were designed to anneal to regions that are identical in both AMY1 and AMY2 so as to ensure equally efficient amplification of both genes. The PCR reactions were optimised for primer concentration, $MgCl_2$ concentration and annealing temperature. The

PCRs for QAMY02 & QAMY03 were conducted as two separate single reactions instead of in a multiplex reaction, in order to maximise the efficiency of the PCR. Once the PCRs had been carried out the PCR products from the two reactions were mixed together and diluted for electrophoresis with the ABI377 / GeneScan$^{TM}$ system.

In order to check that the deletions that the QAMY primers were designed around were real and neither due to sequencing errors or incorrectly interpreted by the computer alignment program, the PCR products were sequenced using the protocol described in section 2.4. The AMY1 and AMY2 PCR products were sequenced in the both the forwards and reverse direction so that the presence of the deletions could be ascertained without the need to separate the two different length fragments. The results of the sequencing confirmed the presence of deletions for both the QAMY02 and QAMY03 systems.

With single target PCRs, quantification results are reliable only when analyses are performed at points in the exponential phase of the PCR amplification curve, before the onset of the plateau phase (Crotty et al. 1994, Jung et al. 2000). The reasons for this are poorly understood but are often attributed to one or more of the key PCR reagents being consumed, effectively halting the reaction. However, many people have found that co-amplification of different concentrations of different targets results in retention of the initial proportions even in the plateau phase (Morrison & Gannon 1994, Hirano 2002). To ensure that quantification results are reliable and elucidate the optimum number of PCR cycles for quantification, the QAMY02 and QAMY03 systems were tested on the samples of known genotype obtained from Prof Jan Pronk, Vrije Universiteit, Amsterdam. It was found that the optimum number of PCR cycles was 30, which is after the end of the exponential phase of the PCR reaction. See Fig 3.6

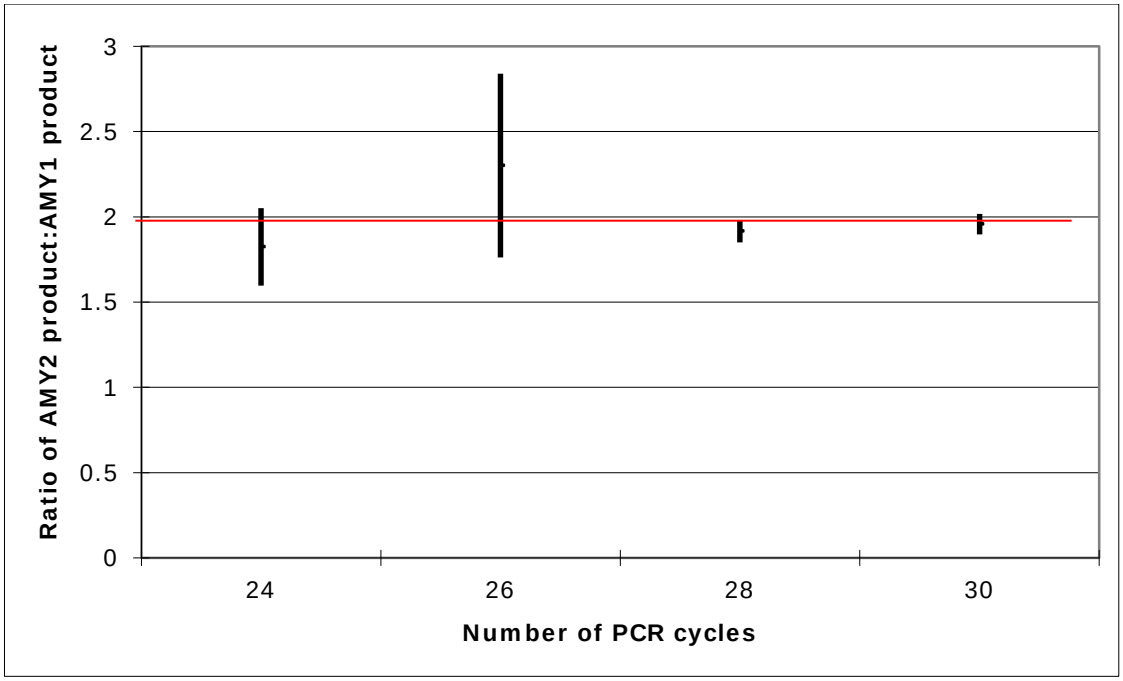


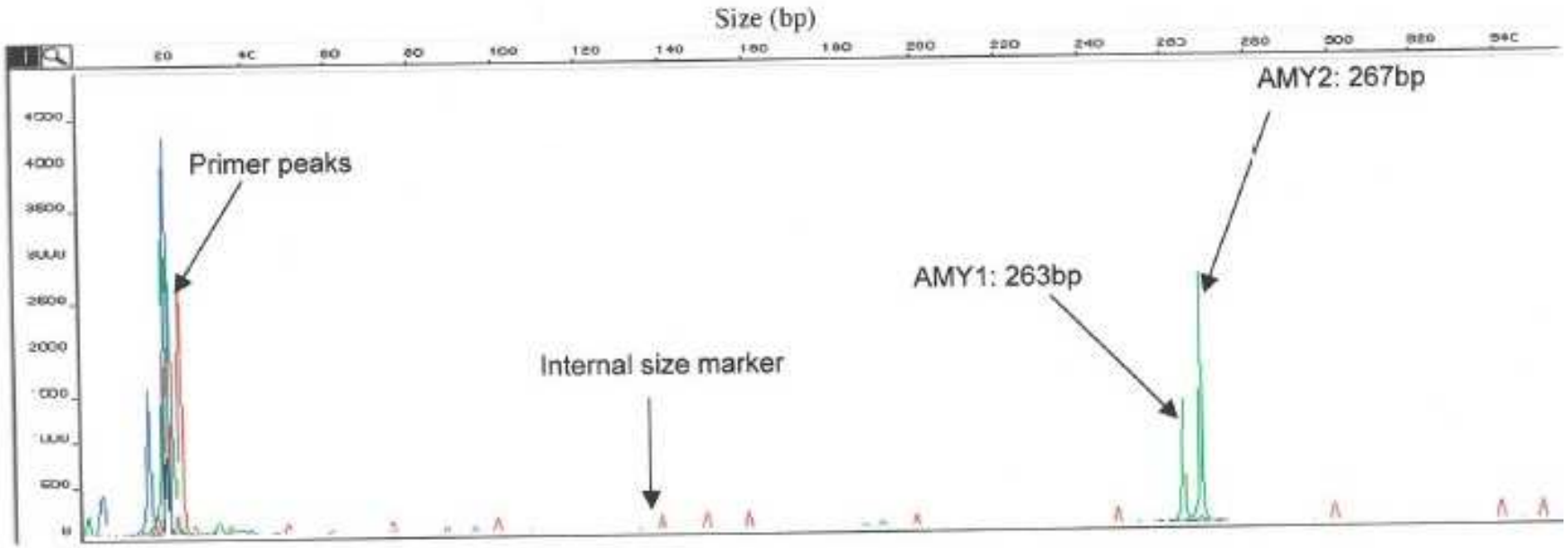

Fig 3.6: Mean ratio of AMY2:AMY1 PCR products from QAMY02 protocol for different numbers of PCR cycles. For all four cycle conditions (24,26,27 & 30), 4 PCRs were carried out and the mean and standard error of the ratios obtained from the PCRs calculated. The mean is shown by the black dot, and the error bars represent plus or minus 1 standard error either side of the mean value. The red line represents the expected ratio of AMY1 product:AMY2 product for the DNA sample used. As can be seen 30 cycles produces both the mean value closest to the expected value, as well as the smallest standard error.

Fig 3.7a: Alignment of AMY1A, AMY2A and AMY2B gene sequences (+1516/1796 bp from start codon). QAMY03 primers are shown in yellow, QAMY03 4bp deletion in AMY1A is shown in red.

```
AMY1A     TCACAT TACTT TCCT TTCACAGT TGATT TTTGA TCTTG TAGG AAAAT AGTTA TAAGG TAT
AMY2A     TCACAT TACTT TCCT TTCAC AGTTG ATTTT TGATC TTGTA GGAAA ATAAT TATAA GATAT
AMY2B     TCACAT TACTT ---C TTCAC AGTTG ATTTT TGATC TTGTA GGAAA ATAGT TATAA GATAT
          ****** *****     ***** ***** ***** ***** ***** ***** ***  ****** * ***

AMY1A     ---GAAATAT TTTGG AATTT TATTA GCACA CTAT AAATT TA-AT CAATA ATTCT TTAAA T
AMY2A     CATGAA ATATT TTGG AGTTT TATTA ACATA CTATA AACTT GCATC AATAA TGCTT TAAAT
AMY2B     CATGAA ATATT TTGG AGTTT TATTA ACATA CTATA AACTT GAATC AATAA TGCTT TAAAT
             *** ***** **** * *** ***** **   ****** ** **   *** ***** * *** *****

AMY1A     TTCTGC CTCTC TGTA AGTCA CACTG AATTA GAAAC TTTGT TTTCT AGGTT CGTAT TTATG
AMY2A     TTCTAC CTCTC TGTA AGTCA CACTG AAGTA GAAAC TTTGT TTTCT AGGTT CGTAT TTATG
AMY2B     TTCTGC CTCTC TGTA AGTCA CACTG AAGTA GAAAC TTTGC TTTCT AGGTT CGTAT TTATG
          **** ****** **** ***** ***** ** ** ***** ****  ***** ***** ***** *****

AMY1A     TGGATG CTGTA ATTA ATCAT ATGTG TGGTA ATGCT GTGAG TGCAG GAACA AGCAG TACCT
AMY2A     TGGATG CTGTA ATTA ATCAT ATGTG TGGTA ACGCT GTGAG TGCAG GAACA AGCAG TACCT
AMY2B     TGGATG CTGTA ATTA ATCAT ATGTC TGGTA ATGCT GTGAG TGCAG GAACA AGCAG TACCT
          ****** ***** **** ***** ****  ***** * *** ***** ***** ***** ***** *****

AMY1A     GTGGAA GTTAC TTCA ACCCT GGAAGTAGG GACTT TCCAG CAGTC CCATATT CTGGA TGGG
AMY2A     GTGGAA GTTAC TTCA ACCCT GGAAG TAGGG ACTTT CCAGC AGTCC CATAT TCTGG ATGGG
AMY2B     GTGGAA GTTAC TTCA ACCCT GGAAG TAGGG ACTTT CCAGC AGTCC CATAT TCTGG ATGGG
          ****** ***** **** ***** ***** ***** ***** ***** ***** ***** ***** *****
```

Fig 3.7b: A GeneScan™ output for the QAMY03 protocol. The ratio of the peak heights from the AMY1 specific product (263bp) and the AMY2 specific product (267bp) as follows:

AMY2 (267bp) : AMY1 (263bp) = 2726 : 1376 = 1.98 : 1
This is rounded to 2:1 which indicates genotype AMY1*H0/*H0.



Details of final primer concentrations, PCR and electrophoresis conditions can be found in table 2.2 and sections 2.3.1 & 2.3.4. After electrophoresis on the ABI377, GeneScan™ outputs the resulting data as peaks of fluorescence (See Fig 3.7b). The heights of the peaks from the AMY1 and AMY2 products were recorded for each marker and the ratio of AMY2:AMY1 peak heights was calculated. This ratio was then compared to the expected ratios for QAMY02 & QAMY03 (see table 3.6) and a genotype was assigned.

*3.5.3 Confirmation Experiments*

The QAMY02 and QAMY03 systems were tested on the samples of known genotype to test the accuracy of the assay for determining AMY1 gene copy number in individuals. Initial experiments showed there was some variation in the ratios of AMY1:AMY2 peak heights obtained from multiple GeneScan™ runs the same PCR reaction. To investigate this, three PCRs were performed on DNA from the same individual of known genotype, and electrophoresis on an ABI377/GeneScan™ system was carried out 4 times on the products of each PCR reaction. The variance in peak height ratios between electrophoresis runs was then calculated.

The acceptable deviation from the expected ratio of AMY2:AMY1 products was set at +/- 0.2. This range represents the range within which genotype could be confidently assigned, leaving a large margin where ambiguous ratios that do not correspond to any known AMY1 genotype would be rejected as bad data (See Table 3.6). The number of electrophoresis runs required to ensure that 99% of the results fell within the acceptable range of the expected ratios was calculated as follows:

Table 3.6 PCR product ratios for AMY1 quantification protocols: QAMY02 and QAMY03 and corresponding genotypes.

Numbers in table are given as number of genes present in individual with genotype shown.

| GENOTYPES | Total number of AMY1 genes per individual | QAMY02 AMY2B: AMY2A+AMY1 | QAMY03 AMY2B+AMY2A: AMY1 |
|---|---|---|---|
| AMY1*H0/H0 | 2 | 2:4 | 4:2 |
| AMY1*H0/H1 | 4 | 2:6 | 4:4 |
| AMY1*H0/H2 or AMY1*H1/H1 | 6 | 2:8 | 4:6 |
| AMY1*H0/H3 or AMY1*H1/H2 | 8 | 2:10 | 4:8 |
| AMY1*H0/H4 or AMY1*H1/H3 or AMY1*H2/H2 | 10 | 2:12 | 4:10 |
| AMY1*H0/H5 or AMY1*H1/H4 or AMY1*H2/H3 | 12 | 2:14 | 4:12 |

Let V= variance across electrophoresis runs on an ABI377/GeneScan™ system

R = acceptable range either side of expected ratios for known AMY1 genotypes

C = Two tailed critical value so that 99% of data fell within the acceptable range

$$\text{Number of runs required} = \frac{V}{\frac{R}{C} \times \frac{R}{C}}$$

Table 3.7 shows the results of the confirmation experiment. For all three of the PCRs the number of electrophoresis runs required to ensure that 99% of the ratios fell within +/-0.2 of the expected ratios for the AMY1 genotypes was never more than two. As a result it was decided to perform two PCRs for both QAMY02 and QAMY03 on each DNA samples, and to run each PCR twice on an ABI377/GeneScan™ system.

*3.5.4 Comparison of the two approaches for AMY1 quantification*

Two different approaches for AMY1 quantification were explored, namely using restriction enzymes to distinguish between AMY1 and AMY2 PCR products as well as updating Bank et al.'s protocol, which uses small differences in length to differentiate the products originating from the AMY1 and AMY2 genes. Once optimised, these protocols were tested for accuracy at assigning genotype using the samples of known genotype provided by Prof Jan Pronk, Vrije Universiteit, Amsterdam. Fig 3.8 and Table 3.8 shows data from these experiments on one individual with the genotype AMY1*H0/H0. As can been seen from figure 3.8, the QAMY02 & QAMY03 systems produce the ratios of AMY1:AMY2 peak heights that are closest to the expected value for that individual, with the smallest variance in ratios across multiple electrophoresis runs. As a result the QAMY02 and QAMY03 systems were used to type the large number of samples from a number of global human populations for AMY1 repeat allele genotype.

Table 3.7 Results of experiments on the QAMY02 system to determine the number of electrophoresis runs required to ensure results fall within the acceptable range of +/- 0.2 either side of expected ratio. The QAMY02 PCR was performed three times on a DNA sample of known genotype obtained from Jan Pronk, Vrije Universiteit, Amsterdam. Electrophoresis was carried out four times on each of the three PCRs that were carried out.

Let V= variance across electrophoresis runs on an ABI377/GeneScan™ system

R = acceptable range either side of expected ratios for known AMY1 genotypes

C = Two tailed critical value so that 99% of data fell within the acceptable range

$$\text{Number of runs required} = \frac{V}{\frac{R}{C} \times \frac{R}{C}}$$

| PCR no. | Standard deviation of results across runs | Acceptable range either side of expected ratios for known AMY genotypes (+/-) | Two tailed critical value (for 99% data to fall within acceptable range) | Variance of results across runs | Number of runs required |
|---|---|---|---|---|---|
| PCR A | 0.109 | 0.2 | 2.57 | 0.011 | 1.968 |
| PCR B | 0.097 | 0.2 | 2.57 | 0.009 | 1.574 |
| PCR C | 0.098 | 0.2 | 2.57 | 0.009 | 1.617 |

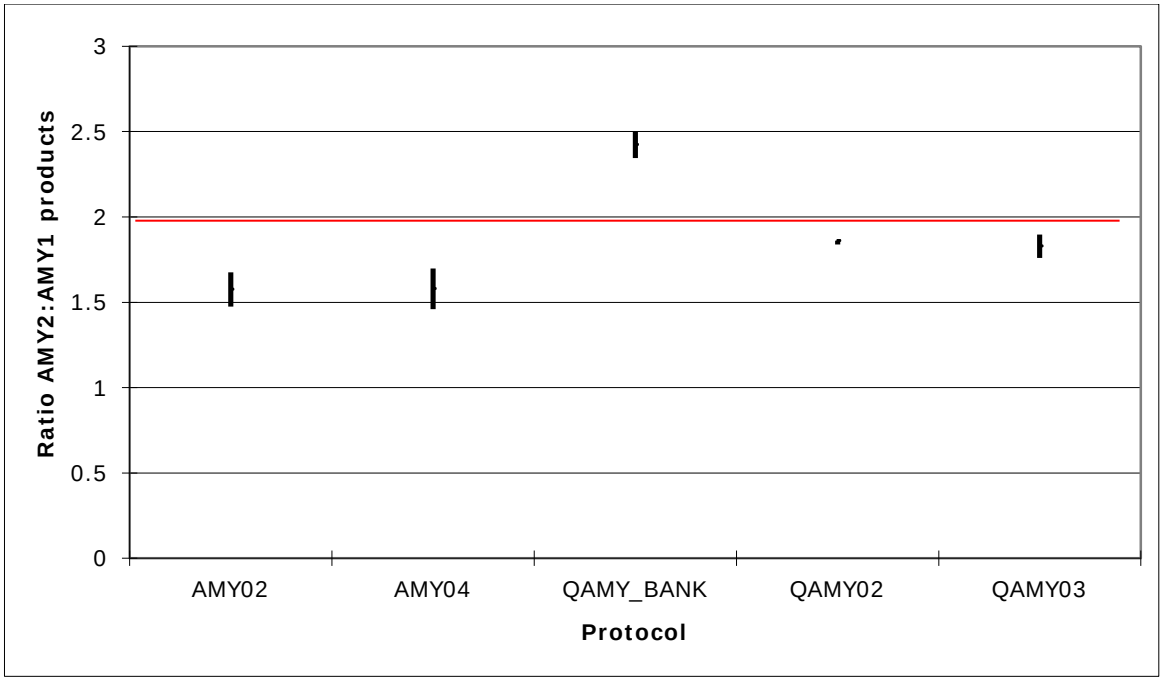
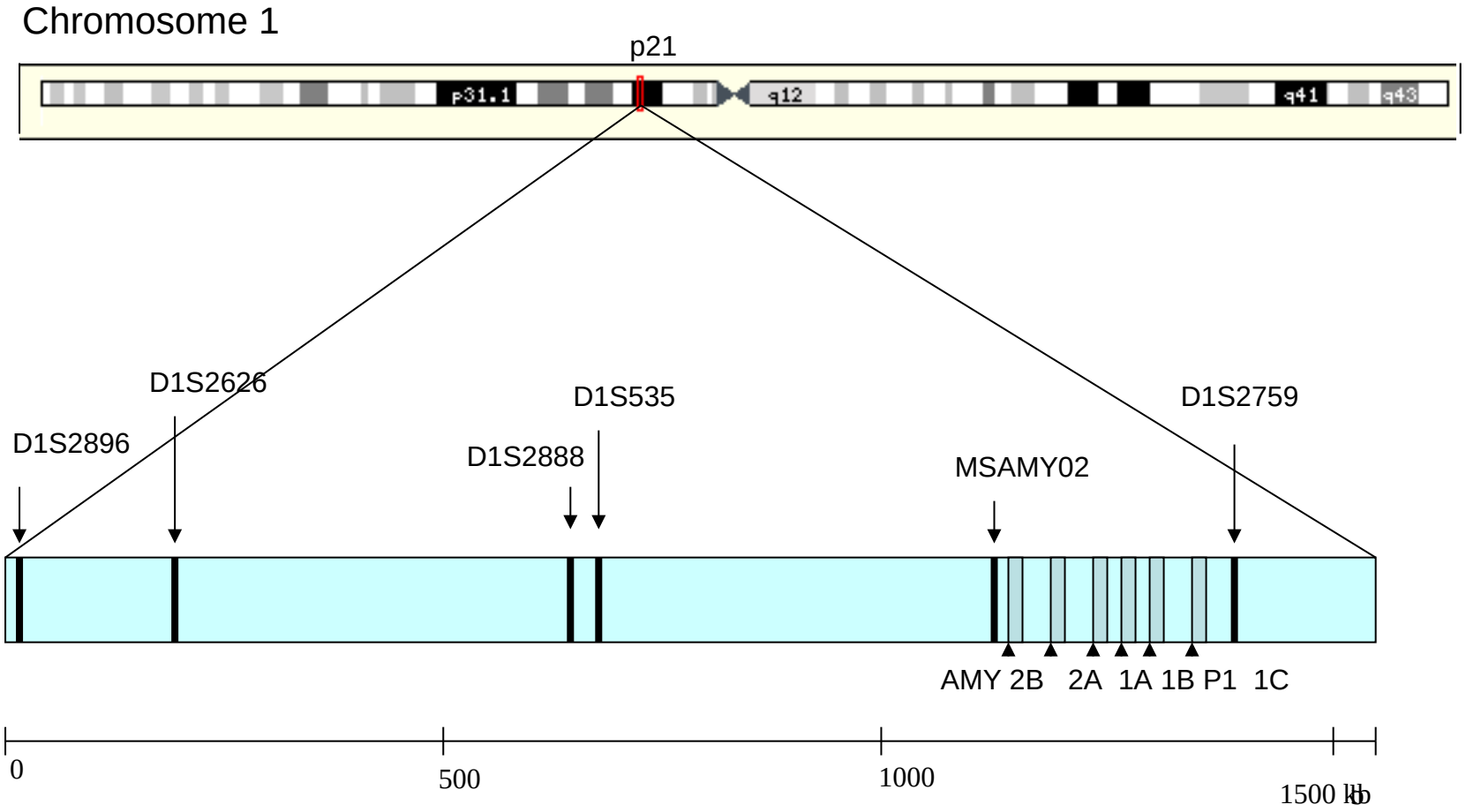



Fig 3.8: Mean ratio of AMY2:AMY1 PCR products fluorescence for one individual from 5 systems used in AMY1 quantification. The expected value for this individual (genotype AMY*H0/H0) is shown in red. AMY02 (n=32) and AMY04 (n=32) are protocols that use restriction enzymes to distinguish AMY1 and AMY2 specific PCR products. QAMY_BANK (n=15) is a PCR that amplifies the area around the same 22bp insertion in AMY1 described in Bank et al (1992). QAMY02 (n=32) & QAMY03 (n=32) are protocols that amplify regions around small insertions in either AMY1 or AMY2 genes.

**Fig 3.9: The Amylase Gene Cluster and location of 6 Closely Linked Microsatellite Markers**



*3.5.5 Detecting the structural arrangement of AMY1 genes on the chromosome.*
The PCR based methods described above can determine the total number of salivary amylase genes present in an individual but cannot, however, determine the apportionment of these genes on to the maternal and paternal chromosomes (see table 3.4). This adds an extra layer of complexity to the problem of assigning phase. It is important to be able to determine phase if data is to be produced that can be analysed using haplotype based tests for selection. Ideally a further assay would distinguish between certain haplotypes, improving the resolution of the existing protocols. For example, a diagnostic assay for AMY1*HO would distinguish between AMY1 HO/H2 and AMY1 H1/H1. However as explained above, the sequence information to design such an assay is to date unavailable.

To overcome the problem of determining the phase of the AMY1 genes a statistical approach, using an expectation maximisation (EM) algorithm was developed. A set of functions, called EMamy, incorporating the EM algorithm was written for the MATLAB programming environment by M. Weale (see Table 2.6). The EM algorithm is a general method for finding the maximum likelihood estimate of the parameters of a model when the dataset is incomplete or has missing values (Bilmes 1998). The algorithm consists of two steps: Firstly, the E step calculates expected values for the missing data from the starting parameters. The M then recalculates the parameters from the data using a maximum likelihood equation. These two steps are repeated until a maximum likelihood is reached. The maximum likelihood is expected to be returned for the estimates of the parameters that would give rise to the observed data.

The EMamy functions were designed to analyse data in the form of the total number of AMY1 repeat units present in an individual, from families consisting of one father, one mother and two children. In some families it is possible to deduce the haplotypes of the parents by following the inheritance of the AMY1 repeat alleles through to the children. Using the information from these families, the EMamy functions return EM estimates of the frequencies of the AMY1 repeat alleles and

therefore assess the relative probability of a total number of AMY1 genes being the result of different combinations of AMY1 repeat alleles (See Appendix A). The EMamy functions report all the possible parental haplotypes, given the children's genotypes, together with the relative probability for each genotype, using all the data, as well as the allele frequencies from the EM estimates.

### 3.6 Microsatellites

As discussed in Chapter 1 (Section 1.3.5.3), haplotype-based tests for selection currently provide the most powerful methods of detecting departures from the neutral expectation (Sabeti et al. 2002). Therefore, in addition to data collected on AMY1 gene copy number, protocols were designed to type six microsatellites closely linked to the amylase gene cluster.

Microsatellites are short tandemly repeated nucleotide motifs of between 2 and 5 bases long that are densely dispersed throughout the genomes of eukaryotes. Many microsatellites have been found to be highly polymorphic in terms of repeat copy number (Goldstein & Pollock 1997). Even in relatively small samples it is typical to find more than 10 repeat number alleles with the result that migrational processes and other demographic events can be studied (Bowcock et al. 1994)

The Human Genome Project databases was searched for known microsatellite markers less than 1 Mb from the amylase gene cluster. Dracopoli & Meisler (1990) identified a polymorphic dinucleotide microsatellite marker 2.3 kb from the AMY2B gene. Additional microsatellites were found through analysis of the contigs of the amylase gene cluster region (see Fig 3.2) using the program ETANDEM from the Human Genome Mapping Project (HGMP) EMBOSS package (telnet://tin.hgmp.mrc.ac.uk). This program searches for tandem repeats and scores them, giving the highest scores to the longest repeated regions as well as those in which the repeat motif continues uninterrupted by bases not contained in the motif. Repeats with the highest scores were analysed visually in Sequencher v.4 (Gene Codes, Ann Arbor Michigan). Using these methods, six microsatellites were selected and primers were designed using Oligo v4.0 to amplify regions around the chosen microsatellite markers (See Fig 3.9).

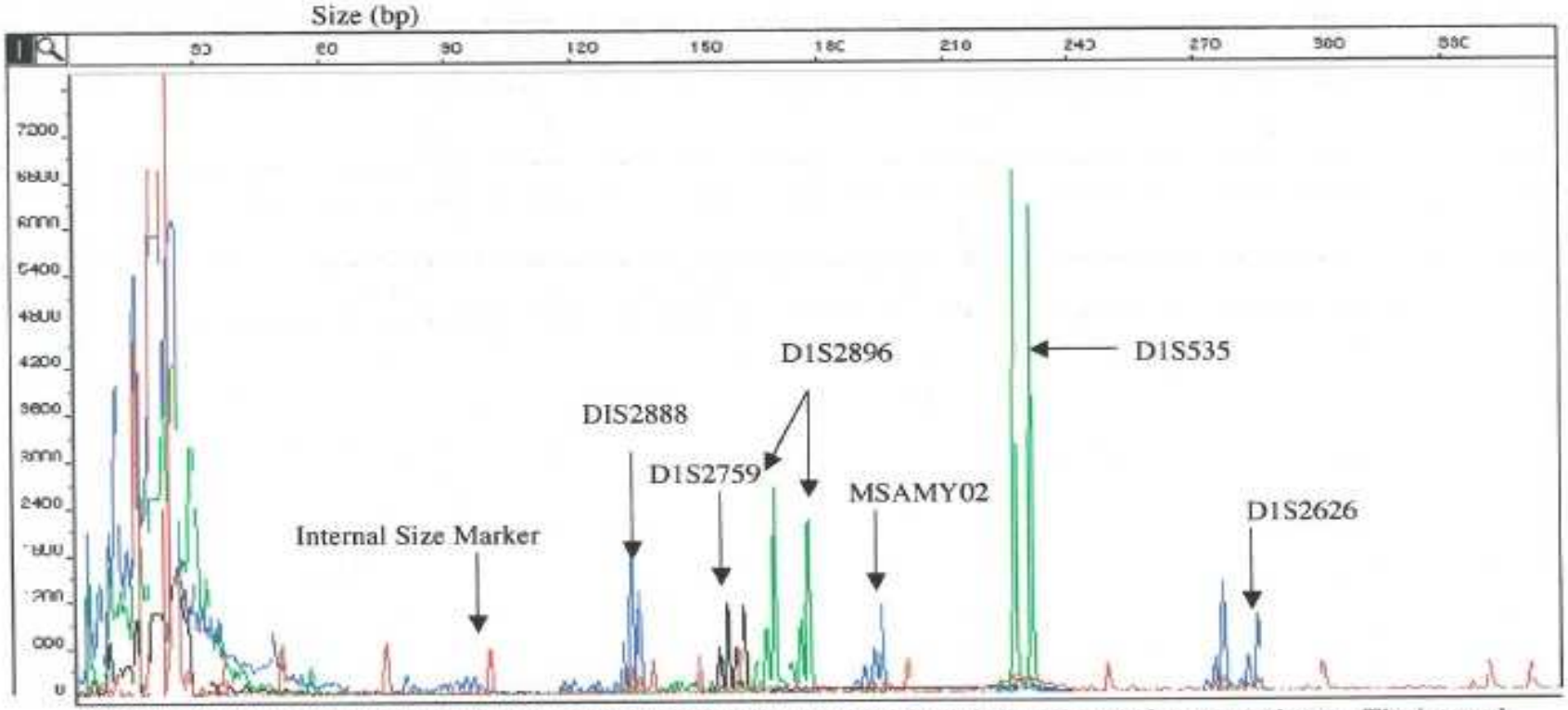



Fig 3.10 GeneScan output for multiplex PCR of 6 microsatellites closely linked to the amylase gene cluster. The internal size marker appears in red.

As with other protocols, primers were optimised for annealing temperature, primer concentration, $MgCl_2$ concentration and for use with DNA extracted from buccal swabs. Once primers were optimised in single PCRs, final concentrations (see Table 2.5**)** of the various primer pairs were optimised for use in a multiplex PCR. Primers were tested in multiplex, initially at equal concentrations (0.2 μM). Following this, the concentrations of individual primer pairs were tested at a range of concentrations from 0.06μM to 0.5μM to achieve the optimal amplification in terms of minimising the amount of primer used to produce as far as possible equal amounts of PCR product for all markers (see Fig 3.10). Final reactions condition for the PCR reaction and electrophoresis are described in sections 2.5.4 & 2.5.6.

Although the ABI377/GeneScan™ system produces very accurate relative estimates of the size of DNA fragments, the sizing not absolute (Thomas et al. 1999, Dr. L. Tagg, PE-Applied Biosytems, *pers. comm.*). The observed product sizes assigned to PCR products using GeneScan analysis software, although consistent across runs, can differ by up to 6 nucleotides from the actual product size. Consequently, a number of DNA samples were sequenced in the region of the 6 microsatellites to provide a way of calibrating the scoring of microsatellite repeat alleles. Sequencing was performed according to the protocol described in section 2.4. The number of repeat motifs contained in the samples was counted from the sequencing trace. This information was combined with the GeneScan™ size estimate and compared to the sequence length and number of repeat motifs found in the reference sequence for the microsatellites obtained from GenBank. Where discrepancy between the GeneScan™ size estimate combined with the number of repeats from sequencing, and the reference sequence was found, a correction was applied to the GeneScan™ size estimates (see Table 3.9).

| Protocol Name | N | Mean ratio of AMY:AMY PCR products | Standard Deviation | Variance | Standard error of the mean |
|---|---|---|---|---|---|
| AMY02 | 32 | 1.575 | 0.567 | 0.321 | 0.100 |
| AMY04 | 32 | 1.579 | 0.672 | 0.452 | 0.119 |
| Bank et al (1992) | 15 | 2.423 | 0.301 | 0.090 | 0.008 |
| QAMY02 | 32 | 1.860 | 0.120 | 0.015 | 0.021 |
| QAMY03 | 32 | 1.828 | 0.382 | 0.146 | 0.068 |

Table 3.8: A comparison of five protocols used in AMY1 quantification. AMY02 and AMY04 are protocols that use restriction enzymes to distinguish AMY1 and AMY2 PCR products prior to quantification. QAMY02 and QAMY03 protocols use small differences in length to distinguish AMY1 and AMY2 PCR products.

Table 3.9 Results of calibration of ABI377/GeneScan™ system for 6 microsatellite markers.

| Microsatellite Marker | Reference sequence length (bp) | Number of repeats in reference sequence | Size of PCR product in sample returned by ABI 377 (bp) | Number of repeats in sample run on ABI377 (from sequencing) | Adjustment (bp) |
|---|---|---|---|---|---|
| D1S2888 (*di*) | 139 | 20 | 133 | 17 | 0 |
| D1S2759 (*di*) | 157 | 19 | 157 | 17 | +4 |
| D1S2896 (*di*) | 163 | 13 | 168 | 14 | +3 |
| AMY-MS02 (*di*) | 191 | 18 | 184 | 17 | +5 |
| D1S535 (*tetra*) | 229 | 10 | 226 | 9 | +1 |
| D1S2626 (*di*) | 282 | 21 | 276 | 17 | +2 |

### 3.7 Summary and discussion

The protocols described in this chapter have been employed successfully to amplify DNA from samples taken as blood or buccal swabs and extracted with standard phenol/chloroform procedures. Some variation was observed in the intensity of signal peaks between different DNA samples. However absence of clear peaks for one or more loci was observed only in samples containing DNA that was either severely degraded or present at very low concentrations.

The combination of the large size (100kb each) of the polygenic repeat regions and the extremely high degree of similarity between the AMY1 genes presented a number of challenges in designing assays for detect the structural arrangements of the AMY1 genes on a particular chromosome. The first challenge was to design an assay to quantify the number of AMY1 genes in an individual by comparing the relative amounts of PCR products from AMY1 and AMY2 genes. It was extremely important that the relative amounts of AMY1 and AMY2 PCR products reflected the starting concentration in the DNA sample and so great care was taken to minimise the risk of unequal amplification efficiency between the two classes of genes. Once a reliable method had been developed to type individuals for AMY1 gene copy number, a further step was required to resolve certain genotypes, which contained equal numbers of AMY1 genes. As reliable sequence information for the region of the amylase gene cluster it was not possible to design further PCR based assays. Consequently a statistical approach, incorporating and EM algorithm, was used to provide AMY1 haplotypes for the individual chromosomes.

To date 1128 individuals from 14 human populations have been typed for salivary amylase gene copy number using the QMAY02 &QAMY03 protocols. 905 of these individuals from 7 populations have also been typed for the six microsatellites closely linked to the AMY gene cluster, using the AMY microsatellite multiplex PCR system outlined above. The methods described in this chapter provide a reliable and cost effective way to type individuals for these markers. The analysis of these data is described in the following chapters to investigate the worldwide distribution of variation

in salivary amylase gene copy number in humans as well as to assess the evidence for selection at the AMY1 locus.

# Chapter 4: Variation in AMY1 gene copy number in humans – does geography or dietary history best explain the patterns found?

## 4.1 Introduction

Bank et al. (1992) first reported that considerable variation in AMY1 gene copy number exists in humans. As outlined in Chapter One, the hypothesis explored in this thesis is that the variation in AMY1 gene copy number in humans may be the result of an adaptation to high starch diets, due to positive selection operating on AMY1 repeat alleles with high number so AMY1 genes. However, the original data set of Bank and colleagues is limited to a small number of Dutch families and therefore cannot inform us on differences in AMY1 gene copy number at the continental level or differences between populations with different dietary histories. The protocol developed in Chapter 3 provided a means of collecting data on variation in AMY1 gene copy number from a wide range of human populations from different geographical origins as well as contrasting dietary histories (see Table 4.1).

There are a number of questions that can be asked about the extent and nature of variation in AMY1 gene copy number. Do all the populations studied show variation in the number of salivary amylase genes in individuals? Do all populations have the same modal number of AMY1 genes? Do the populations show significant differences in AMY1 haplotype frequencies? Are the frequency differences in excess of, or less than would be expected under neutrality? And finally, does the distribution of AMY1 haplotype frequencies fit the predictions from the hypothesis that high AMY1 gene copy number haplotypes will be at higher frequency in populations with a long history of high starch diets than in those populations that have only adopted high starch diets in recent times?

Differences in allele frequencies between populations can be influenced by a number of factors including genetic drift shaped by demographic history as well as natural selection. As outlined in chapter 1, large data sets of both SNPs (Sachidanandam 2001) and microsatellites (see Kayser et al. 2003) are now available which form a

| Population | N | Approx date of farming (ybp) | Major prehistoric crops | Modern dependence on agriculture (after to Murdock 1968) | | Classification for AMOVA |
|---|---|---|---|---|---|---|
| Ethiopian families – Amharic speakers | 156 | 13,000 (Harlan 1989)<br>8500 (Ehret 2002) | Wild grain collection<br>Ensete cultivation | 60% | Intensive permanent - cereals | Agriculturalist |
| Nigerians - Ibibio | 94 | 8000<br>(Ehret 2002) | Yam cultivation | 60% | Extensive / shifting cultivation – roots or tubers | Agriculturalist |
| Algerian families | 36 | 7000<br>(Camps 1975) | Sheep, goats | 70% | Intensive cultivation dependent on irrigation, cereals | Agriculturalist |
| Malawi - Chewa | 96 | 2000<br>(Cavalli-Sforza et al 1994) | Millet & squash | 50% | Extensive / shifting cultivation - cereals | Non or recent agriculturalist |
| Armenian families | 100 | 10,000<br>(Hilman 1989) | Wheat, barley | 70% | Intensive permanent - cereals | Agriculturalist |
| Kuwait families | 32 | 10,000<br>(Harris 1981) | Wheat, barley | 60% | Intensive cultivation dependent on irrigation, cereals | Agriculturalist |
| Ashkenazi Jewish families | 116 | Unknown | Wheat, barley | 60% | Intensive permanent - cereals | Unknown |
| British families | 94 | 6000<br>(Thorpe 1996) | Wheat, barley | 60% | Intensive permanent - cereals | Agriculturalist |
| Irish families | 120 | 5000-5500<br>(Thomas 1996) | Wheat, barley | 50% | Intensive permanent – roots or tubers | Agriculturalist |
| German families | 120 | 5.800-7,000<br>(Diamond 1997) | Wheat, barley | 60% | Intensive permanent - cereals | Agriculturalist |
| Swedish Saami | 52 | N/a | N/a | 0% | Absence of agriculture | Non or recent agriculturalist |
| Mongolians | 96 | 6000-5000<br>(Morgan 1990) | Sheep. Horses, cattle | 10% | Casual, sporadic or slight cultivation | Agriculturalist |
| Yakut | 82 | Unknown<br>(Forsyth 1992) | Cattle, horses | 10% | Casual, sporadic or slight cultivation - cereals | Unknown |
| Singaporean Chinese families | 128 | 6000 - 3000<br>(Cavalli-Sforza et al 1994) | Rice, pigs | 70% | Intensive permanent cultivation & Horticulture | Agriculturalist |

Table 4.1: A summary of agricultural history of populations under study. Ybp = years before present. Power calculations were carried out that suggested that a minimum sample number of 50 was necessary for statistical analysis to have adequate power. Kuwait and Algerian samples have a sample size of less than 50 as they were added at a late stage in the project.

null distribution of allele frequency differences, against which to compare the data from the locus under investigation. As non-equilibrium population wide processes should affect all regions of the genome in a roughly equal fashion (Payseur et al. 2002), significant departures from the null distribution of frequency differences for sets of genome wide markers, forms the basis of method for testing hypotheses of local selection (Akey et al. 2002, Lewontin & Krakauer 1973, Cavalli-Sforza 1966). In this chapter data from AMY1 repeat allele frequency difference between populations will be compared to null-distributions generated from genome-wide SNP and microsatellite markers to see whether the AMY1 locus shows unusual allele frequency differences between populations, compared to the rest of the genome.

## 4.2 Methods

### *4.2.1 Sample collection and typing*

Samples were collected from 14 different populations from four geographical areas: Africa (Ethiopian families, Algerian families, Nigeria, Malawi), the Middle East/Western Asia (Armenian families, Kuwaiti families), Europe (German families, British families, Irish families, Ashkenazi Jewish families, Saami), East Asia (Singapore Chinese families, Yakut, Mongolia). The subject's ethnicity was self identified and noted along with other biographical information such as place of birth, current residence, first and second languages, cultural identity and religion. The same information was collected for the subject's mother, maternal grandmother, father and paternal grandfather. All samples were extracted and typed for AMY1 gene copy number according to the protocols described in section 2.3 & 2.4.2 respectively.

The following section gives a breakdown of samples typed for AMY1 gene copy number by country and ethno-linguistic group, as well as a brief overview of the population's origins and agricultural history (See also Table 4.1 for summary)

**Africa**

Ethiopia/Amharic speaking families (n=78)

In the Ethiopian highlands there exists some of the oldest evidence for the collection of wild grains and grasses for use as food, in the world (Harlan 1989). Deliberate cultivation of ensete dates from approx 8500-7500 ybp and Ethiopia eventually became either the primary or secondary point of dispersion for 36 crops including teff, a small kernel grass whose flour is often baked into large round flat breads and is remains a major crop in Ethiopia today (Marcus 1994).

Nigeria/Ibibio (n=47)

The Niger-Congo speaking peoples are thought to have developed the intensive collection of wild yams, which have a high starch content in the area covered by modern day Nigeria. West African planting culture developed from 8,000 years ago in response to a reduction in the availability of wild yams due to the spread of a wetter and warmer climate and woodlands. The West African planting tradition included the deliberate cultivation of yams, black-eyed peas and voandzeia (an African groundnut) (Curtain et al. 1995).

Algeria/Arab families (n=18)

The beginnings of agriculture in Algeria are poorly documented (Camps 1975). However, the remains of domesticated sheep and goats have been found in the Haua Fteah cave in eastern Libya dating from 7000 ybp (Rogerson 1998). The present day inhabitants of Algeria are thought to have resulted from migrations of Arabs and Bedouins from the middle East, which admixed with local Berber groups (Cavalli-Sforza et al. 1994).

Malawi/ Chichewa (n=48).

Around 2000 years ago Bantu speaking peoples began migrating into the area around Lake Malawi, bringing with them an entire economy combining techniques for iron working with a range of crops such as millet and squash (Needham et al. 1984). They soon displaced and over-ran the hunter-gatherers they encountered. Between the 14th and 19th centuries, many more Bantu tribes migrated to Malawi (Fage 1988).

**Middle East**

Armenia/Armenian families (n=50)

Modern Armenia is a fraction of the size of Ancient Armenia, which included modern day North East Turkey, Armenia, and parts of Iranian Azerbaijan (Reigate 2000). This large area includes the early eastern arm of the expansion of farming from the Fertile Crescent (Harris 1981). From 8000ybp, irrigation techniques and new tools such as hoes extended the areas brought under agriculture. Wheat and barley as well as sheep, goats and cattle formed the basis of the new economy.

Kuwait/Arabic (n=14)

Ancient Mesopotamia, which includes the area covered by modern day Kuwait, became the linchpin of ancient international trade. The fertile soil between the Tigris and the Euphrates, as well as the development of irrigation around 8000 ybp, produced a large surplus of food, such as wheat and barley, which was used to trade for minerals (such as copper from Magana in present day Oman) and timber from the Indus valley (Cavalli-Sforza et al. 1994). From 4000 years ago, regular incursions from the nomads of the interior caused the gulf coast to take on a distinctly Arab flavour.

**Europe**

Ashkenazi Jewish families (n=60)

In tenth century Christian Europe, Jewish communal and social life as well as Jewish scholarship developed in the three Rhineland communities of Speyer, Worms and Mayence. From there they spread westwards to France and eastwards to Eastern Germany and Bohemia, establishing a unity of custom, ritual and law. These communities became known as the Ashkenazim. The word is now generally applied to all Jews of European origin and customs (apart from small groups of Spanish and Portuguese Jews who follow the Eastern or Sephardi tradition) (Werblowsky & Wigoder 1997). Studies using both classical and molecular markers have shown evidence for both the common genetic origin

of Jewish communities and admixture between Jewish communities and their geographical neighbours (Thomas et al. 2002).

Germany/German (n=62)

As farming spread westwards and northwards from Anatolia and the Middle East, it followed first the Danube and then the Rhine valleys (Harris 1981). By 7000 ybp the Neolithic was well established in the area that is now modern Germany. In Roman times, Germanic tribes repulsed their invaders to the banks of the Rhine, and then went on to conquer territories in much of northern Europe. Their legacy can be seen in the modern speakers of the Germanic languages – the Dutch, Danish, English, Swiss, Flemish and Austrians (Cavalli-Sforza et al. 1994).

UK/English (n=44)

Before farming arrived, at around 6000 ybp, Britain was populated by hunter-gatherers who colonised the island following the retreat of the last ice age (Price 2000). The history of the British Isles has been marked by a series of invasions from mainland Europe. Weale et al. (2002) analysed Y chromosomes from males in 7 British towns and found evidence of a substantial migration of Anglo-Saxon Y chromosomes into central England, but not Wales. There is also evidence that the Danish Vikings made a significant contribution to the gene pool of the British Isles, especially in the North and East coastal regions (Capelli et al. 2003).

Eire/Irish (n=58)

Ireland is on the very western most fringe of Europe. The earliest evidence for agriculture in Ireland comes from sites such as Cashelkeety, C. Kerry where cereal like pollen have been found dating to 5,500 ybp (Woodman 2000). It is generally thought that the transition to an agrarian economy in Ireland was largely the result of acculturation of the indigenous Mesolithic communities, rather than colonisation by near eastern farmers. In a study by Hill et al. (2000) 98% males sampled had the putative ancestral Palaeolithic Y chromosome

haplogroup (hg1). This haplogroup is found at frequencies of 89% in the Basque but as low as 1.8% in Turkey.

Sweden/ Saami (n=27).

It is believed that the Saami arrived on the Fenno-Scandinavian peninsula just over 10,000 ybp. They are considered to be the first residents of this area since the last ice age and are thought to have followed their prey northwards as the glaciers retreated (Torroni et al. 2001). Approximately 60,000 Saami live in the northern regions of Norway, Sweden and Finland today. The traditional Saami diet consisted almost entirely of fish and meat with very little carbohydrate until the twentieth century (Haglin 1991, 1999)

**East Asia**

Mongolia/Mongolian (n=48).

The steppes of Central Asia were a difficult environment for agriculture but the open grasslands lent themselves to pastoral nomadism and animal husbandry. Goat, sheep and cattle remains are found from 6000 ybp. As is often the case with pastoral nomads, the Mongolian nomads traded with settled societies to the South for grain, tea and textiles (Morgan 1990). Murdock (1968) estimates that Mongolian nomads today cultivate approximately 10% of their food, with the remaining 90% being sourced through trade and from their livestock.

Russia/Yakut (n=41),

The Yakut live in central Siberia, among the Tungus people. Their language belongs to the Turkik family of languages along with modern Turkish. They are semi-nomadic pastoralists who keep cattle and horses as well as practicing agriculture. This subsistence pattern combined with the origin of their language suggests an origin from the steppes to the South rather than the Siberian forest. As Forsyth (1992) describes: “although it appears obvious that the Yakuts must have come from a steppe environment south of the Siberian forest, no convincing explanation exists of the reason for this or the time when it occurred”.

Singapore/ Singaporean Chinese (n=60)

The Chinese make up approximately 77% of the population of Singapore today. They are a relatively heterogeneous population in terms of dialect origin with more than 20 dialects represented. These dialects fall into three main groups – the Hokkiens, Teochews and Cantonese which all originate from Southern China. In the Yangtze area of Southern China, rice was cultivated in large quantities from 7000ybp. Rice and pigs also formed the basis of the Asian South coastal cultures from 6000ybp. The spread of rice cultivation to the islands of South East Asia took place from 3000ybp (Cavalli-Sforza et al. 1994).

### *4.2.2 Statistical analysis*

Estimates of allele frequency were obtained from the total number of AMY1 genes in individuals using the *EMamy* functions implemented in MATLAB (Mathworks, Natick, MA) as described in sections 2.8 & 3.3.5. Statistical analyses were performed on the AMY1 gene copy number data as detailed in sections 2.9.1 – 2.9.4.

## 4.3 Results

### *4.3.1 Similarities and differences between populations under study*

Considerable variation was found in AMY1 gene copy number between individuals, in all populations studied (See Fig 4.1). Some individuals had a total of only two AMY1 genes (one copy of AMY1C on each chromosome with none of the 100kb polygenic repeat unit (see Fig 1.9 and Table 4.2). In contrast, individuals were found with as many as 16 AMY1 genes. The distribution of the different AMY1 total gene counts in the 14 populations under study can be seen in Fig 4.1.

The population with the highest mean number of AMY1 gene copies per individual was the Mongolian sample which had an average of 7.9 AMY1 gene copies per individual (see Fig 4.2) The populations with the lowest average number of AMY1 gene copies per individual (5.7) was the Saami.

| | Population | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No. AMY1 genes | Algeria | Malawi | Nigeria | Ethiopia | British | Ireland | Germany | Saami | Ashkenazi Jews | Armenia | Kuwait | Singapore Chinese | Yakut | Mongolia | Total |
| 2 | 0 (0.000) | 0 (0.000) | 0 (0.000) | 1 (0.013) | 1 (0.023) | 1 (0.017) | 3 (0.048) | 3 (0.111) | 2 (0.033) | 2 (0.040) | 0 (0.000) | 0 (0.000) | 4 (0.098) | 0 (0.000) | 17 |
| 4 | 4 (0.222) | 16 (0.333) | 6 (0.128) | 28 (0.359) | 10 (0.227) | 12 (0.207) | 17 (0.274) | 8 (0.296) | 14 (0.233) | 10 (0.200) | 6 (0.429) | 8 (0.133) | 7 (0.171) | 4 (0.083) | 150 |
| 6 | 4 (0.222) | 22 (0.458) | 20 (0.426) | 27 (0.346) | 16 (0.364) | 24 (0.414) | 22 (0.355) | 11 (0.407) | 30 (0.500) | 16 (0.320) | 6 (0.429) | 24 (0.400) | 13 (0.317) | 17 (0.354) | 252 |
| 8 | 8 (0.444) | 6 (0.125) | 13 (0.277) | 14 (0.179) | 5 (0.114) | 11 (0.190) | 14 (0.226) | 3 (0.111) | 5 (0.083) | 15 (0.300) | 0 (0.000) | 19 (0.317) | 9 (0.220) | 14 (0.292) | 136 |
| 10 | 0 (0.000) | 3 (0.063) | 6 (0.128) | 7 (0.090) | 7 (0.159) | 4 (0.069) | 1 (0.016) | 1 (0.037) | 7 (0.117) | 3 (0.060) | 1 (0.071) | 5 (0.083) | 2 (0.049) | 7 (0.146) | 54 |
| 12 | 1 (0.056) | 1 (0.021) | 2 (0.043) | 1 (0.013) | 2 (0.045) | 2 (0.034) | 4 (0.065) | 0 (0.000) | 2 (0.033) | 3 (0.060) | 1 (0.071) | 4 (0.067) | 4 (0.098) | 5 (0.104) | 32 |
| 14 | 1 (0.056) | 0 (0.000) | 0 (0.000) | 0 (0.000) | 2 (0.045) | 4 (0.069) | 1 (0.016) | 0 (0.000) | 0 (0.000) | 0 (0.000) | 0 (0.000) | 0 (0.000) | 1 (0.024) | 1 (0.021) | 10 |
| 16 | 0 (0.000) | 0 (0.000) | 0 (0.000) | 0 (0.000) | 1 (0.023) | 0 (0.000) | 0 (0.000) | 1 (0.037) | 0 (0.000) | 1 (0.020) | 0 (0.000) | 0 (0.000) | 1 (0.024) | 0 (0.000) | 4 |
| Tot | 18 | 48 | 47 | 78 | 44 | 58 | 62 | 27 | 60 | 50 | 14 | 60 | 41 | 48 | 655 |
| Mean | 7.22 | 5.96 | 7.06 | 6.03 | 7.23 | 6.93 | 6.29 | 5.70 | 6.23 | 6.84 | 5.86 | 7.10 | 6.93 | 7.79 | |
| SeM | 0.63 | 0.28 | 0.30 | 0.24 | 0.48 | 0.37 | 0.32 | 0.54 | 0.29 | 0.38 | 0.64 | 0.27 | 0.51 | 0.35 | |
| Var | 7.12 | 3.66 | 4.15 | 4.31 | 10.27 | 7.89 | 6.41 | 7.91 | 4.89 | 7.20 | 5.82 | 4.40 | 10.83 | 5.83 | |
| SeV | 2.44 | 0.75 | 0.86 | 0.69 | 2.22 | 1.48 | 1.16 | 2.19 | 0.90 | 1.45 | 2.28 | 0.81 | 2.42 | 1.20 | |

Table 4.2: Total counts of the number of AMY1 genes per individual for the 14 populations under study. Figures in brackets are frequencies for the number of AMY1 genes per individual. SeM= standard error of the mean. Var = Variance, SeV = standard error of the variance.

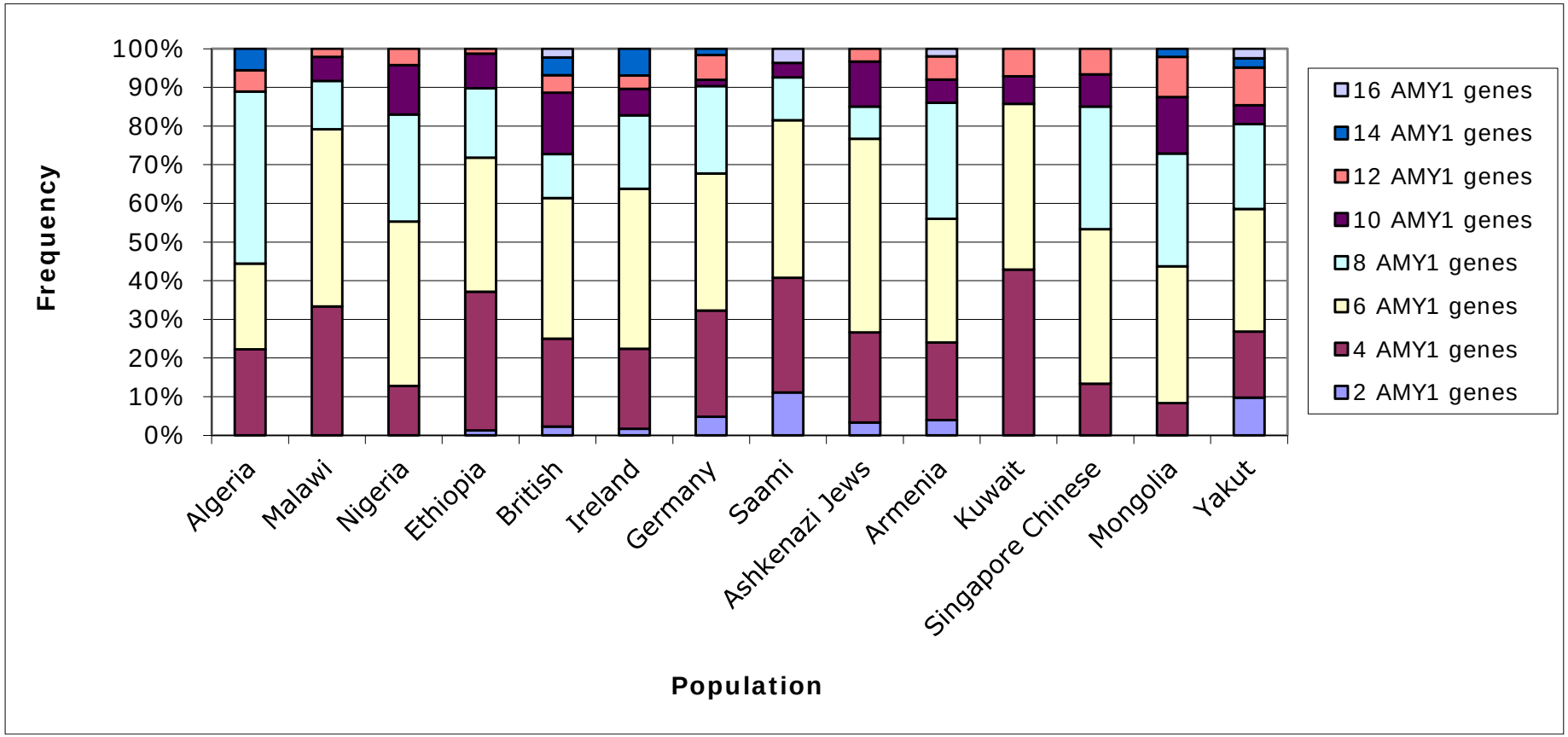



Fig 4.1 Bar Chart to show the distribution of AMY1 gene counts in individuals in the 14 populations under study

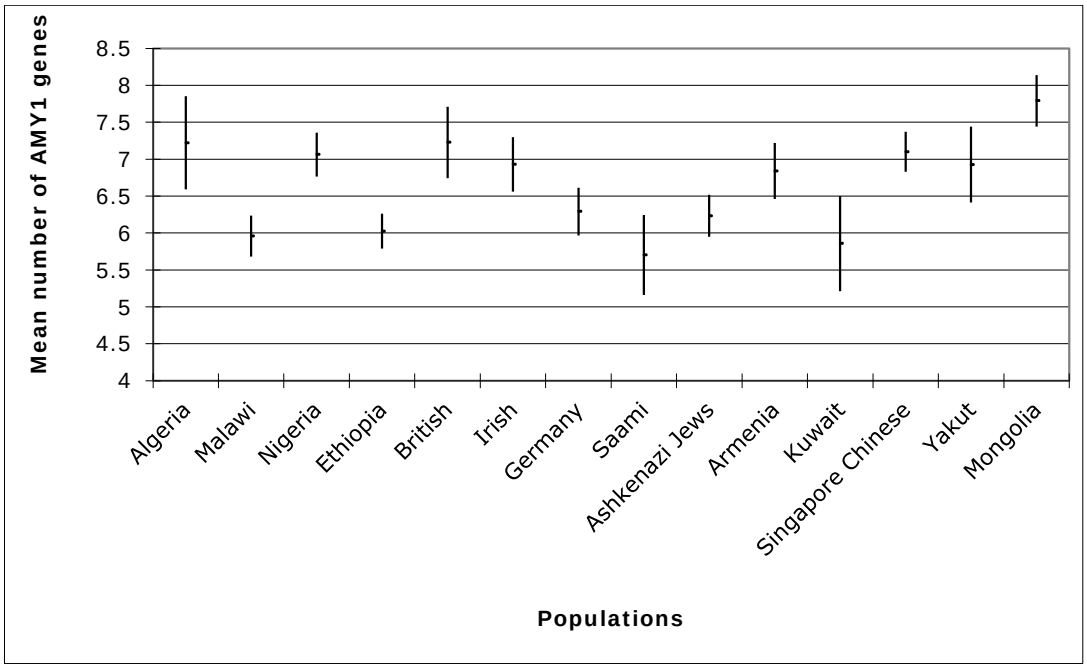



Fig 4.2: Mean AMY1 gene copies per individual in 14 populations under study. Mean value is plotted with +/- 1 standard error of the mean.

The modal number of AMY1 gene copies per individual in most of the populations studied, was 6 AMY1 genes per individual. There are three exceptions however. Algeria had a modal number of AMY1 genes of 8, and both Kuwait and Ethiopia had modal values of 4 AMY1 genes per individual. Six of the 14 populations (Algeria, Malawi, Nigeria, Kuwait, Singapore Chinese and Mongolians) had a minimum of 4 AMY1 genes per individual. All other populations had a minimum of 2 AMY1 genes per individual. The highest number of AMY1 genes found was 16 AMY1 genes per individual, a genotype found in the British, Saami, Armenians, Yakut and Mongolians. Algerians, Irish and Germans had a maximum of 14 AMY1 genes per individual and a maximum of 12 AMY1 genes per individual were found in Malawi, Nigeria, Ethiopia, Ashkenazi Jews, Kuwait and Singapore Chinese.

*Significant differences between populations under study*

As the total AMY1 gene count data appears non-normally distributed (see Fig 4.3 Line graph of counts) a Kruksal-Wallis (non parametric ANOVA) test (with a Dunn-Sidak correction for pairwise comparisons) was used ascertain whether there were significant differences in the average AMY1 gene count between populations. Pair-wise comparisons showed that the average AMY1 count in the Mongolian sample was significantly different from the Malawi sample ($p<0.005$), Ethiopian sample ($p<0.01$) and the Saami sample ($p<0.01$).

Fig 4.4 shows the variance in AMY1 gene count in the different populations. The variance is expected to reflect effective population size, which in turn can reflect the amount of drift in a population. This is because in small populations alleles are quickly lost from the population through drift, whereas in large populations there is a balance between the loss of alleles through drift and the arrival of new ones through mutation. The population with the highest variance in AMY1 gene count is the Yakut (10.8) where as the lowest variance is found in Malawi (3.66). This is an interesting result there are many loci where a greater genetic diversity in Africa has been found than anywhere else in the world, and that non-Africans

Fig 4.3a: Frequency of the different total AMY1 gene counts per individual in African populations. Algeria n=18; Malawi, n=48; Nigeria n=47; Ethiopia n=78.

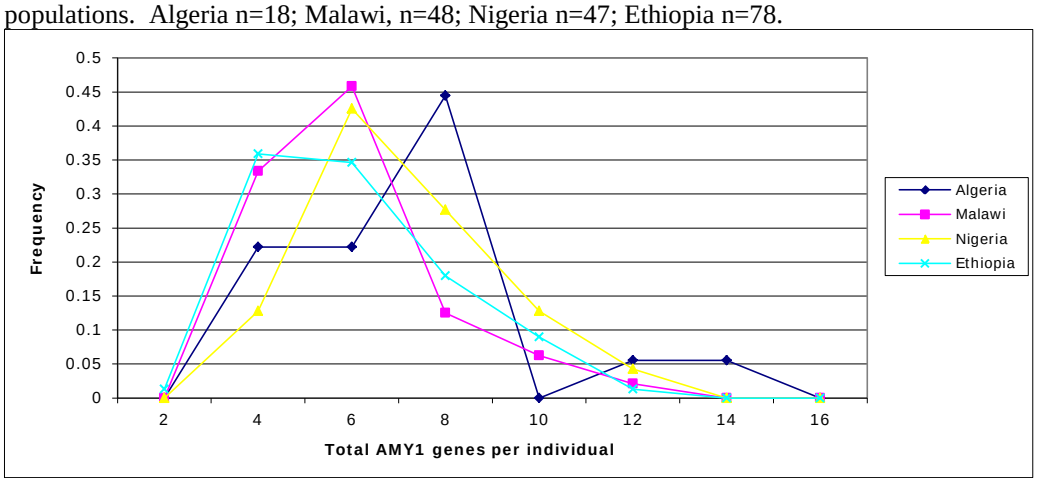



Fig 4.3b: Frequency of the different total AMY1 gene counts per individual in European populations. British n=44; Irish n=58; Germany n=62; Saami n=27; Ashkenazi Jews n=60.

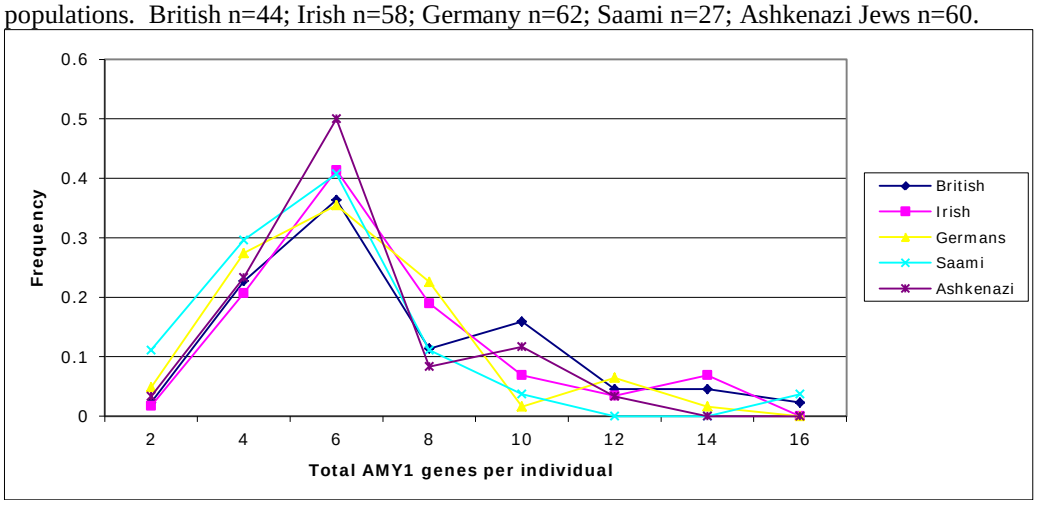
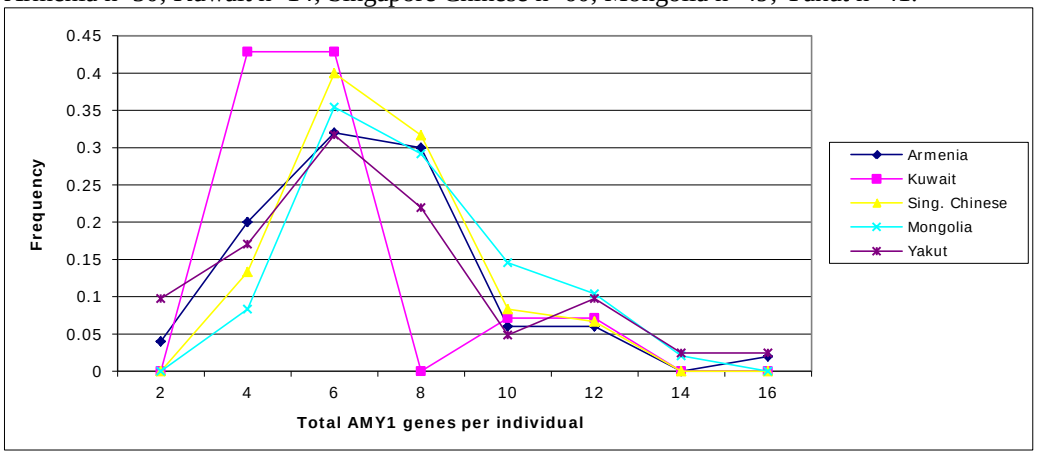



Fig 4.3c: Frequency of the different total AMY1 gene counts per individual in Asian populations. Armenia n=50; Kuwait n=14; Singapore Chinese n=60; Mongolia n=49; Yakut n=41.



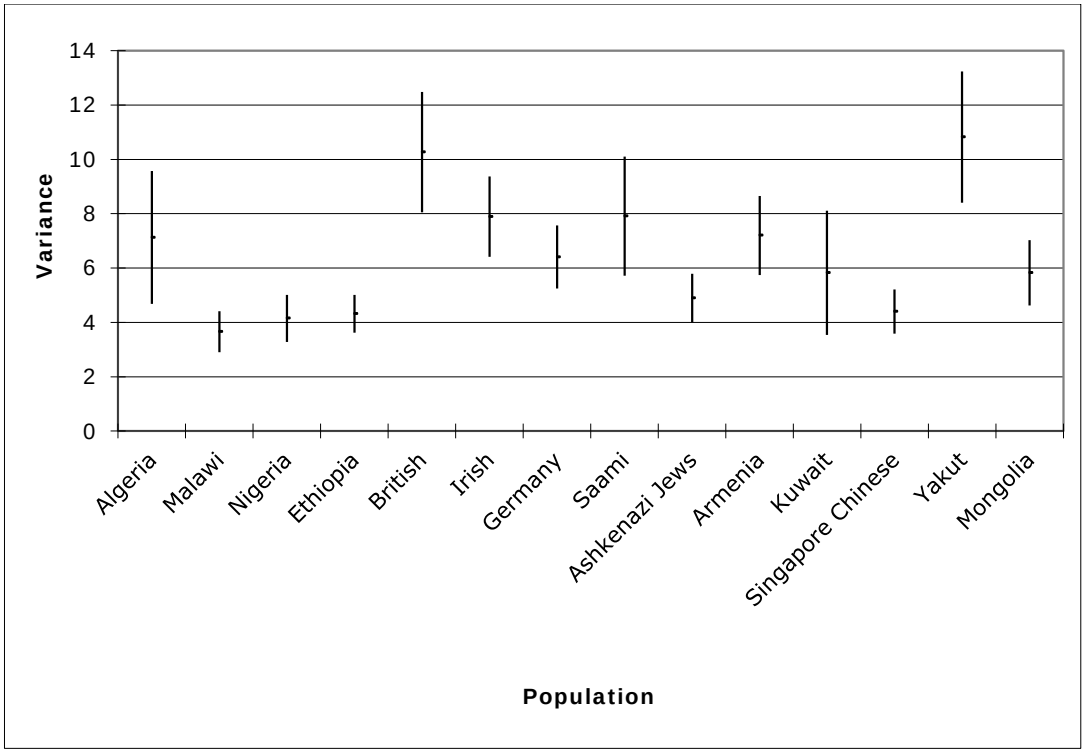



Fig 4.4: Variance in AMY1 gene copies per individual in the 14 populations under study. Variance is plotted with +/-1 standard error of the variance.

carry a small subset of African diversity (Bowcock et al. 1994, Tishkoff et al. 1996, Kaessman et al. 1999, Vigilant et al. 1991).

F statistics (i.e. variance in one population divided by variance in another) were calculated for pair-wise comparisons of significant differences in the variances between the different populations. After applying the Dunn-Sidak multiple comparison correction it was found that the Yakut has a significantly greater variance in AMY1 gene count from both Malawi (p=0.00043) and the Ethiopians (p=0.00054).

*AMY1 repeat allele frequencies in different populations*

As was discussed in Chapter 3, different combination haplotype of AMY1 repeat alleles that make up the various AMY1 genotypes, can give the same total number of AMY1 genes in an individual (See Table 3.4). In order to resolve the phase of the AMY1 repeat alleles, a set of functions incorporating an EM algorithm were developed by M.E. Weale (See 2.8 & 3.3.5). The *EMamy* functions return estimates of AMY1 repeat allele frequencies as shown in Table 4.3 along with expected heterozygosity values. The resulting *EMamy* estimates for the AMY1 repeat allele frequencies were used for performing an analysis of molecular variance (AMOVA) and genetic distance measures such as $F_{ST}$

*4.3.2 Is the variation between populations best structured with geography or agricultural history?*

To test whether geography or agricultural history best corresponds with the observed pattern of extant genetic diversity in AMY1 gene copy number, an AMOVA was performed on the AMY1 repeat allele frequency estimates from the 14 populations grouped into different classifications. First the populations were grouped into continents (Africa, Middle East, Europe, East Asia) and an AMOVA carried out. Subsequently the populations were reclassified as either hunter-gatherers and very recent agriculturalists (in the last 2000 years with an assumed previous history of hunter-gathering), or established agriculturalists (See Table 4.1).

| | Population | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AMY1 repeat allele | Algeria | Malawi | Nigeria | Ethiopia | British | Ireland | Germany | Saami | Ashkenazi Jews | Armenia | Kuwait | Singapore Chinese | Yakut | Mongolia |
| AMY1*H0 | 0.1004 | 0.1892 | 0.0764 | 0.2644 | 0.1949 | 0.2383 | 0.2312 | 0.2766 | 0.2271 | 0.1892 | 0.2176 | 0.1004 | 0.2361 | 0.0674 |
| AMY1*H1 | 0.5804 | 0.6864 | 0.6328 | 0.5102 | 0.5205 | 0.4274 | 0.5532 | 0.6117 | 0.543 | 0.5343 | 0.5516 | 0.5215 | 0.4721 | 0.5695 |
| AMY1*H2 | 0.2567 | 0.0789 | 0.2225 | 0.1801 | 0.1408 | 0.2666 | 0.1831 | 0.0603 | 0.1819 | 0.1417 | 0.1367 | 0.3288 | 0.2143 | 0.2423 |
| AMY1*H3 | 0 | 0.0368 | 0.0682 | 0.0454 | 0.1165 | 0 | 0.0153 | 0.0286 | 0.0406 | 0.1173 | 0.0583 | 0.0396 | 0 | 0.0809 |
| AMY1*H4 | 0.0312 | 0.0087 | 0 | 0 | 0.0273 | 0.0677 | 0.0082 | 0 | 0.0074 | 0.0175 | 0.0357 | 0.0096 | 0.0775 | 0.0809 |
| AMY1*H5 | 0.0312 | 0 | 0 | 0 | 0 | 0 | 0.009 | 0.0227 | 0 | 0 | 0 | 0 | 0 | 0 |
| n | 36 | 96 | 94 | 156 | 88 | 116 | 124 | 54 | 120 | 100 | 28 | 120 | 96 | 82 |
| heterozygosity | 0.585 | 0.485 | 0.540 | 0.635 | 0.657 | 0.685 | 0.607 | 0.544 | 0.619 | 0.645 | 0.625 | 0.608 | 0.604 | 0.669 |

Table 4.3 AMY1 repeat allele frequency estimates from *EMamy* functions and expected heterozygosity (h) values for the 14 populations under study. Expected heterozygosity (*h*) (equivalent to genetic diversity, see Nei 1987) was calculated using the formula:

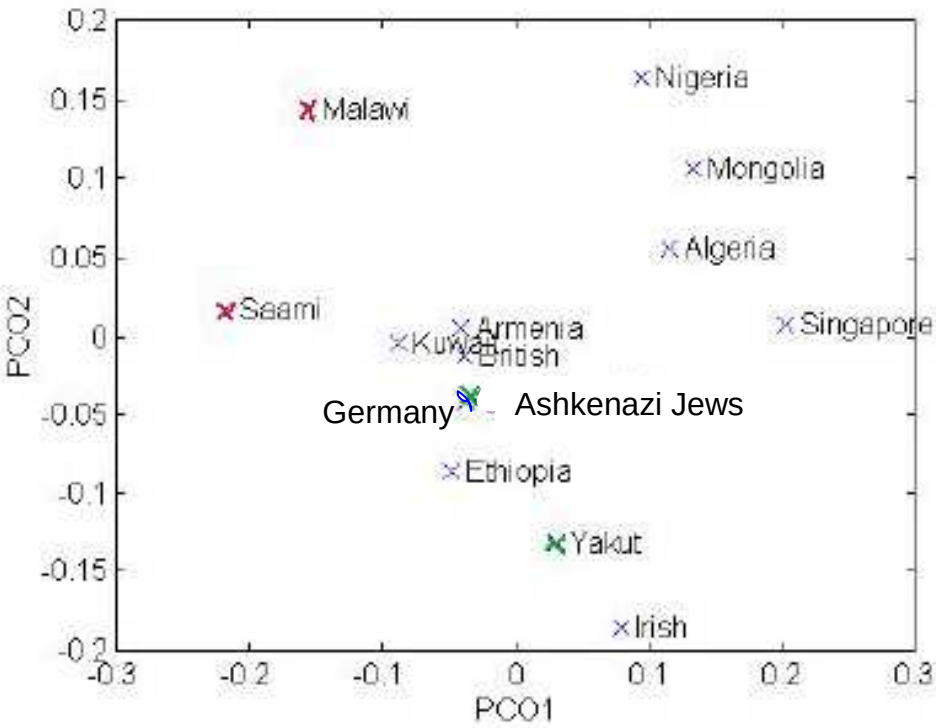
$$h = 1 - \sum_{i=1}^{m} x_i^2$$

The AMOVA method apportions the total variance in the data into three hierarchical levels: within populations, between populations within groups and between groups. The best classification of these populations is expected to maximise the amount of variance that is apportioned between groups (see Hurles et al. 2002). The results (Table 4.4) indicate that the best grouping is obtained when the populations are grouped by farming history rather than by continent.

Table 4.4 AMOVA on populations using two different groupings. The amount of variance apportioned to each of the three levels of classification is given for two different classifications of 12 global populations, based on farming history and geography. Populations are abbreviated as follows: Saa = Saami, Mal = Malawi, Alg = Algeria, Nig = Nigeria, Eth = Ethiopia, Kuw = Kuwait, Arm = Armenia, UK = British, Irl = Irish, Ger = German, Mon = Mongolian, Sing-Chi = Singapore Chinese.

| Grouping | Grouping rationale | Total % Variation | | |
|---|---|---|---|---|
| | | Within Populations | Between populations within groups | Between groups |
| {Saa, Mal}<br>{Alg, Nig, Eth, Kuw, Arm, UK, Irl, Ger, Mon, Sing-Chi} | Farming history | 96.4 | 1.37 | 2.2 |
| {Alg, Mal, Nig, Eth}<br>{UK, Irl, Ger, Saa}<br>{Kuw, Arm}<br>{Mon, Sing-Chi} | Geography | 97.9 | 1.50 | 0.59 |

Fig 4.6: A principal co-ordinate plot using pairwise $F_{ST}$ values for the 14 populations under study. The pair-wise $F_{ST}$ comparisons for the populations, estimated using Arlequin (Schneider et al 2000), were compiled as a matrix and subjected to a principal co-ordinate analysis using Genstat v3.2 (VSN, Hemel Hempstead, UK). Similar to Principal Component Analysis, this procedure explains the principal vectors of variance between population groups and extracted as many vectors as required to account for these differences. Populations with a long history of agriculture are shown with blue crosses, non (or recent) agriculturalists are shown with red crosses and unclassified populations are shown with green crosses.



From the brief survey of agricultural histories of the populations in this study in Section 4.2.1 (and summarised in Table 4.1) in some case the origins and history of agriculture are unclear. For example, the Ashkenazi Jewish sample were not included as they are an admixed group of Europeans with Middle Eastern origins (Thomas et al. 2002). In addition, the Yakut were not included in this analysis as their origin is uncertain (Forsyth 1992).

Fig 4.6 is a principal co-ordinate plot based on pairwise $F_{ST}$ values for AMY1 allele frequencies in the 14 populations under study (see Table 4.5). The points on the PCO plot do not form recognisable geographical clusters. However, it should be noted that the points do not form clusters according to agricultural history either.

*4.3.3 Does the mean number of AMY1 genes in different populations follow what we would expect from there farming history?*

If there has been adaptation to high starch diets in terms of salivary amylase gene copy number, we would expect to see an increase in AMY1 gene copy number in those populations which have a long history of high starch diets. Populations that still practices hunter-gathering as their principle means of subsistence or have only adopted agriculture recently would be expected to have low starch diets and therefore a lower number of AMY1 genes on average.

The groupings of populations into old and new agriculturalists as in the hierarchical AMOVA (see Table 4.4) was explored to see if there were significant differences in mean AMY1 gene copy number between populations from the old and new group. Significant differences using a Kruskal-Wallis non parametric ANOVA, were found in the average AMY1 gene copy number between Malawi and Mongolia ($p<0.05$) and the Saami and Mongolia ($p<0.01$). Significant differences were also found within the old agriculturalist group between Ethiopia and Mongolia ($p<0.01$).

Table 4.5: $F_{ST}$ and P values for pairwise comparisons between the 14 population under study calculated with the Arlequin program (Scheider et al 2000) using *EMamy* estimates of AMY1 allele frequency. $F_{ST}$ values are in the lower left of the table, P values are in the upper right. Significant comparisons (P<0.05) are shown by the shaded boxes.

| | ALG | MAL | NIG | ETH | UK | IRL | GER | SAA | ASH | ARM | KUW | SCH | YAK | MON |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALG | * | 0.084+-0.0026 | 0.655+-0.0048 | 0.177+-0.0038 | 0.193+-0.0038 | 0.148+-0.0032 | 0.367+-0.0054 | 0.071+-0.0024 | 0.398+-0.0045 | 0.199+-0.0042 | 0.429+-0.0055 | 0.568+-0.0046 | 0.327+-0.0048 | 0.705+-0.0048 |
| MAL | 0.024 | * | 0.034+-0.0017 | 0.013+-0.0009 | 0.036+-0.0019 | 0.001+-0.0001 | 0.051+-0.0022 | 0.441+-0.0058 | 0.060+-0.0024 | 0.051+-0.0021 | 0.454+-0.0051 | 0.001+-0.0001 | 0.004+-0.0006 | 0.006+-0.0007 |
| NIG | -0.010 | 0.023 | * | 0.005+-0.0007 | 0.034+-0.0018 | 0.002+-0.0004 | 0.037+-0.0020 | 0.014+-0.0012 | 0.042+-0.0022 | 0.046+-0.0018 | 0.188+-0.0034 | 0.101+-0.0034 | 0.010+-0.0010 | 0.647+-0.0051 |
| ETH | 0.012 | 0.028 | 0.034 | * | 0.314+-0.0052 | 0.094+-0.0031 | 0.751+-0.0046 | 0.155+-0.0035 | 0.625+-0.0046 | 0.279+-0.0048 | 0.746+-0.0043 | 0.005+-0.0007 | 0.396+-0.0048 | 0.005+-0.0007 |
| UK | 0.010 | 0.021 | 0.021 | 0.001 | * | 0.025+-0.0015 | 0.267+-0.0045 | 0.125+-0.0032 | 0.205+-0.0041 | 0.999+-0.0003 | 0.914+-0.0028 | 0.009+-0.0011 | 0.178+-0.0041 | 0.071+-0.0027 |
| IRL | 0.015 | 0.073 | 0.053 | 0.009 | 0.021 | * | 0.079+-0.0029 | 0.005+-0.0007 | 0.073+-0.0027 | 0.015+-0.0013 | 0.183+-0.0043 | 0.020+-0.0013 | 0.758+-0.0044 | 0.005+-0.0007 |
| GER | 0.000 | 0.017 | 0.021 | -0.005 | 0.003 | 0.012 | * | 0.233+-0.0044 | 0.997+-0.0005 | 0.272+-0.0046 | 0.831+-0.0037 | 0.017+-0.0011 | 0.428+-0.0051 | 0.031+-0.0015 |
| SAA | 0.035 | -0.003 | 0.047 | 0.010 | 0.014 | 0.053 | 0.006 | * | 0.222+-0.0044 | 0.131+-0.0032 | 0.640+-0.0049 | 0.001+-0.0004 | 0.053+-0.0020 | 0.005+-0.0007 |
| ASH | -0.001 | 0.017 | 0.020 | -0.004 | 0.005 | 0.013 | -0.008 | 0.006 | * | 0.199+-0.0036 | 0.753+-0.0041 | 0.020+-0.0013 | 0.395+-0.0044 | 0.029+-0.0016 |
| ARM | 0.009 | 0.017 | 0.018 | 0.002 | -0.010 | 0.025 | 0.003 | 0.013 | 0.005 | * | 0.907+-0.0027 | 0.009+-0.0009 | 0.122+-0.0030 | 0.074+-0.0024 |
| KUW | -0.003 | -0.004 | 0.011 | -0.012 | -0.017 | 0.013 | -0.015 | -0.013 | -0.014 | -0.017 | * | 0.078+-0.0027 | 0.491+-0.0046 | 0.206+-0.0039 |
| SCH | -0.008 | 0.068 | 0.011 | 0.031 | 0.030 | 0.022 | 0.024 | 0.077 | 0.023 | 0.030 | 0.027 | * | 0.047+-0.0021 | 0.227+-0.0037 |
| YAK | 0.003 | 0.046 | 0.036 | -0.001 | 0.007 | -0.007 | -0.001 | 0.029 | -0.001 | 0.010 | -0.005 | 0.019 | * | 0.030+-0.0019 |
| MON | -0.011 | 0.037 | -0.005 | 0.031 | 0.014 | 0.037 | 0.022 | 0.056 | 0.022 | 0.013 | 0.010 | 0.004 | 0.024 | * |

Fig 4.7 shows the average number of AMY1 genes per individual plotted against the time since the development of agriculture for the 14 populations under study. There is a slight correlation ($r^2 = 0.1947$) between the time since adoption of agriculture and the mean number of AMY1 genes in individuals.

*4.3.4 Selection or drift? The interregional differentiation approach to identifying selection*

Although significant differences have been found in the mean AMY1 gene copy number between an old agricultural populations (Mongolia) and the new agricultural populations Saami and Malawi, these differences could also be explained by genetic drift. The allele frequency difference, quantified by $F_{ST}$ for Mongolia vs Saami comparison is 0.056 ($p<0.01$) and Mongolia vs Malawi is 0.03713 ($p<0.01$).

To examine whether the observed differences in AMY1 repeat allele frequency between the Saami and non-European populations are within the range expected under neutrality for the genome as a whole, allele frequency difference were quantified using the genetic distance measure $F_{ST}$ (Weir 1996), and compared to $F_{ST}$ values for large numbers of presumed neutral loci elsewhere in the genome, typed in comparable populations (See Cavalli-Sforza 1966, Lewontin & Krakauer 1973). As non-equilibrium population wide processes should affect all regions of the genome in a roughly equal fashion (Payseur et al. 2002), significant departures from the range of $F_{ST}$ values found in the rest of the genome would be consistent with local selection operating at the candidate locus. Thus the large data sets of SNPs and microsatellites that are now available effectively form a null-distribution against which to compare the data from the locus in question (Akey et al. 2002, Kayser et al. 2003, Sachindanandam et al. 2001).

Currently, large data sets of SNPs, from which null distributions can be constructed, are only available for Europeans, East Asians and African Americans (Sachidanandam et al. 2001). Although the populations examined in this study are not identical to those for which large data sets of SNPs are

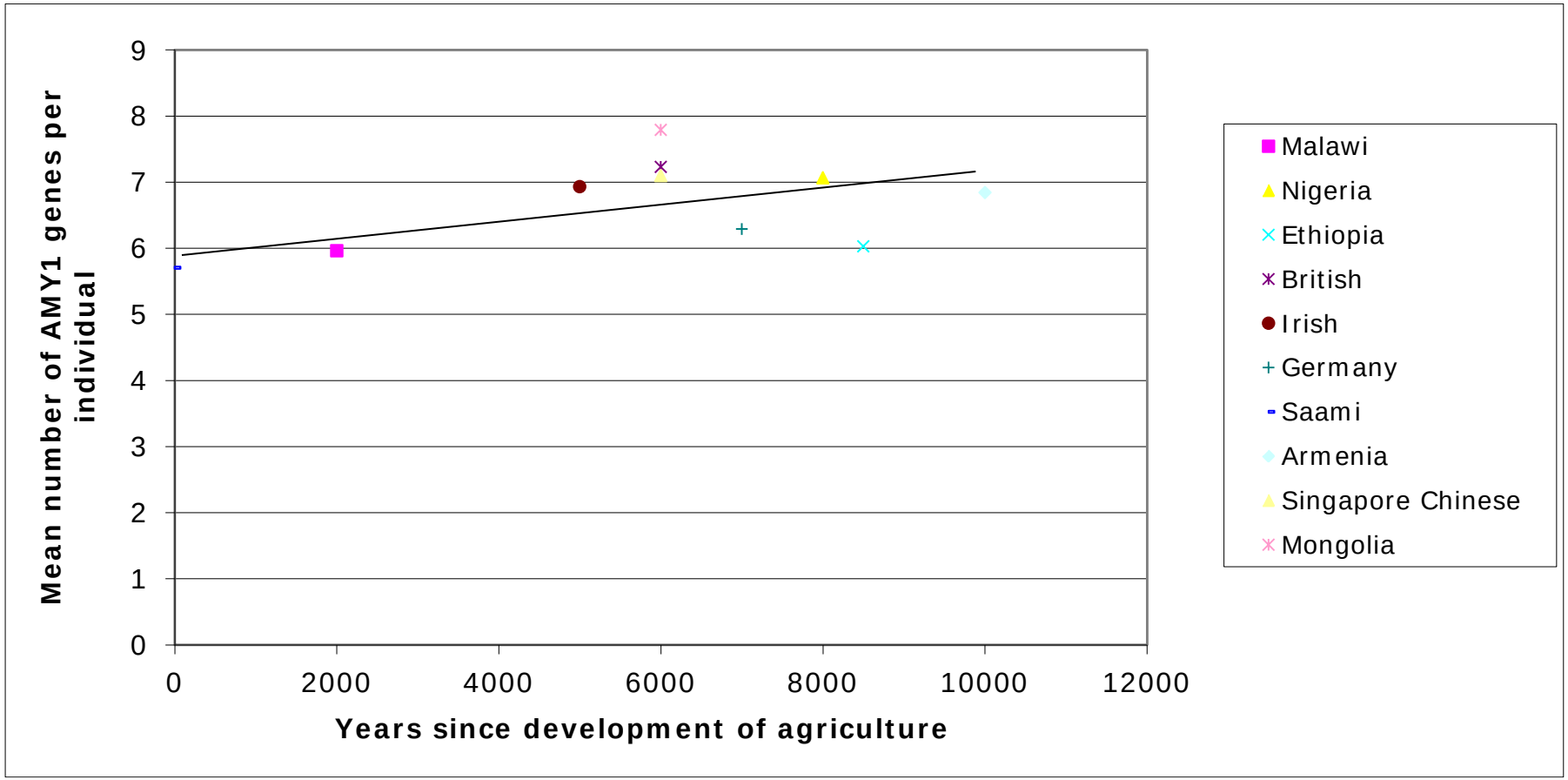



Fig: 4.7 Graph of mean number of AMY1 genes per individual and time since the development of agriculture. Algeria and Kuwait were removed from this analysis due to small sample size. The solid line shows the least order of squares line of best fit.

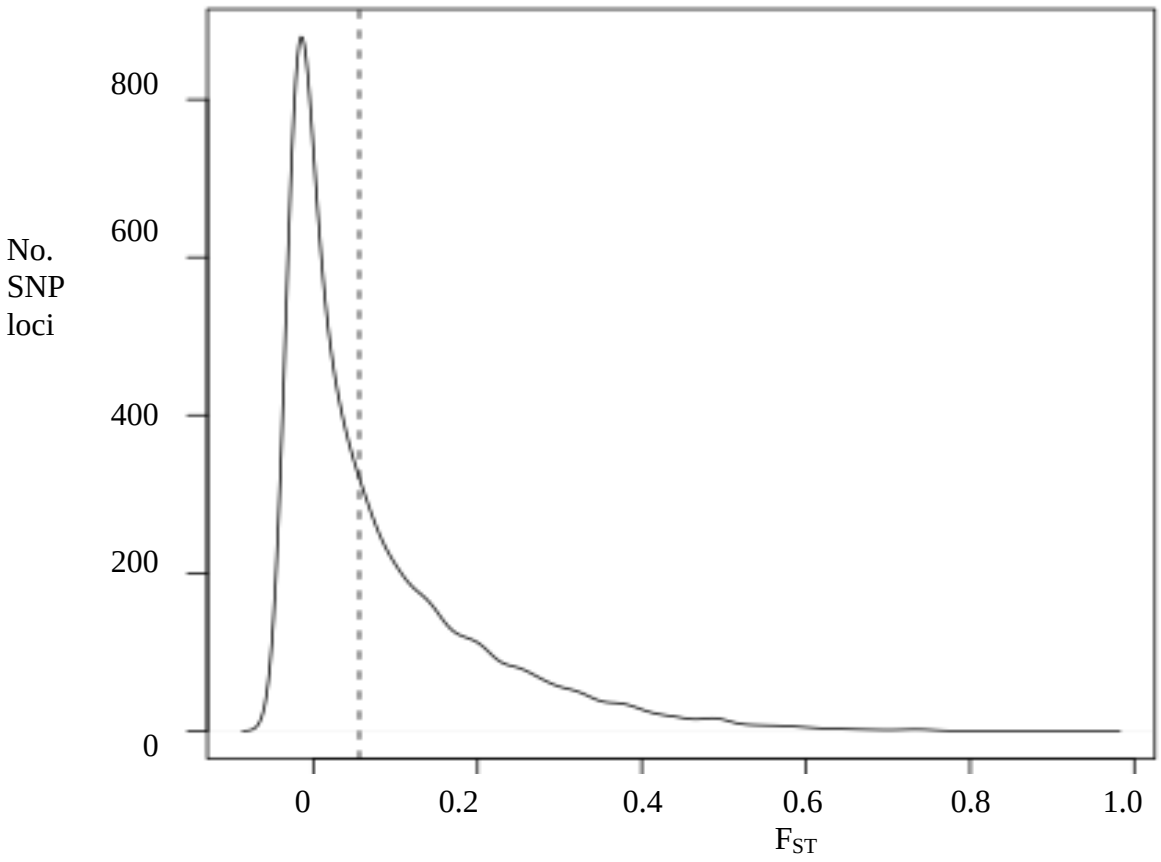
available, it can be argued that a comparison of the Saami / Mongolian AMY1 repeat allele $F_{ST}$ against the null-distribution of $F_{ST}$s for the European / East Asian data sets is a conservative one. This is because previous studies of a large number of classical polymorphic markers (Cavalli-Sforza et al. 1988; Cavalli-Sforza et al. 1994) have shown that the $F_{ST}$s between Saami and East Asian populations are typically lower than those between continental European and East Asian populations.

A total of 11,024 SNPs was carefully selected from a larger dataset of 33,487 SNPs typed in 42 East Asians and 42 European Americans by Sachidanandam et al. (2001). The selection criteria were such that each SNP was (a) either polymorphic or variable between populations, (b) mapped only once onto the genome, and (c) separated by at least 50 kb from the next nearest SNP, to minimise correlation in $F_{ST}$ values. It is important to point out that this SNP set will, by chance, contain some loci that are under selection, but unless the proportion of loci under balancing selection is large then this will have only a conservative effect on the comparison presented here. (see Caldwell et al. 2004)

When compared with the empirical genomic distribution of $F_{ST}$ values based on 11,024 SNPs, the $F_{ST}$ value for Mongolia vs Saami was found to lie in the top 41.6% of the distribution (Figure 4.8). The Mongolians and Saami represent the most extreme difference in AMY1 repeat allele frequency between populations. However as no significant departure from the neutral expectation was found, even using the most extreme comparison in terms of AMY1 repeat allele frequencies, it was not necessary to repeat this test for the other populations.

One major problem of comparing the AMY1 $F_{ST}$ values with data from SNPs is that the AMY1 gene copy polymorphism is not simply a single nucleotide polymorphism. Instead the repeat alleles are made up of a complex, polygenic repeat unit. The creation of novel AMY1 repeat alleles is thought to occur through unequal crossover events (Groot et al. 1990). In addition, many SNPs exist as biallelic markers, where as there are at least 5 different AMY1 repeat

Fig 4.8: Data on 11,024 SNPs typed in 42 East Asians and 42 European Americans (Sachidanandam et al 2001) was taken from a dataset of 33,487 SNPs typed by the Orchid Laboratory, publicly available at the SNP Consortium web site (http://snp.cshl.org/allele_frequency_project/panels.shtml). All $F_{ST}$ values were calculated using the unbiased 'random populations' formula for haploid data given by Weir (1996). The $F_{ST}$ value for AMY1 between Mongolians and Saami (0.056) is shown by the dotted line. 41.6% of SNPs in this distribution have a higher $F_{ST}$ than the AMY1 comparison.



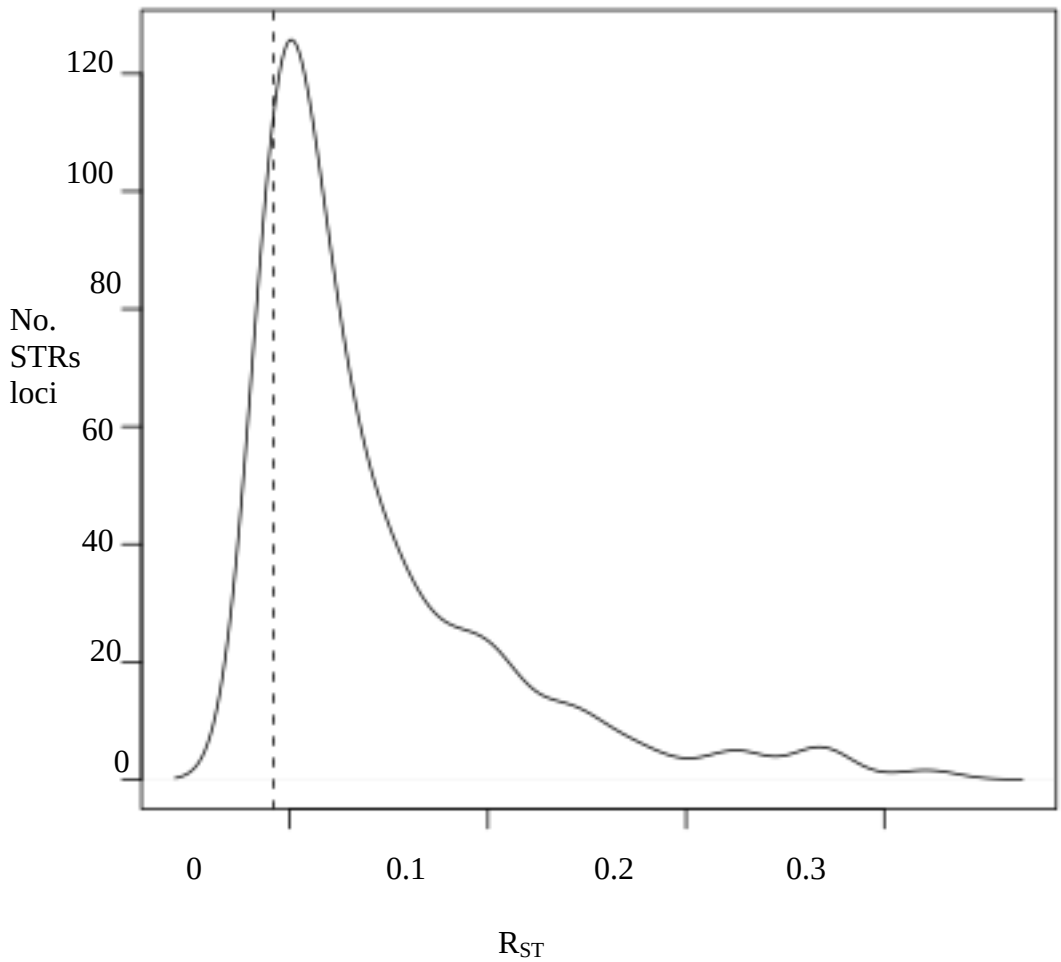
alleles. Unfortunately, data on systems elsewhere in the genome with a similar mutational process is not currently available. However data does exist for numerous multiallelic microsatellite loci spread throughout the genome. Microsatellites are thought to mutate in a step-wise fashion, as a result of DNA replication slippage (see Stumpf & Goldstein 2001 for a review), with most mutations involving an increase or decrease of a single repeat unit (Goldstein et al. 1995). It is a reasonable assumption that AMY1 repeat alleles also mutate in a step-wise fashion, and therefore worth comparing the allele frequency differences in AMY1 gene copy number between populations, with allele frequency difference calculated from microsatellites spread throughout the genome. The conventional measure of allele frequency differences between populations is $F_{ST}$, as used above. However, Slatkin (1995) developed a statistic known as $R_{ST}$, which is analogous to $F_{ST}$, but also incorporates information about the molecular distances between alleles from the step-wise mutation model.

In a comparison of 332 microsatellites between Europeans and Africans from across the genome, the $R_{ST}$ for AMY1 repeat allele frequency difference between the Ethiopians and German populations ($R_{ST}$= -0.008). A total of 84.6% of the 332 microsatellites had a higher $R_{ST}$ value (See Fig 4.9). It can be seen from Fig 4.9 that the $R_{ST}$ value for the AMY1 locus is not an outlier compared to the other $R_{ST}$ values for the 332 microsatellites.

4.3.5 *Estimating the mutation rate of AMY1 repeat alleles*

Slatkin (1995) showed that under an unbounded step-wise mutation model, the expected squared difference in repeat size (D) between two chromosomes separated by *t* generations is $t\mu\sigma^2$, where $\mu$ is the per generation stepwise mutation rate and $\sigma^2$ is the variance of the change in repeat size (assuming symmetric mutation). Within a single population, the expected coalescence time (*t*) is twice the effective population size ($N_e$). In a populations of $N_e$ diploid individuals, then the average square difference (D) between two chromosomes picked at random is $4N_e\mu$ (Goldstein et al. 1995). The average squared

Fig 4.9: $R_{ST}$ data for 332 STRs typed in 48 Europeans (blood donors from Leipzig, Germany) and 23 East Africans (from Gondar, Ethiopia) as well as 24 Southern Africans (from the Nguni, Sotho-Tswanga and Tsonga groups of South Africa) was taken from a dataset of 332 STRs typed by Kayser et al (2003). The $R_{ST}$ value for the AMY1 locus between Ethiopians and Germans (-0.008) is shown by the dotted line.



difference (D) is equal to 2V, where V is the variance in repeat alleles among chromosomes. Thus $V = 2N_e\mu$.

There is however a special case concerning the AMY1 repeat allele data. As discussed in chapter 3, the structural arrangement of the AMY1 repeat alleles and the genes they contain meant that an EM algorithm was applied to determine phase of the AMY1 repeat alleles. As this adds an extra layer of complexity to obtaining data on AMY1 repeat alleles it is desirable to incorporate into calculations the total count of AMY1 genes in an individual, which can be determined experimentally. Fortunately in this case this can be done as follows:

The average variance (V*) in 2 chromosome combined counts of AMY1 genes, $V^*=2V$.
Thus $V^* = 4N_e\mu$.

Average variance for AMY1 repeat alleles across all populations, $V^* = 6.4288$. Current human effective populations size ($N_e$) is approximately 10,000 (Wall 2003). Thus $\mu = 0.00016072$. This result is considerably smaller than the mutation rate that has been found for both autosomal and Y chromosome microsatellites of between 0.002 and 0.004 per locus per generation (Weber & Wong 1993, Kayser et al. 2003).

### 4.4 Summary and discussion

This is the first study to have investigated salivary amylase gene copy number variation in large numbers of individuals from a wide range of human populations. All populations studied showed extensive quantitative variation in AMY1 gene copy number among individuals. Significant differences in AMY1 gene copy number were found between the Mongolian sample compared to Malawi ($p<0.005$), Ethiopia ($p<0.01$) and the Saami ($p<0.01$).

The method used in this chapter for comparing allele frequency difference between populations for the AMY1 locus with a genome-wide distribution of allele frequency differences was designed to identify loci that have unusual frequency differences compared to the rest of the genome. One explanation for such differences would be local selective pressures in one population, but not the other, which would drive allele frequency differences. The hypothesis in this study was that local selective pressures in populations that adopted cultivation of high starch content crops would affect allele frequencies in those populations, so that there would be a large difference when compared to populations that had not adopted agriculture. However, the extent of inter-population differentiation at the AMY1 locus was within the range for presumed neutral loci for both SNPs and microsatellites between similar populations. Therefore drift cannot be rejected as an explanation for the differences in allele frequencies observed between populations.

If high AMY1 gene copy number is an adaptation to agriculture, which developed at the most 10,000 years ago, then perhaps not enough time has elapsed for a signal of selection to be detected. In fact, the strongest signals of selection in the human genome found to date have been detected for the lactase persistence associated allele at the lactase gene (Bersaglieri et al. 2004). Lactase persistence is thought to be an adaptation to milk drinking as a result of pastoralism, which arose less than 10,000 years ago. Salivary amylase gene copy number polymorphisms present an interesting and unusual case, however, as there is no null allele that stops the expression of salivary amylase enzyme,

causing a detrimental effect on the individual. The higher number of salivary amylase genes only increase the expression of salivary amylase enzyme (Bank et al. 1992). Although high numbers of genes would be an advantage to individuals consuming high starch diets, having a low number of genes may not present a great disadvantage. It is likely that the selective pressures acting on AMY1 are weaker than those acting on lactase, and other loci with alleles conferring resistance to fatal diseases. The inter-regional differentiation approach does not distinguish between weak selection that would, over short time periods, cause intermediate allele frequency differences within the range found in the rest of the genome, and drift which can also cause allele frequency differences. The method only provides a way of distinguishing loci with extreme allele frequency difference between populations (presumed to be the result of very strong differential selection) when compared with the rest of the genome.

In addition, it must be asked how appropriate a comparison of data from the AMY1 locus is with either SNPs or microsatellites, due to their very different mutation processes and rates. It is clear that more sensitive haplotype based methods using additional data from closely linked loci are required to rule out the possibility of selection having operated at the AMY1 locus.

Despite the lack of evidence found for selection through the inter regional differentiation approach, some evidence was found for larger average gene copy number occurring in populations with a long history of agriculture. In addition a hierarchical analysis of molecular variance (AMOVA) showed that the best grouping to maximise the amount variance between groups was when the populations were groups according to farming history rather than by geography. As it is very difficult to ascertain the average amount of starch eaten in the past by populations, farming of high starch crops was used as a proxy for a high starch diet. This rests on the assumption that on average, populations that cultivate high starch crops consume larger quantities of starch than they did before they started to cultivate, as well as those populations who do not cultivate them. Archaeological evidence shows that once cereal agriculture has been adopted by a population, then the starchy staples tend to dominate the diet (Cassidy 1980).

There are, however, less clear cut examples where the amount of high starch crops cultivated by a population is not an indicator of the amount of starch ingested by that population. These examples include pastoral nomads, who occasionally cultivate and often trade with neighbouring farmers, and hunter-gatherers who gather large amounts wild grains, roots and tubers. In these cases it is very difficult to determine the amount of starch eaten, which presents problems for classification of these populations for analysis.

There are many populations that were not available for this study that would provide additional information for a global survey of variation in AMY1 gene copy number in humans. In particular it would be desirable to increase the number of hunter-gatherer and non-agricultural populations, such as Australian Aboriginies, !Kung San, and the Arctic Inuit populations in order to facilitate the agricultural vs non-agricultural population comparisons. It would also be an interesting avenue to extend the study to include populations from the Americas where a number of high starch crops, such as maize, were domesticated.

The data presented in this chapter have provided an initial survey of variation in salivary amylase gene copy number in human populations. Extensive quantitative variation in gene copy number has been found in all populations studied to date. The question still remains: How did this distribution of variation arise? The following chapters will look in more detail for evidence of selective forces acting on the AMY1 locus using data from closely linked microsatellites, as well as from a sample of chimpanzees.

# Chapter 5: Microsatellites as tools for exploring variation and evolution in the human amylase gene cluster.

## 5.1 Introduction

Microsatellites have proved a valuable tool for inferring features of evolution and demographic processes as well as mapping complex diseases, linkage analysis and forensics in many species including humans. As outlined in chapter 3 (section 3.7) microsatellites are highly polymorphic in terms of repeat motif copy number as a result of having many potential alleles, and have a high mutation rate. Consequently, microsatellites have proved to be informative in dating mutational events (see Stephens et al. 1998, Goldstein et al. 1999), studying demographic events and processes (see DiRienzo et al. 1998, Bowcock et al. 1994), and detecting natural selection (see Slatkin & Bertorelle 2001). The range of applications of microsatellite data to the study of human evolution are illustrated in the following recent studies:

*Population structure and demographic history*

Rosenberg et al. (2002) studied 1056 individuals from 52 human populations using 377 autosomal microsatellite loci. Without using prior information about the origins of individuals, they identified up to six major genetic clusters, five of which correspond to major geographic regions, as well as subclusters that often correspond to individual populations. Zhivertovsky et al. (2003) used the same data set to infer splits and expansions in modern human populations. They estimated a populations tree based on the $T_D$ estimator of divergence time (see Zhivotovsky 2001) and a stepwise mutation model, that suggests that the branches leading to the present sub-Saharan African populations of hunter-gatherers were the first to diverge from a common ancestral population approx 71-142 thousand years ago. The branches corresponding to sub-Saharan farming populations and those that left Africa diverge next, with subsequent splits of branches for Eurasia, Oceana, East Asia and America. In addition, they were able to use the data to infer that African hunter-gatherer populations and the

populations of Oceana and America exhibit no statistically significant signature of growth.

*Using microsatellite data to infer the age of alleles*

Stephens et al. (1998) estimated the date of the CCR5-Δ32 deletion that inactivates the chemokine receptor on lymphoid cells. This receptor serves an entry point for a number of pathogens, including the human immunodeficiency virus (HIV-1). They performed haplotype analysis of 192 Caucasian individuals by typing for the CCR5 deletion and two closely linked microsatellite loci. They identified the most likely ancestral CCR5-Δ32 haplotype and then estimated the proportion of CCR5-Δ32 haplotypes that exhibit no change from the ancestral haplotype. Assuming that mutation and recombination occur at a combined rate *r*, they then used the proportion of unchanged haplotypes to estimate the date of origin. Stephens and colleagues (1998) estimated the age of the CCR5-Δ32 containing haplotype to be approximately 700 years old.

*Detecting regions of the genome that have experienced selection*

Both natural selection and demographic processes can lead to a skew in the frequency distribution of polymorphisms (Payseur et al. 2002). However, demographic processes such as population bottle-necks are expected to affect all loci in the genome in a roughly equal fashion. These population level processes cause a skew in the average genomic frequency distribution of polymorphisms, whereas selection causes localised skews in the frequency distribution for particular genomic regions. Payser et al. (2002) analysed publish data from 5,257 mapped microsatellites in individuals of European descent, with a sliding window analysis of the frequency distribution of microsatellite polymorphisms across the human genome. They identified 43 regions that had unusually skewed frequency distributions, which may have been subject to positive selection. Kayser et al. (2003) used data from 332 microsatellite loci in both Europeans and Africans to identify genomic regions with significantly larger than average genetic distances (measured using $R_{ST}$) between populations. They identified 11 regions of the genome that exhibited larger genetic difference between populations than average, consistent with selection.

*Analysis of Intra allelic variability*

As outlined in Chapter 1, intra-allelic variability is the joint distribution of the frequency of a neutral allele and the extent of variability at closely linked marker loci (Slatkin & Bertorelle 2001). It can be modelled in three ways: 1) as the number of chromosomes carrying the ancestral allele at a linked marker locus; 2) as the length of a conserved haplotype shared by all copies of the allele; and 3) as the number of mutations at one or more linked marker loci. Slatkin & Bertorelle (2001) illustrate this using the data from the CCR5 locus and two closely linked microsatellite markers, produced by Stephens et al. (1998) (See Section1.3.5.3).

The frequency of the Δ32 allele at CCR5 exceeds 10% in Europeans, yet it appears to be relatively young (see Stephens et al. 1998). In their study, Stephens et al. (1998) surveyed 46 chromosomes carrying Δ32 and found that found 44 carried the 197 allele at one closely linked microsatellite, and 41 carried the 215 allele at another closely linked microsatellite locus. The combination of the high Δ32 allele frequency and extremely low variability at the two microsatellite loci is consistent with the pattern expected for a locus that has experienced recent selection. Slatkin & Bertorelle (2001) conducted a formal analysis of intra-allelic variability on the CCR5 data and confirmed the findings of Stephens and colleagues (1998). In addition, Slatkin & Bertorelle (2001) used a method described by Slatkin (2001) to estimate the selection intensity from these data and estimated the selection coefficient in favour of Δ32 to be at least 0.2. The selection coefficient is used to define the relative fitness of alleles in a population, where a selection coefficient of 0.1 represents a 10% decrease in fitness compared to the fittest allele (see Jobling et al. 2004).

This chapter will describe the analysis of data from 6 microsatellites closely linked to the amylase gene cluster to examine the evidence for selection at the AMY1 locus and investigate the evolution of the amylase gene cluster in humans. Two approaches will be used: the first involves comparing genetic distances between populations based on data from 6 microsatellites closely

linked to the AMY gene cluster, to a null distribution based on the data set of 332 microsatellite loci used by Kayser et al. (2003). The second approach uses an analysis of intra-allelic variability on data from compound haplotypes comprising of AMY1 repeat alleles and 6 closely linked microsatellites.

## 5.2 Methods

### *5.2.1 Sample collection*

DNA samples from families consisting of two parents and at least one child were typed for the six microsatellite loci closely linked to the amylase gene cluster as well as for AMY1 repeat alleles (see section 2.3.1b). Buccal cells were collected, and the DNA extracted (See section 2.2) from the following populations: Africa (Ethiopia), the Middle East/Western Asia (Armenia), Europe (Germany, UK, Eire, Ashkenazi Jews), East Asia (Singapore Chinese). The multiplex PCR protocol (see sections 2.3.2) was used to genotype individuals for microsatellite repeat number for 6 microsatellite closely linked to the AMY1 gene cluster. Section 3.6 describes the development of this protocol as well as the process of selecting the microsatellites to be included. Family samples were used so that compound haplotypes of both microsatellite and *EMamy* estimates of AMY repeat alleles (see Sections 2.7 & 3.5) could be inferred by following the pattern of co-inheritance of the alleles from the parents to the children, (see Nehati-Javeremi & Smith 1996) (See Fig 2.1).

### *5.2.2 Statistical Analysis*

Statistical analyses were performed on the microsatellite data and compound haplotype as detailed in sections 2.8.4, 2.8.5 and .2.8.6.

## 5.3 Results

### *5.3.1 Variation in microsatellite allele frequencies in different populations*

As would be expected with microsatellite loci, all markers showed high levels of polymorphism. Table 5.1 shows the distribution of microsatellite alleles in the 7 populations under study. For all microsatellite loci except D1S2888, the smallest range of microsatellite alleles present in a population is found in the

Singapore Chinese sample. In the case of D1S2888 the largest range of microsatellite alleles was found in the British sample. The largest range of alleles for the six microsatellite loci were not consistently found in any one population: Ashkenazi Jews for D1S2888, Ethiopians for D1S2759, British for D1S2896, Ashkenazi Jews for AMY-MS02, and Armenians for D1S2626 (See table 5.1)

Microsatellite haplotype data was analysed by an AMOVA, as well as using an exact test of population differentiation based on haplotype frequencies, and genetic distances measured using $R_{ST}$ implemented using the Arlequin program (Schneider et al. 2000). $R_{ST}$ is analogous to $F_{ST}$ but incorporates into the step-wise mutation model of microsatellite evolution (see Slatkin 1995). Genetic distances between populations from the microsatellite data, using $R_{ST}$, show that the greatest difference was between the Armenians and Singapore Chinese ($R_{ST}$ = 0.11756). Significant differences were found between the Singapore Chinese and all population except Ireland (see Table 5.2). The exact test of population differentiation based on haplotype frequencies did not find any significant differences between any of the populations under study (See Table 5.3). AMOVA analysis found that 1.07% of the variation was to be found among populations, with 98.93% to be found within populations (See Table 5.4).

Table 5.1: Microsatellite repeat allele range, mode and variance for each population group. “High” indicates the largest allele for the locus found in each population; “low” indicates the smallest allele for the locus found in each population. “Mode” indicates the most frequent allele size found in each population. “Variance” indicates the variance in allele size for the population. The population with the lowest values for the summary statistics for each microsatellite locus are marked in light grey; the population with the highest values for the summary statistics for each microsatellite locus are marked in dark grey.

| | Armenia | Ashkenazi | British | Ethiopia | Germany | Ireland | Sing Chi. | All |
|---|---|---|---|---|---|---|---|---|
| D1S2888 | | | | | | | | |
| High | 21 | 21 | 21 | 25 | 22 | 21 | 21 | 25 |
| Low | 16 | 11 | 17 | 16 | 16 | 15 | 16 | 11 |
| Range | 5 | 10 | 4 | 9 | 6 | 6 | 5 | 14 |
| Mode | 18 | 20 | 19 | 19 | 19 | 19 | 19 | 19 |
| Variance | 1.355 | 2.182 | 0.744 | 1.933 | 1.295 | 1.821 | 2.624 | 1.701 |
| D1S2759 | | | | | | | | |
| High | 22 | 22 | 20 | 20 | 19 | 20 | 19 | 22 |
| Low | 12 | 14 | 14 | 7 | 14 | 14 | 17 | 7 |
| Range | 10 | 8 | 6 | 13 | 5 | 6 | 2 | 15 |
| Mode | 17 | 17 | 17 | 17 | 17 | 17 | 17 | 17 |
| Variance | 4.627 | 1.453 | 1.347 | 3.221 | 1.652 | 1.659 | 0.476 | 2.310 |
| D1S2896 | | | | | | | | |
| High | 18 | 19 | 20 | 18 | 18 | 18 | 18 | 20 |
| Low | 11 | 11 | 11 | 11 | 11 | 11 | 13 | 11 |
| Range | 7 | 8 | 9 | 7 | 7 | 7 | 5 | 9 |
| Mode | 14 | 14 | 14 | 14 | 14 | 14 | 14 | 14 |
| Variance | 2.774 | 3.887 | 5.706 | 3.489 | 4.645 | 5.833 | 2.544 | 4.105 |

| | Armenia | Ashkenazi | British | Ethiopia | Germany | Ireland | Sing Chi. | All |
|---|---|---|---|---|---|---|---|---|
| AMYMS02 | | | | | | | | |
| High | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| Low | 15 | 8 | 15 | 11 | 15 | 15 | 16 | 8 |
| Range | 5 | 12 | 5 | 9 | 5 | 5 | 4 | 12 |
| Mode | 17 | 17 | 17 | 17 | 17 | 17 | 17 | 17 |
| Variance | 0.586 | 1.900 | 1.335 | 1.137 | 1.197 | 1.007 | 0.952 | 1.223 |
| D1S535 | | | | | | | | |
| High | 12 | 11 | 12 | 11 | 11 | 11 | 11 | 12 |
| Low | 8 | 8 | 8 | 7 | 9 | 9 | 9 | 7 |
| Range | 4 | 3 | 4 | 4 | 2 | 2 | 2 | 5 |
| Mode | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |
| Variance | 0.470 | 0.746 | 0.577 | 0.548 | 0.624 | 0.351 | 0.444 | 0.555 |
| D1S2626 | | | | | | | | |
| High | 25 | 24 | 25 | 24 | 24 | 24 | 25 | 25 |
| Low | 16 | 17 | 17 | 17 | 17 | 17 | 20 | 16 |
| Range | 9 | 7 | 8 | 7 | 7 | 7 | 5 | 9 |
| Mode | 21 | 21 | 21 | 21 | 21 | 21 | 21 | 21 |
| Variance | 4.044 | 2.709 | 3.193 | 3.413 | 3.452 | 3.588 | 1.358 | 3.294 |

Table 5.2: Comparisons of pairs of populations microsatellite data measured using $R_{ST}$. $R_{ST}$ values are on the lower left of the table, associated P values are on the upper right. Comparisons that are significant (P=0.005) are in bold and shaded. $R_{ST}$ were calculated according to Slatkin (1995) and P values were estimated by permutation, implemented in the Arlequin software (Schneider et al 2000).

| | Armenia | Ashkenazi Jews | British | Ethiopians | Germans | Irish | Singapore Chinese |
|---|---|---|---|---|---|---|---|
| Armenia | * | 0.50223 +-0.0050 | 0.14335 +-0.0033 | 0.50581 +-0.0054 | 0.29621 +-0.0041 | 0.20414 +-0.0037 | **0.00020 +-0.0001** |
| Ashkenazi Jews | -0.00225 | * | 0.60172 +-0.0047 | 0.73329 +-0.0043 | 0.76478 +-0.0042 | 0.88526 +-0.0031 | **0.00307 +-0.0006** |
| British | 0.01228 | -0.00477 | * | 0.18315 +-0.0041 | 0.55994 +-0.0055 | 0.61766 +-0.0053 | **0.02871 +- 0.0013** |
| Ethiopians | -0.00221 | -0.00551 | 0.00803 | * | 0.37551 +-0.0045 | 0.30839 +-0.0044 | **0.00010 +-0.0001** |
| Germans | 0.00518 | -0.01073 | -0.00702 | 0.00162 | * | 0.69181 +-0.0041 | **0.01802 +-0.0012** |
| Irish | 0.00956 | -0.01299 | -0.00797 | 0.00355 | -0.01264 | * | 017711 +-0.0043 |
| Singapore Chinese | 0.11756 | 0.06228 | 0.04750 | 0.09506 | 0.06208 | 0.01659 | * |

Table 5.3 Exact Test of Population Differentiation – Microsatellite data
Non differentiation exact P values. Comparisons that are significant (P<0.05) are shaded.
Table 5.2: Comparisons of pairs of populations microsatellite data measured using $R_{ST}$. $R_{ST}$ values are on the lower left of the table, associated P values are on the upper right. Comparisons that are significant (P=0.005) are in bold and shaded. $R_{ST}$ were calculated according to Slatkin

| | Armenia | Ashkenazi Jews | British | Ethiopians | Germans | Irish | Singapore Chinese |
|---|---|---|---|---|---|---|---|
| Armenia | * | | | | | | |
| Ashkenazi | 0.27374+-0.0214 | * | | | | | |
| British | 0.74511+-0.0153 | 0.06649+-0.0173 | * | | | | |
| Ethiopians | 0.89590+-0.0129 | 0.30402+-0.0306 | 1.00000+-0.0000 | * | | | |
| Germans | 0.55671+-0.0257 | 0.13072+-0.0138 | 1.00000+-0.0000 | 1.00000+-0.0000 | * | | |
| Irish | 0.54834+-0.0180 | 0.28874+-0.0199 | 1.00000+-0.0000 | 1.00000+-0.0000 | 1.00000+-0.0000 | * | |
| Singapore Chinese | 0.30987+-0.0143 | 0.03573+-0.0044 | 0.77358+-0.0177 | 0.62196+-0.0275 | 0.48987+-0.0189 | 0.48964+-0.0206 | * |

Table 5.4 AMOVA for microsatellite data (Wier & Cockerham 1984, Excoffier et al. 1992, Wier 1996)

| Source of variation | Percentage of Variation | P value |
|---|---|---|
| Among populations | 1.07 | 0.03842+/- 0.00204 |
| Within populations | 98.93 | |

The distributions of repeat alleles from the six microsatellites for the different AMY1 polygenic repeat alleles in all populations are shown in Fig 5.1 a-f. For AMY1*H0, H1 and H2, the modal repeat allele for each microsatellite is the same. However for AMY1*H3 the modal repeat allele for D1S2888 is 20 repeats and for AMY-MS02 is 18, whereas for the other AMY1 polygenic repeat alleles the modal microsatellite allele is 19 and 17 repeats respectively.

*5.3.2 Comparison with other microsatellite loci in the human genome*

As explained in chapter 4, one method to distinguish between differential selection in different populations and drift is to look not only at the locus in question but also at the rest of the genome. Drift and demographic events affect the whole genome, whereas selection operates on a particular locus. An outlier on a genome wide distribution of $R_{ST}$ vales would suggest that a process other than drift had influenced the allele frequencies at that particular locus.

$R_{ST}$ data for microsatellites typed in 48 Europeans (blood donors from Leipzig, Germany) and 23 East Africans (from Gondar, Ethiopia) as well as 24 Southern Africans (from the Nguni, Sotho-Tswanga and Tsonga groups of South Africa) was taken from a dataset of 332 microsatellites typed by Kayser et al. (2003). Fig 5.2a-f show comparisons of the $R_{ST}$ values for Germany vs Ethiopia for the AMY microsatellites with $R_{ST}$ values for 332 microsatellites from the human genome typed in Europeans (Germans) and Africans. None of the microsatellite show unusually high $R_{ST}$s compared to the distribution from the 332 loci typed by Kayser et al. (2003).

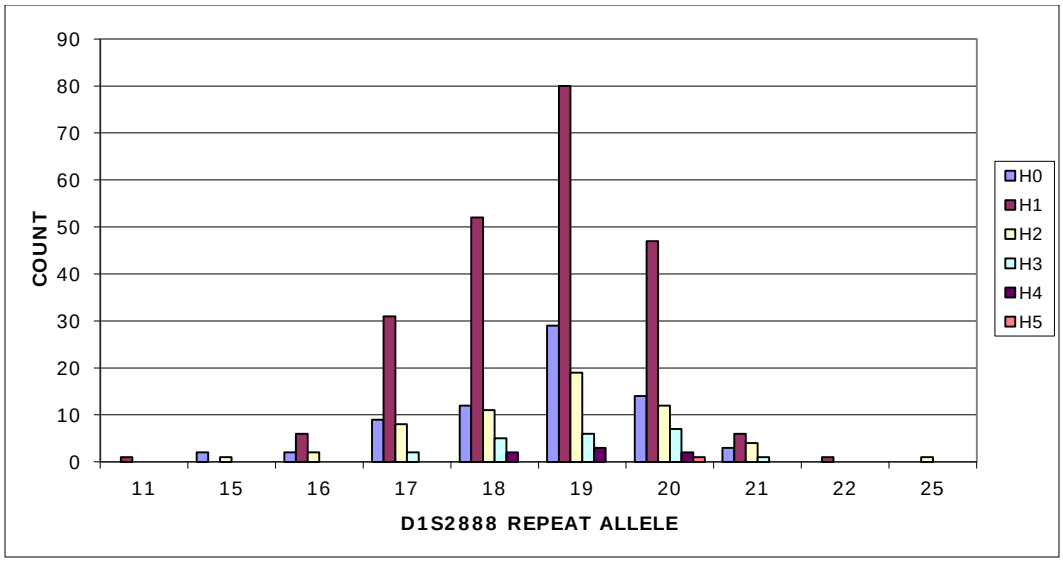



Fig 5.1a: The distribution of D1S2888 alleles for the AMY1 repeat alleles,

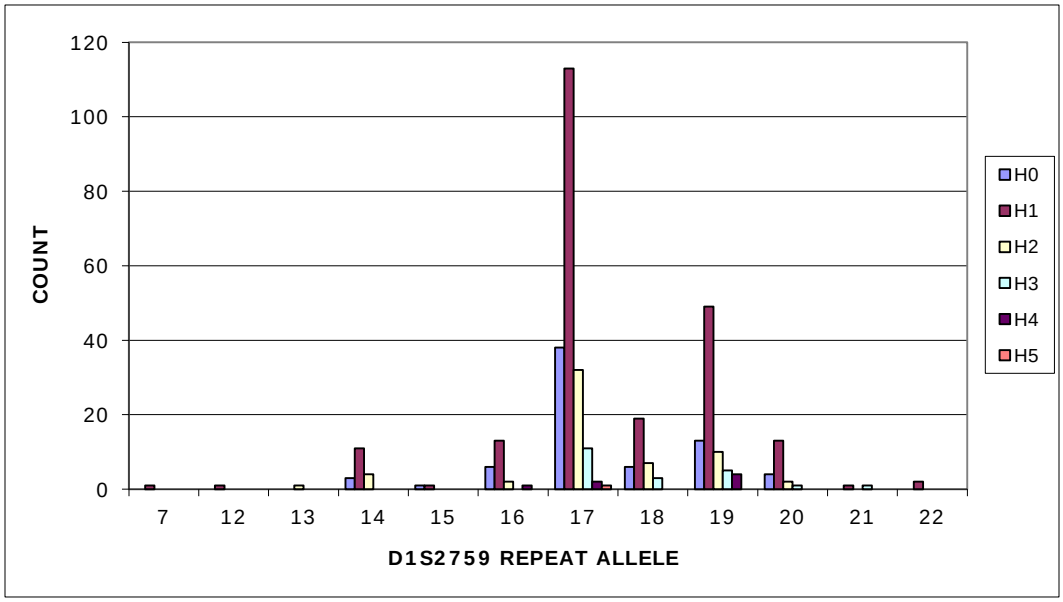



Fig 5.1b: The distribution of D1S2759 alleles for the AMY1 repeat alleles

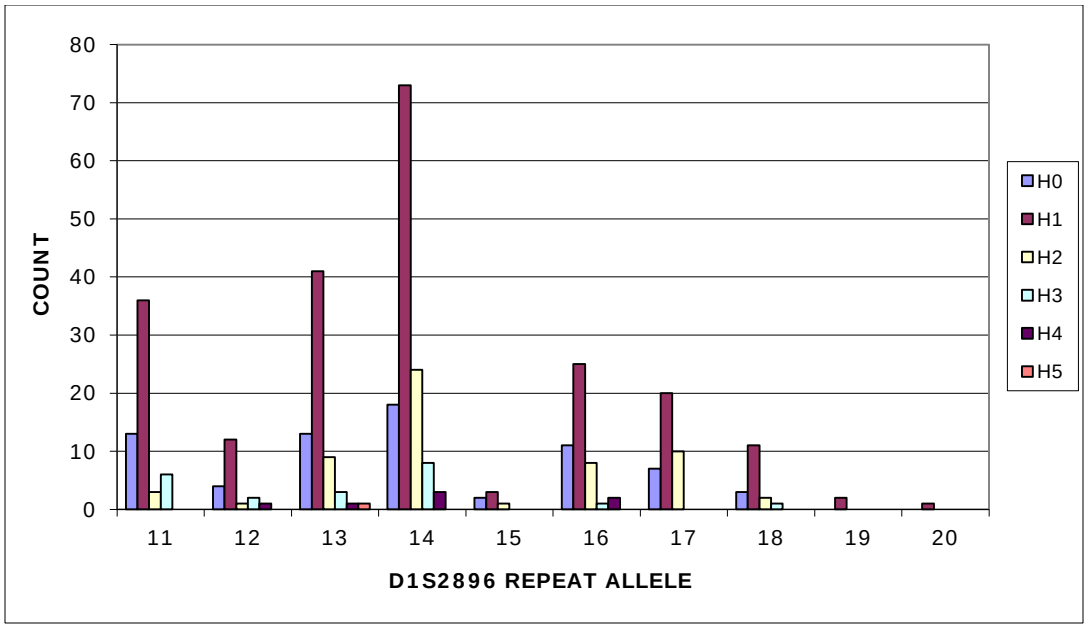



Fig 5.1c: The distribution of D1S2896 alleles for the AMY1 polygenic alleles

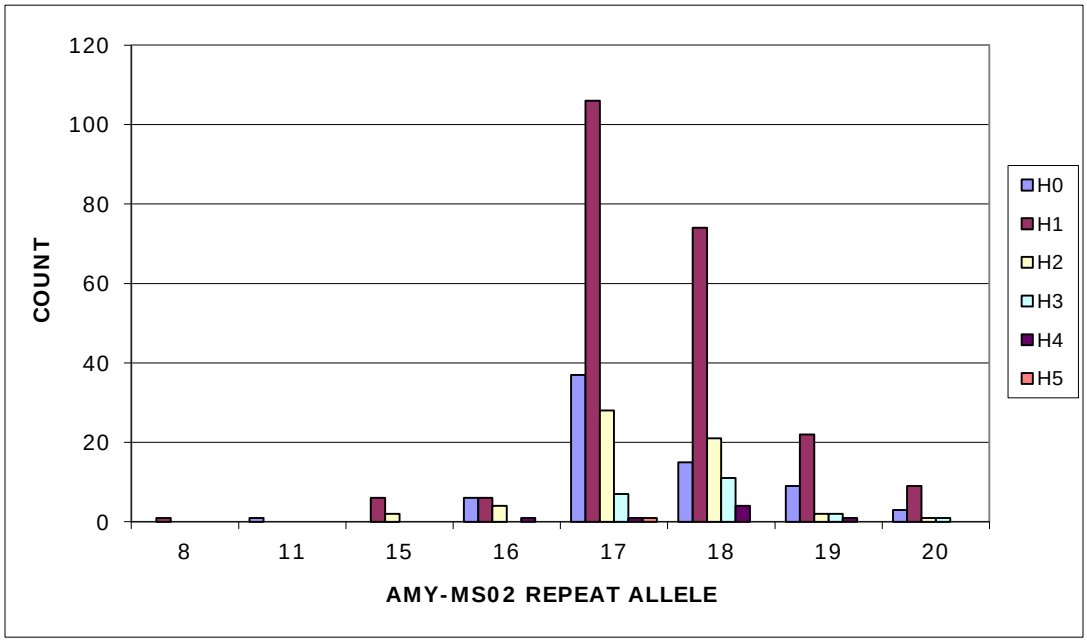



Fig 5.1d: The distribution of AMY-MS02 alleles for the AMY1 repeat alleles

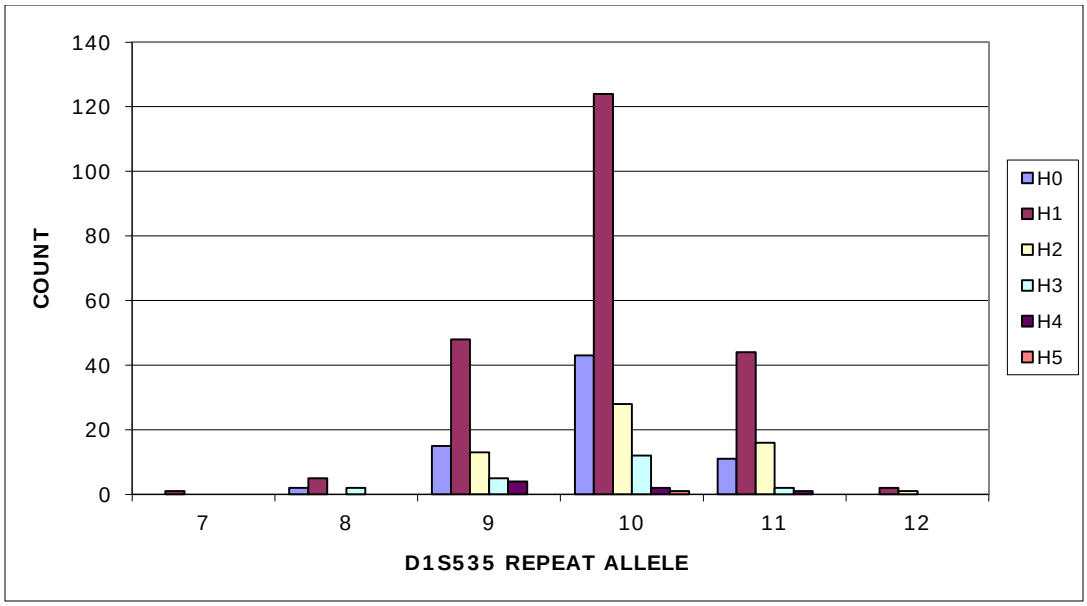



Fig 5.1e: The distribution of D1S535 alleles for the AMY1 polygenic alleles

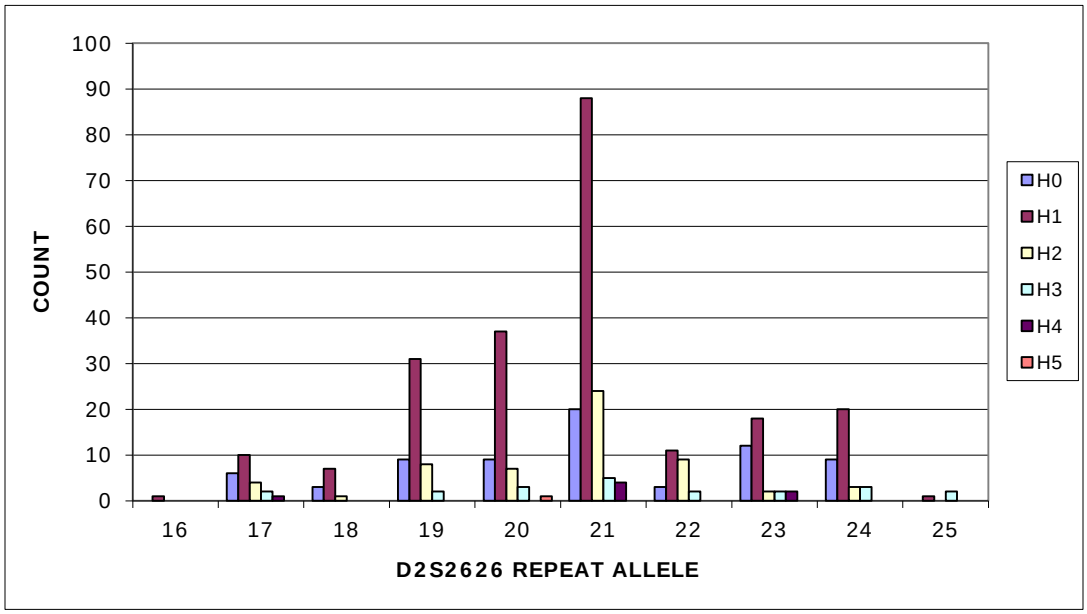



Fig 5.1f: The distribution of D1S2626 alleles for the AMY1 repeat alleles

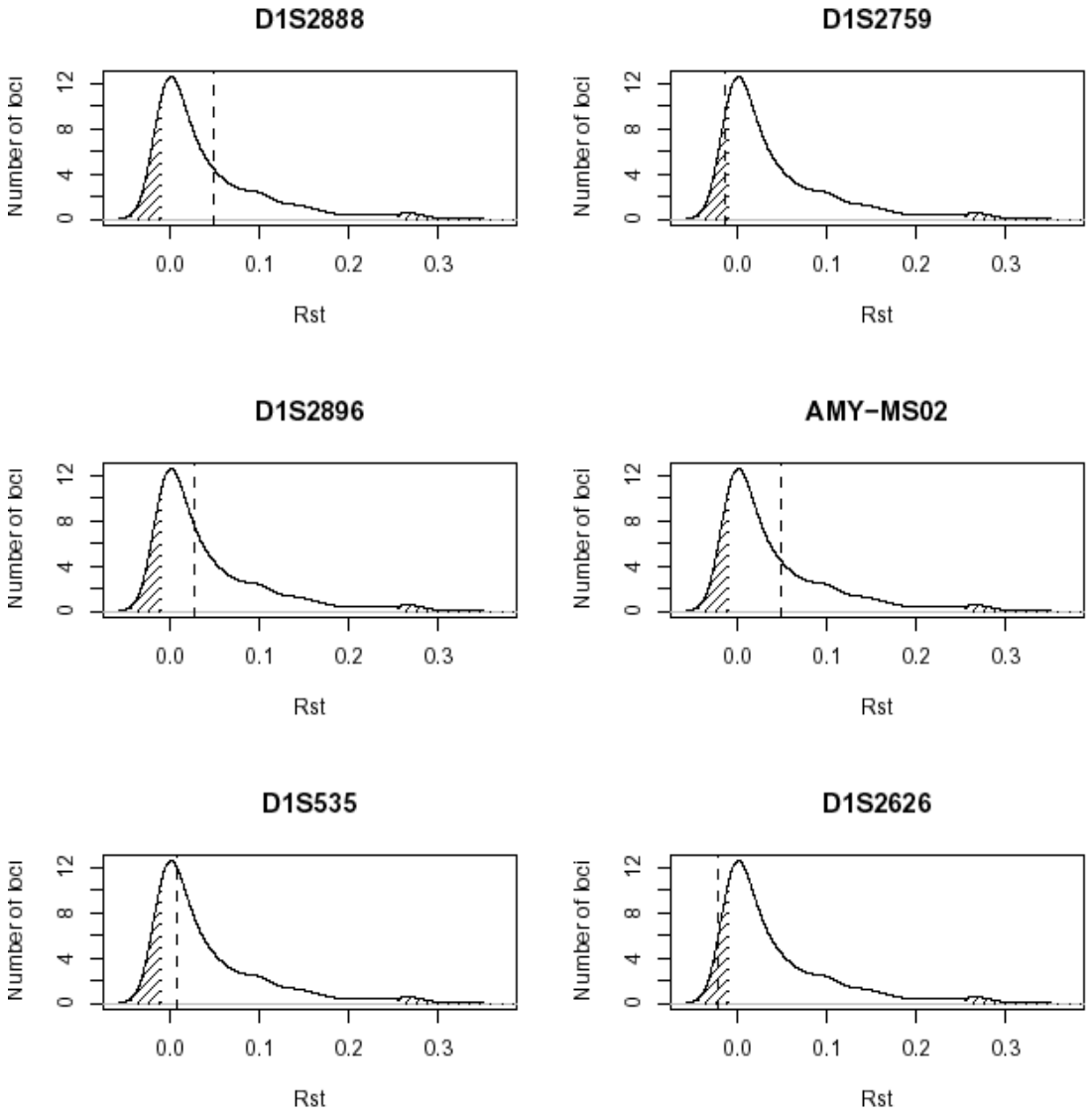



Fig 5.2a-f: Density plots of the distribution of $R_{ST}$ values for 332 microsatellites typed in Africans and Germans (Kayser et al 2003). Statistical analysis of the microsatellite data was performed using the statistics package 'R' (URL: http://www.R-project.org/). The $R_{ST}$ values for the 6 microsatellites closely linked to the amylase gene cluster between Ethiopians and Germans (D1S2888 RST= 0.04928, D1S2749 RST= -0.01341, D1S2896 RST=0.02737, AMY-MS02 RST= 0.04906, D1S535 RST = 0.0069, D1S2626 RST= -0.02181) are shown by the dotted lines. The shaded areas represent the 2.5% (-0.011) and 97.5% (0.24575) quantiles of the distribution.

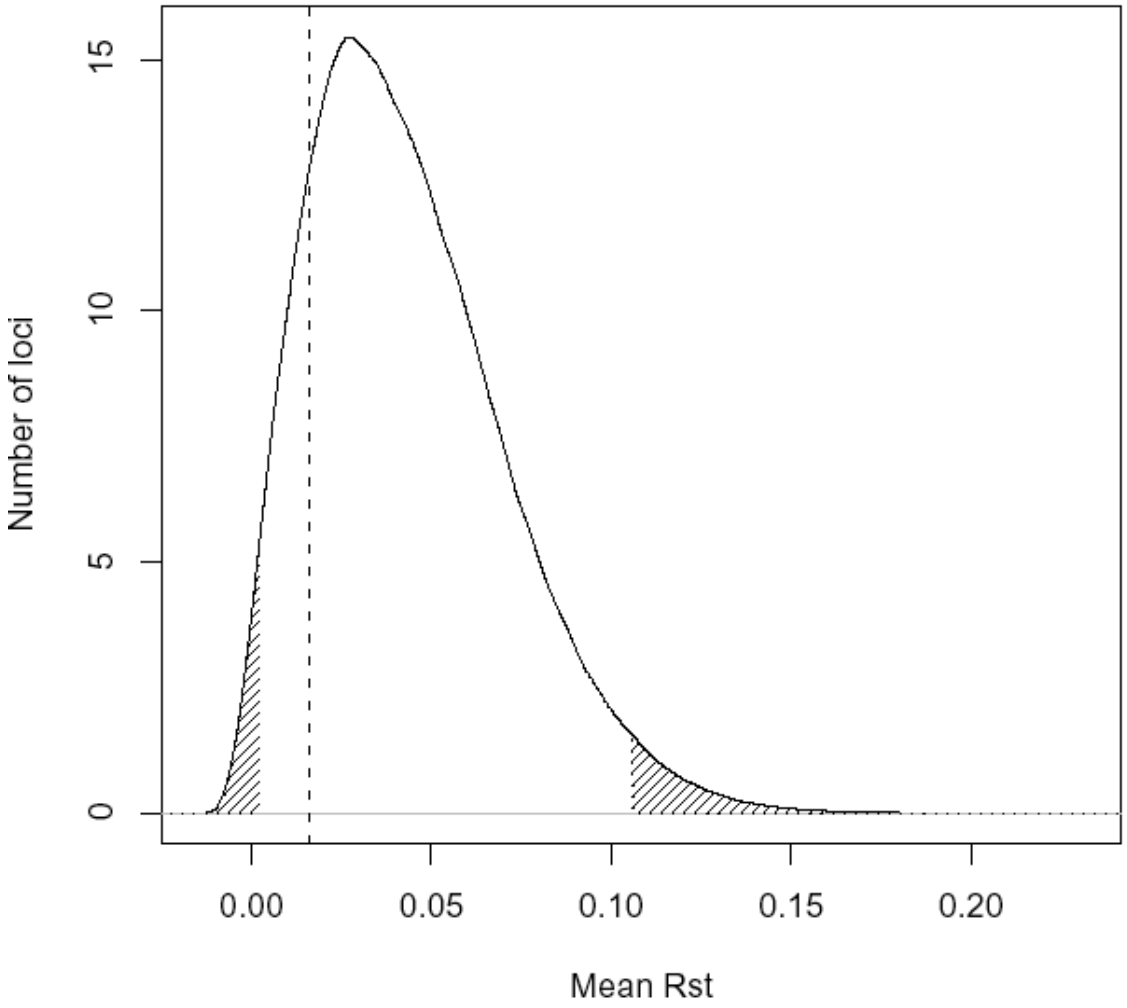



Fig 5.2g: Plot of the distribution of average $R_{ST}$ values for randomly chosen groups of six loci from a dataset of 332 microsatellites typed in Africans and Germans (Kayser et al 2003). Statistical analysis of the microsatellite data was performed using the statistics package 'R' (URL: http://www.R-project.org/). The average $R_{ST}$ value for the 6 microsatellites closely linked to the amylase gene cluster between Ethiopians and Germans ($R_{ST}$=0.0162316) is shown by the dotted line.

However, two of the microsatellites (D1S2759 and D1S2626) did have unusually low $R_{ST}$ values compared to the distribution of Kayser et al.'s (2003) dataset. Unusually low $R_{ST}$s can be interpreted signals of balancing selection (see Akey et al. 2002). This pattern was not found when the $R_{ST}$ values for the six AMY microsatellites were averages and compared to a distribution of average $R_{ST}$ values from groups of six loci randomly chosen from the Kayser et al. (2003) dataset (See Fig 5.2g).

### *5.3.3 Microsatellite variance and AMY1 repeat allele frequencies*

The average variance in microsatellite repeat alleles for each AMY1 allele can be viewed as a crude proxy for age, as the older an allele is, the more variation there will be at linked microsatellites, both as a result of mutation and recombination. Old alleles should be at a higher frequency in the population than young ones, as it takes time for alleles to drift from low to high frequency (see Slatkin & Bertorelle 2001). If an allele high frequency, but has low microsatellite variability associated with it (i.e. it is young) it must have gone from low to high frequency very quickly. This scenario is consistent with selection operating on the allele, as strong selection would drive an allele to high frequency in less time than it would take through drift. Fig 5.3 shows the average variance in microsatellite repeat alleles for the AMY1 repeat alleles, plotted against their frequency. Fig 5.3a shows data form all 6 microsatellites, and in this figure AMY1*H1 stands out as having a similar average microsatellite variance as the other alleles, but is at a markedly higher frequency compared with the other AMY1 repeat alleles. However, using data from only the two closest microsatellites to the AMY gene cluster, this pattern does not appear so striking (Fig 5.3b).

### *5.3.4 Analysis of intra allelic variability*

Analysis of intra-allelic variability was carried out on compound haplotypes consisting of the AMY1 repeat allele as well as the 6 microsatellites using the program SYSSIPHOS written by Dr Michael Stumpf, Imperial College London. The program estimates the likelihood of the data over a range of selection coefficients ($s$) and population growth rates ($r$) for a given allele age.

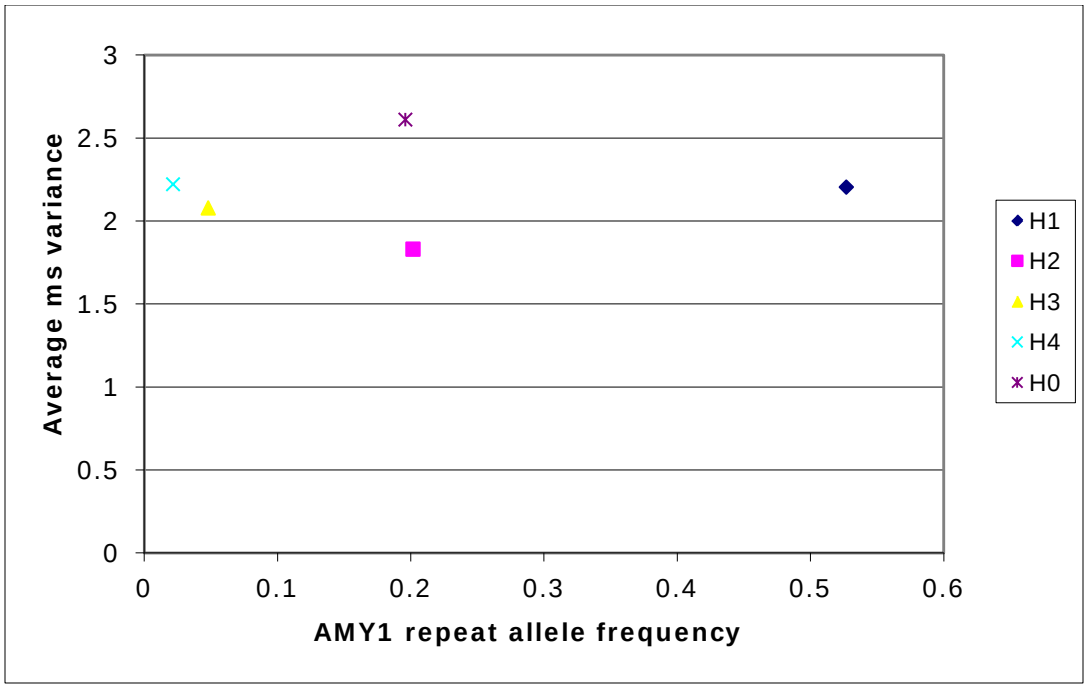



Fig 5.3a: AMY1 repeat allele frequency and the variance in microsatellite repeat alleles averaged over 6 microsatellites closely linked to the amylase gene cluster.

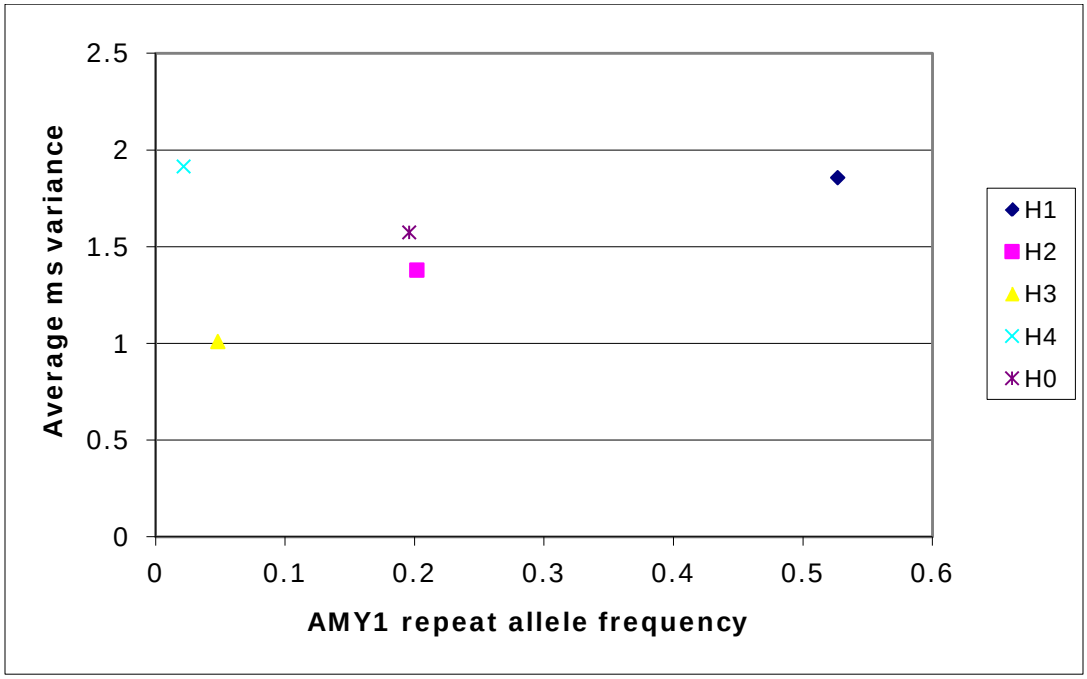



Fig 5.3b: AMY1 repeat allele frequency and the variance in microsatellite repeat alleles averaged over the 2 microsatellites (D1S2759 & AMY-MS02) closest to the amylase gene cluster.

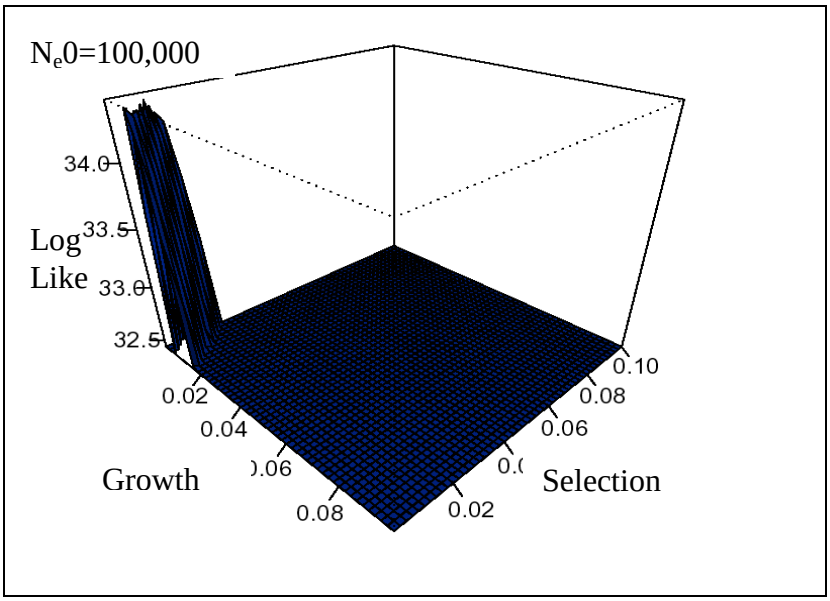
The following parameters were used in the analysis of intra-allelic variability using the SYSSIPHOS program (see Appendix C for an example of a SYSSIPHOS parameter input file):

Initially the likelihood surface for selection and growth parameters was explored on a 50x50 grid with growth rate parameters (r) ranging from 0 to 1.0 in increments of 0.02, and selection parameters of 0.0001 to 0.1 with interval increases of 0.002 (where 1-s is the relative fitness of homozygotes for the allele under investigation and 1+s/2 is the relative fitness of herterozygotes and 1 = the relative fitness of homozygotes for all other genotypes. The initial wide incremental interval of selection coefficients was used in order observe where the dataset suggested the highest likelihood values. Following this, a second fine scale analysis was then carried out using selection parameters of 0.0001 to 0.006.

The maximum depth of the amylase gene cluster gene trees (tmax) was set to 100,000 generations. Preliminary tests suggested that the method is relatively insensitive to current population size (See Fig 5.4a,b). The current population size ($N_e0$) was set at 10,000,000. A microsatellite mutation rate (mu) of 0.0012 per locus per generation (Weber & Wong 1993) was assumed.

The probability of recombination per generation (rho) (See table 5.5) for each microsatellite was calculated using both physical distances from the July 2003 assembly of the UCSC Human Genome working draft, and sex averaged recombination rates from deCODE (http://www.decode.com). In addition departures from the stepwise mutation model (Slatkin 1995) such that a length dependent microsatellite mutation rate (see Stumpf and Goldstein 2001) is taken into account. The slope (a) and intercept (b) of the length dependence were calculated to be –3.1 and 0.62 respectively, where k = allele length:

$$mu(k) = mu(a\ k+b)$$



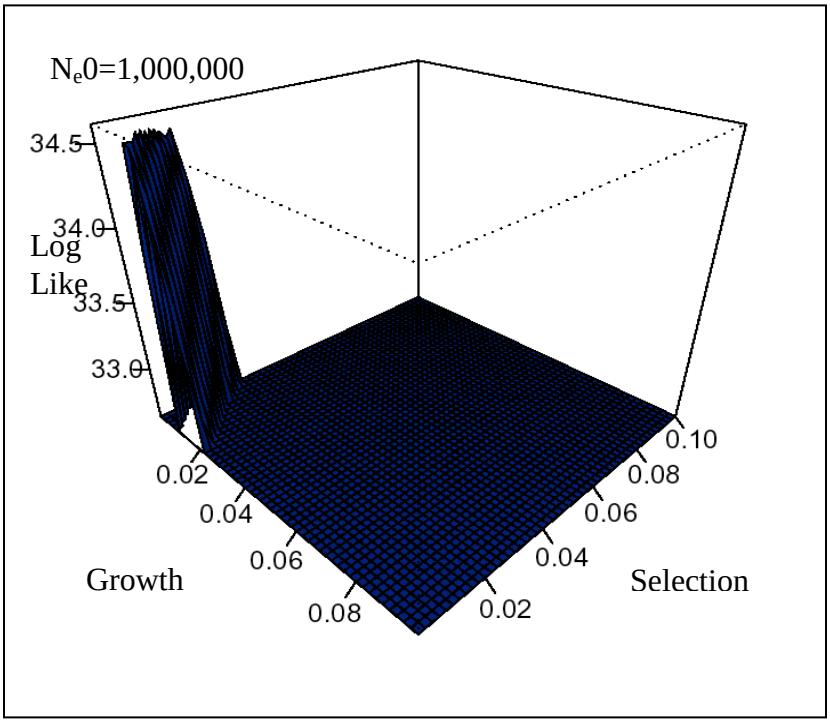




Fig 5.4: The effect of current population size ($N_e0$) on maximum likelihood estimates for selection (s) and growth (r) for the World data set. Post-processing of the SYSSIPHOS output files was carried using the statistics package 'R' (URL: http://www.R-project.org/).

| Recomb Block (and recomb rate) / Marker Name and Position (bp). | | Block 1: 1.3 cM/Mb | Block 2: 0.2 cM/Mb | Block 3: 0.3 cM/Mb | Block 4: 1.2 cM/Mb | Block 5: 0.9 cM/Mb | cM |
|---|---|---|---|---|---|---|---|
| D1S2896 | 101435399 | 564601 | 1000000 | 553028 | - | - | 1.0998897 |
| D1S2626 | 102458557 | - | 541443 | 553028 | - | - | 0.274197 |
| AMY-MS02 | 103449926 | - | - | 103103 | - | - | 0.0309309 |
| AMY2B | 103452226 | - | - | 100803 | - | - | 0.0302409 |
| AMY2A | 103514651 | - | - | 38378 | - | - | 0.0115134 |
| AMY1A | 103553029 | - | - | 0 | - | - | 0 |
| D1S535 | 103731135 | - | - | 178106 | - | - | 0.0534318 |
| D1S2888 | 104444355 | - | - | 446971 | 444354 | - | 0.6673161 |
| D1S2759 | 105199386 | | | 446971 | 1000000 | 199385 | 1.5135378 |

Table 5.5: Recombination distances for amylase genes and six microsatellites closely linked to the amylase gene cluster. Numbers in the table indicate the number of base pairs in each recombination block between the marker and the start codon of the AMY1A gene. Marker positions are taken from the July 2003 assembly of the UCSC Human Genome working draft. Recombination rates for the blocks are taken from deCODE sex average recombination rates (http://www.decode.com). Recombination distances (cM) are calculated by multiplying the number of base pairs in each block by the recombination rate for that block, and then finding the sum of the results for the blocks for each marker. This is then converted to cM by multiplying the sum by 1,000,000.

The log likelihoods and selection coefficients (*s*) for the AMY1 repeat alleles in the Ethiopian, Armenian and Western European (British, Irish & German) populations are shown in Fig 5.5a,b,c, when the growth rate is set to zero. Only populations with sufficient data (n>9 haplotypes for each AMY1 repeat allele) were used in the analysis of intra-allelic variability. In the Ethiopian sample the AMY1*H1 allele showed the highest value for *s* (*s*=0.0281), compared to the other AMY1 repeat alleles (AMY1*H1 > AMY1*H2 > AMY1*H0) (See Table 5.6). With the Armenians, the AMY*H0 allele showed the highest value for *s* (*s*=0.0361) compared to the other AMY1 repeat alleles (AMY1*H0 > AMY1*H1 > AMY1*H2). In the Western European population group, the AMY1*H2 allele showed the highest value for *s* (*s*=0.0221) compared to the other AMY1 repeat alleles (AMY1*H2 > AMY1*H1 = AMY1*H0). However none of the populations showed significant differences in *s* between the AMY1 repeat alleles using a likelihood ratio test (see section 2.8.6).

The maximum likelihood estimates of growth, when s=0 (see Table 5.6) are in agreement with previous estimates of global and European growth rates of 1.2% and 1.6% respectively (Cavalli-Sforza et al. 1994, see also Wilson et al. 2003 for a review).

### 5.3.5 Estimating the time to the most recent common ancestor for the AMY1 repeat alleles.

Dates for the TMRCA were obtained through both the analysis of intra-allelic variability using the SYSSIPHOS program. Having estimated the maximum likelihood for parameters of s and r given the data, for a fixed AMY1 repeat allele age, the average age in generation for each AMY1 repeat allele is estimated using SYSSIPHOS, given the microsatellite data for fixed maximum likelihood values of *s* and *r* (See Slatkin 2001).

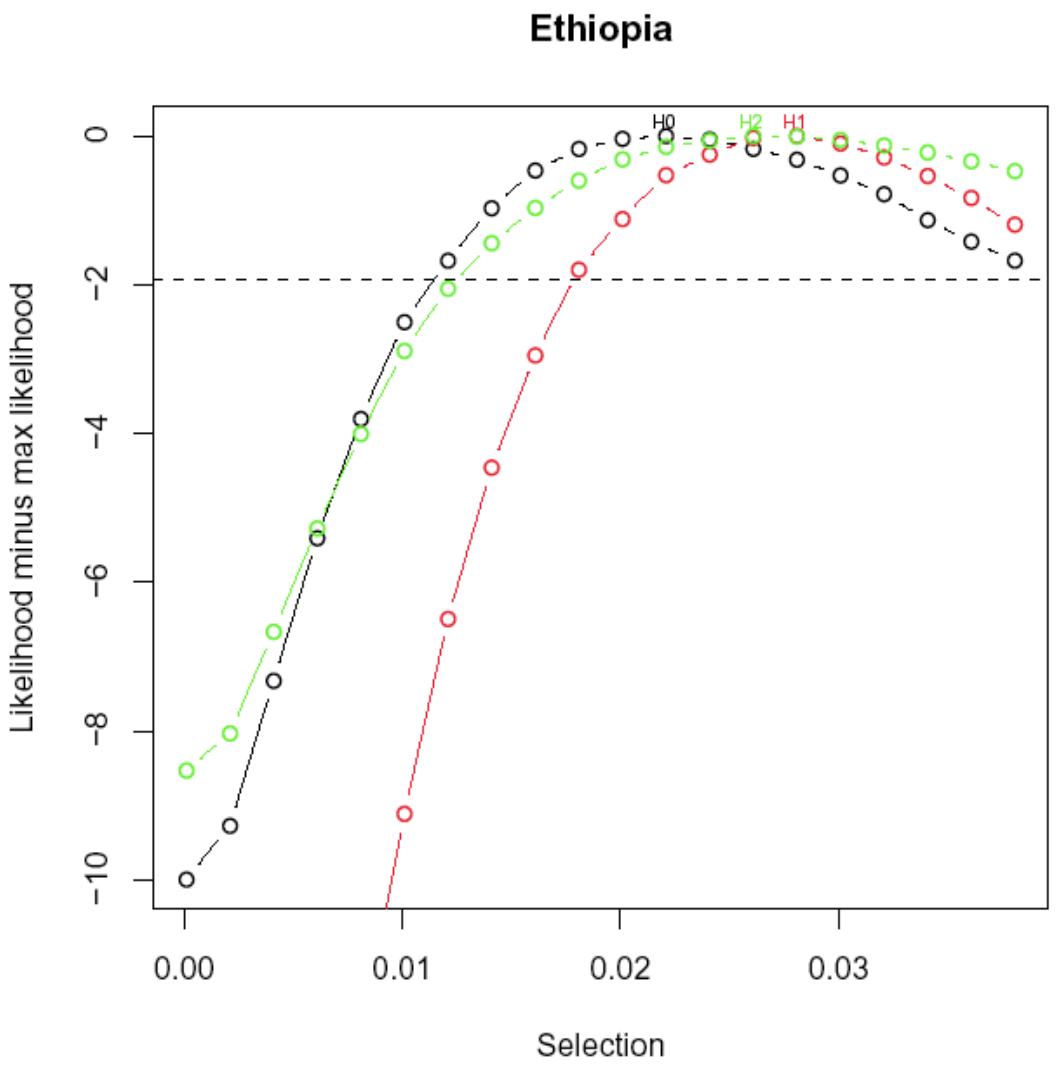



Fig 5.5a: A plot of the trasnformed log likelihoods (L) vs selection coefficients (s) from the Ethiopian data set for AMY1 repeat alleles when growth = 0. Log likelihoods for each value of s were calculated for each of the AMY1 repeat alleles separately from microsatellite haplotype data using the SYSSIPHOS program (M.Stumpf, Imperial College London). The dotted line represents a –1.92 reduction in likelihood from the maximum.

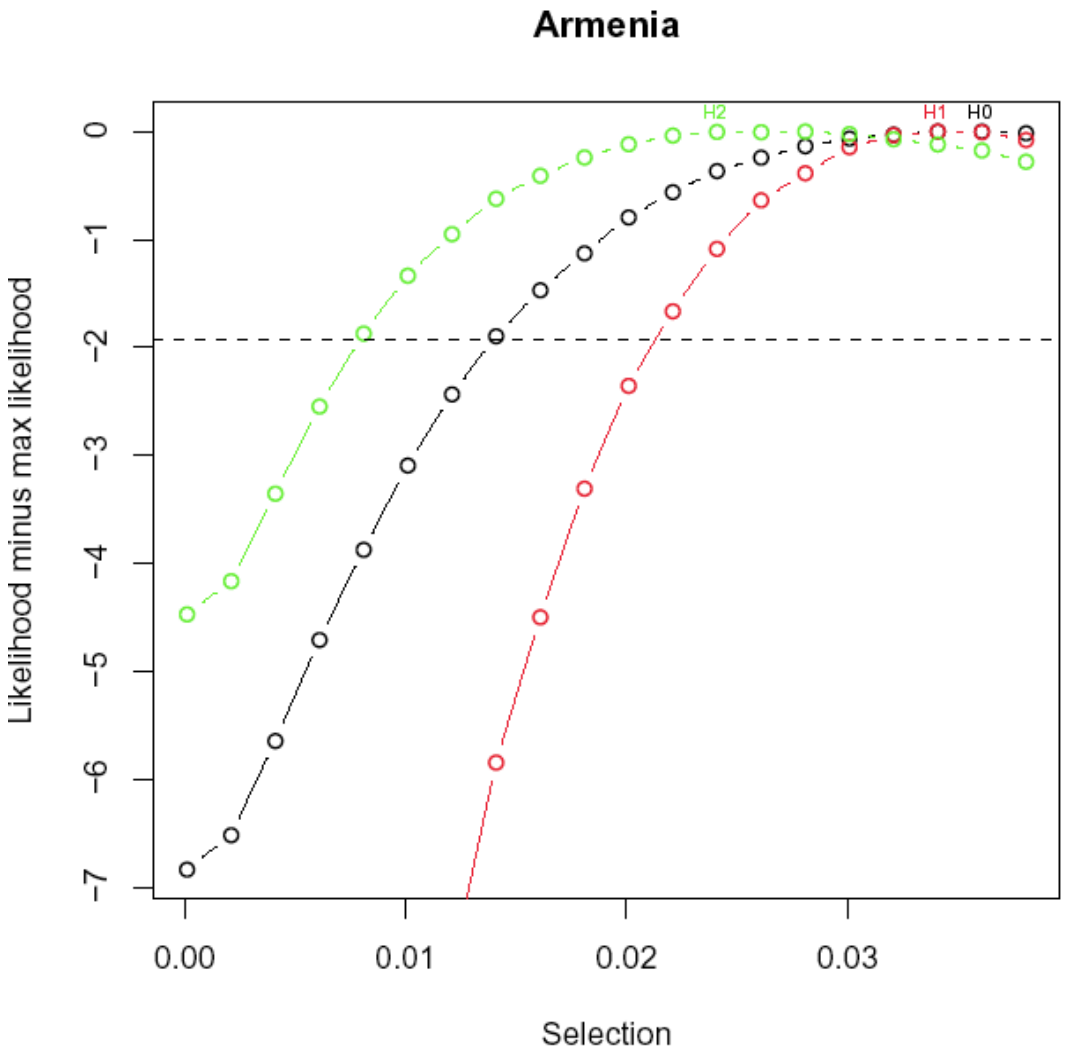



Fig 5.5b: A plot of the transformed log likelihoods (L) vs selection coefficients (s) from the Armenian data set for AMY1 repeat alleles when growth = 0. The dotted line represents a –1.92 reduction in likelihood from the maximum.

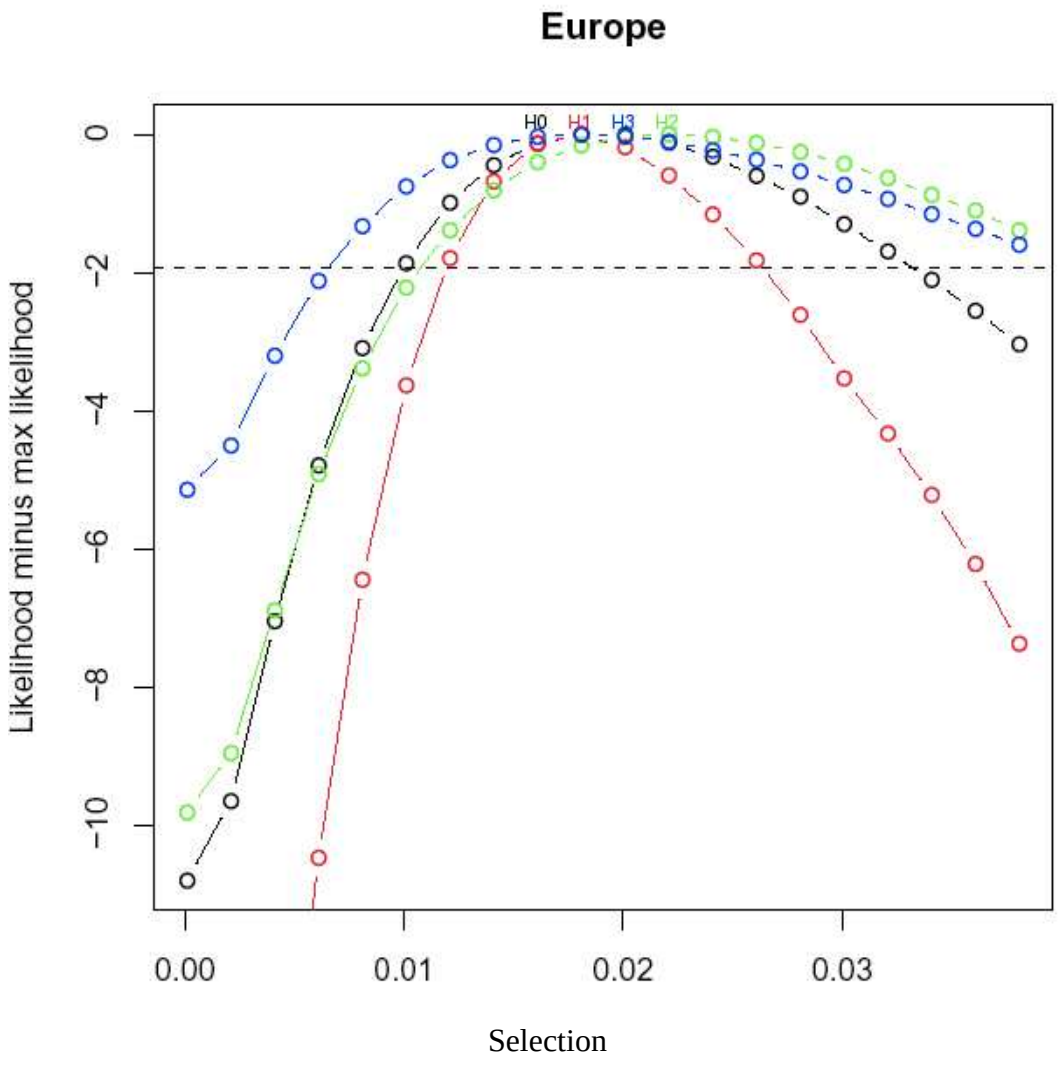



Fig 5.5c: A plot of the transformed log likelihoods (L) vs selection coefficients (s) from the Ethiopian data set for AMY1 repeat alleles when growth = 0. The dotted line represents a –1.92 reduction in likelihood from the maximum.

| Population / AMY1 repeat allele | Armenia | | | | Ethiopia | | | | Western Europe | | | | World | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Max like estimate of s when r=0 | Log Like | Max like estimate of r when s=0 | Log Like | Max like estimate of s when r=0 | Log Like | Max like estimate of r when s=0 | Log Like | Max like estimate of s when r=0 | Log Like | Max like estimate of r when s=0 | Log Like | Max like estimate of s when r=0 | Log Like | Max like estimate of r when s=0 | Log Like |
| AMY1*H0 | 0.036 | 6.839 | 0.032 | 6.854 | 0.022 | 9.994 | 0.020 | 10.02 | 0.018 | 10.80 | 0.016 | 10.84 | 0.026 | 34.64 | 0.024 | 34.69 |
| AMY1*H1 | 0.034 | 23.01 | 0.030 | 23.06 | 0.028 | 26.89 | 0.024 | 26.95 | 0.018 | 23.86 | 0.016 | 23.92 | 0.036 | 91.73 | 0.020 | 98.20 |
| AMY1*H2 | 0.028 | 4.473 | 0.024 | 4.493 | 0.026 | 8.528 | 0.024 | 8.550 | 0.022 | 9.818 | 0.020 | 9.843 | 0.020 | 21.69 | 0.018 | 29.17 |

Table 5.6: Maximum likelihood estimates of selection (s) and growth (r) for different AMY1 repeat alleles in different human populations, calculated using the SYSSIPHOS program (M.Stumpf, Imperial College London). Western Europe = British, Irish and German samples.

In addition, a microsatellite variance (ASD) method was used to estimate the age of the AMY1 repeat allele, using the YTIME program written by Dr Michael Weale, University College London (see Thomas et al. 2002). The Average Squared Distance (ASD) between an ancestral microsatellite haplotype and all other haplotypes has been shown to be linearly related to the time since the most recent common ancestor (TMRCA) under an unbound stepwise mutation rate (Goldstein et al. 1995, Slatkin 1995)

$$ASD = \mu t$$

Where $\mu$ = mutation rate and t = time.

ASD is calculated from the data without the need to construct a phylogeny. However an ancestral haplotype must be specified. In this analysis the modal haplotype was taken to be the ancestral (See Stumpf & Goldstein 2001 for a review). Growth was set at zero, and a length-dependent (see Calabrese et al. 2001, Kruglyak et al. 1998) microsatellite mutation rate where $\mu$=0.0012 (Weber & Wong 1993) and average microsatellite repeat length for the 6 microsatellites was 16.407.

Table 5.7 shows the estimates of TMRCA for the AMY1 repeat alleles from all the chromosomes typed for the six microsatellite loci and AMY1 repeat alleles. The point estimates obtained from SYSSIPHOS are within the 95% confidence intervals obtained from the ASD method implemented in the YTIME program (Dr M. Weale, University College London). AMY1*H0 appears to have the largest TMRCA (506.5 – 2725.1 generations), followed by AMY1*H1 (416 – 2314.5 generations) and AMY1*H2 (338.3 – 1963.3 generations). All of the estimates place the TMRCA for the chromosomes within the last 70,000 years.

| AMY1 repeat allele | | SYSSIPH max like point estimate when r=0 | SYSSIPH max like point estimate when s=0 | YTIME Point estimate | YTIME upper quantile | YTIME lower quantile |
|---|---|---|---|---|---|---|
| AMY*H0 | gens | 594.06 | 586.16 | 1,089.1 | 506.5 | 2,725.1 |
| | yrs | 14,851 | 14,654 | 27,228 | 12,663 | 68,128 |
| AMY1*H1 | gens | 416.37 | 625.82 | 921.51 | 416 | 2314.5 |
| | years | 10,409 | 15,646 | 23,038 | 10,400 | 67,863 |
| AMY*H2 | gens | 644.01 | 632.97 | 773.89 | 338.3 | 1,963.3 |
| | years | 16,100 | 15,824 | 19,347 | 8,457.5 | 49,083 |

Table 5.7: Time to the most recent common ancestor (TMRCA) estimates for the AMY1 repeat alleles from a global sample of all chromosomes typed for the six microstellite loci and AMY1 repeat allele. Generation time was taken as 25 years. YTIME upper quantile corresponds to 97.5% and the lower quantile to 2.5%, so that 95% of the data are within the upper and lower quantiles.

## 5.4 Discussion

The comparisons of $R_{ST}$ for the AMY microsatellites with 332 microsatellite loci in the genome failed to show any unusually large differences in the distribution of microsatellite alleles between populations for the AMY microsatellites. However, this method is somewhat crude and only loci with extreme allele frequency differences would stand out from the null distribution.

In addition, there are a number of problems with the comparative data set of 332 microsatellites from the study by Kayser et al. (2003). Firstly, the 332 microsatellites were typed in Europeans and Africans. The European sample was identified as 48 blood donors from Leipsig, Germany. This sample matches well with the German sample typed for the AMY markers. However, the African sample typed by Kayser and colleagues (2003) consisted of 23 Ethiopians and 23 Bantu speaking South Africans. It is well established that higher genetic diversity occurs in Africa than anywhere else in the world and that non-Africans carry a small subset of African diversity (Bowcock et al. 1994, Tishkoff et al. 1996, Kaessman et al. 1999, Vigilant et al. 1991). Therefore it is especially important to ensure that African groups are not lumped together. Using a mixed African population would serve to make the comparison more conservative, as Ethiopians are more closely related to Europeans than a mixed group of Sub-Saharan Africans are to Europeans (see Wilson et al. 2001). Unfortunately it was not possible to obtain the data from Kayser et al. (2003) with the African populations separated out.

The second problem with using Kayser et al.'s (2003) data set is the relatively small number of microsatellite markers. This dataset of 332 microsatellite loci gives an average of only 14 markers per chromosome, which leaves much of the genome unrepresented.

As explained in chapters 1 and 4, analysis of intra-allelic variability currently provides the most powerful method for detecting selection (see Sabeti et al. 2002). The analysis of intra-allelic variability unfortunately failed to show compelling evidence for selection. In the Ethiopian and Western European groups, the chromosomes with duplicated amylase

elements (AMY1*H1, H2, H3 etc) showed a slightly stronger signal of positive selection than those without. This is consistent with the hypothesis outlined in Chapter 1 that AMY repeat alleles with higher number of AMY1 genes, confer an selective advantage in populations that consume high starch diets. In contrast the Armenian sample shows the opposite pattern, where the highest selection coefficient was found for the AMY1*H0 allele. It must be noted however that the number of AMY1*H0 chromosomes in the Armenian sample is very low (n=9). Interestingly the strongest signal for selection ($s$=0.0361, when r=0) appears in the Armenians, which have the longest history of cereal agriculture of all the populations in this study.

The range of TMRCA estimates for the AMY1 repeat alleles are all with the last 70,000 years, with point estimates in the region of 20,000 years. This is an intriguing result since there is clear evidence in the archaeological record of an increase in consumption of high starch content foods around this time, such as the charred remains of wild grass species found at Wadi Kubbaniya that have been dated to 17,000-16,000 years (Hillman 1989). However the estimates of TMRCA are also consistent with a population bottleneck due to a common and recent origin African origin for all non-African human populations within the last 44,000 – 200,000 years (Tishkoff et al. 1996, Tishkoff & Williams 2002). A population bottleneck may reduce microsatellite diversity in such a way as to give the impression of a more recent origin than the actual age. Thus the microsatellite diversity associated with the AMY1 repeat alleles may result from population demography rather than by the actual age of the mutation and the selective pressures that it has been subjected to (if any). It is clear, therefore, that caution should be applied when interpreting the results of compound haplotype based dating methods.

# Chapter 6: Variation in salivary amylase gene copy number in Chimpanzees

## 6.1 Introduction

Variation in salivary amylase gene copy number has been found in all human populations studied to date (see chapter 4, section 4.3). The question remains, however, of whether this inter-individual variation in AMY1 gene copy number is found in other primate species. If it is the case that AMY1 gene copy number variation is only present in humans, then it is likely that the duplication events that gave rise to the different AMY1 repeat alleles have occurred since the human/chimpanzee lineages diverged. The aim of this chapter was to investigate the extent of variation in AMY1 gene copy number in chimpanzees.

### *6.1.1 What can chimpanzees tell us about human evolutionary genetics*?

Chimpanzees and humans are estimated to have shared a common ancestor between 4.6 and 6.2 million years ago (Chen & Li 2001). Early comparative studies of humans and chimpanzee genomes estimated the extent of DNA sequence difference to be around 1.6% (see Ebersberger 2002 for review). More recent studies, sampling large numbers of sequences from across the genome (Chen & Li 2001, Ebersberger 2002), have shown the average extent of sequence divergence to be in the region of 1.24%. These figures seem relatively low considering the phenotypic differences, such as large brains, bipedalism and language, that we can see between humans and our closet living relatives. However the key is not in the number of differences, but where they are in the genome and what traits they affect. It has been suggested that differences between chimpanzees and humans lie not in different genes, but in differences in the control of gene expression (Enard et al. 2002*a*). Thus relatively small genetic changes could have a large impact on the resulting organism. Enard et al. (2002*a*) identified species-specific gene expression patterns in chimpanzees and humans, which indicate that changes in protein and gene expression have been particularly pronounced in the human brain. Fortna et al. (2004) studied gene duplication in five hominoid species and found that genes showing copy number expansions were most common in humans, and included a number of gene thought to be involved in the structure and function of the brain.

Enard et al. (2002*b*) have also studied a gene that is involved in language acquisition in humans. It seems that two functional copies of the FOXP2 gene are required for normal speech development. Through a comparison with FOXP2 from chimpanzee and other primates, as well as studying the variation in humans they were able to suggest that human FOXP2 has been a target of natural selection in recent human evolution. However it not possible to distinguish between recent natural selection in the hominid lineage, or a relaxation of selective constraints based on the data of Enard et al. (2002*b*).

Studies using chimpanzee sequence data can also be used to shed light on human demographic history and evolution. The rate of sequence divergence can be used to estimate the age of the most recent common ancestor for a given locus, the ages of individual lineages and the timings of prehistoric migrations as well as the ages of demographic events such as population expansions and bottlenecks. For example, comparisons of human and chimpanzee sequence information has provided evidence for reduced genetic diversity in humans and signatures of population expansion (Kaesmann et al. 2001).

As our closest living relatives, chimpanzees have also been used as an outgroup in many human phylogenetic studies. An outgroup is a lineage known to be more distantly related to the other lineages under study than they are to each other. The outgroup serves to root phylogenetic trees, and in addition provides a method for testing whether the rate of evolution is constant over all evolutionary lineages. Yi et al. (2002) used this method to provide evidence that a reduction has occurred in the rate of mitochondrial evolution in the hominoids compared to other higher primates.

*6.1.2 How can chimpanzees inform us in the study of human dietary adaptation?*

Chimpanzees also provide us with an invaluable tool in the study of human dietary adaptation. Chimpanzee diets are relatively unaffected by human inventions such as agriculture and are assumed to have remained the same since their evolution (see Milton 1999, 2000). It is therefore likely that the diets of chimpanzees are closer to the common ancestor of humans and chimpanzees than that of humans living today. It can therefore be predicted that chimpanzees have not adapted at the genetic level to human foods such as cows milk or large quantities of cereal products.

Studies of chimpanzees in the wild have revealed that the chimpanzee diet is made up of 65% fruit, 20 % leaves from arboreal plant species and 5% meat and insects (Newton Fisher 1998, Basabose 2002). They remainder of their diets is made up of honey, sap, leaves and stems of terrestrial plant species, nuts, seeds and bird's eggs. They exhibit a wide range of feeding behaviours, including hunting of prey such as wild bush pigs, monkeys and even small antelope (Stanford et al. 1994). Chimpanzees also use tools to acquire food such as hammer stones to crack nuts (Boesch & Boesch 1983) and twigs to 'fish' termites out of nests and to dip honey (Boesch & Boesch 1993).

As was discussed in Chapter 1 (section 1.2) , adaptive explanations are often used to explain differences in eating patterns between different populations, and are similarly used to explain differences in feeding behaviours between humans and chimpanzees. If, when we explore the genetic variation in humans and chimpanzees for these traits, we do find not any difference between the two species, we can reject such adaptive explanations.

*6.1.3 Chimpanzees and Amylase*

There have been only a small number of studies of amylase in non-human primates. McGeachin & Akin (1982) reported that amylase was found in the saliva and pancreas in gorillas, orang-utans and chimpanzees. As was outlined in chapter 1 (section 1.5.1) Samuelson et al. (1990, 1996) investigated amylase gene promoter regions in New-World monkeys, Old-World monkeys and apes. They found that a γ–actin pseudogene was

integrated after the divergence of the New-World monkeys from the primate ancestral tree and a retroviral sequence was integrated later after the divergence of the Old-World monkeys. They found that all human amylase (pancreatic and salivary) genes contain the γ-actin insert and therefore conclude that all the human amylase genes diverged from each other after this insertion event approximately 40 million years ago (Samuelson et al. 1990).

Samuelson et al. (1990) also reported the relative amounts of the amylase hybridising fragments they obtained. The relative intensities from the AMY2B and AMY2A were identical in the human and the chimpanzee samples. However the AMY1 fragment in the human DNA hybridised three times more intensely than the fragment in the chimpanzee DNA. They suggested that the human salivary gene had been amplified from one copy to three copies after the divergence of humans and chimpanzees. They also pointed out that this recent date for the salivary gene duplication is consistent with the lack of sequence divergence between the three human salivary amylase genes. However, this study was based on a single DNA sample from each species. Chapter 4
(Section 4.3) shows data from a large number of humans from 14 different populations. Variation in AMY1 gene copy number was found in all human populations studied. This chapter aims to study a number of chimpanzee DNA samples in order to explore the extent of variation in AMY1 gene copy number in chimpanzees.

**6.2 Methods:**

Five chimpanzee (*Pan troglodytes*) DNA samples (see Ruano et al. 1992) were obtained with kind permission from Prof Dallas Swallow, Dept. Biology University College London. The human QAMY02 and QAMY03 protocols were performed on the five samples according to the methods described in section 2.3.1(b).

PCR and cloning were performed as detailed in Section 2.5. Sequencing of the cloned PCR products was carried out according to the protocol described in section 2.4. The final reaction conditions for the chimpanzee QAMY02 assay are detailed in section 2.6.

Chromatograms produced from sequencing, by the ABI3100, were analysed using Sequencher v4.0 (GeneCodes, Ann Arbor, MA). Human and chimpanzee sequences were aligned and converted to NEXUS format. Phylogenetic analysis was conducted using both UPGMA distance, neighbour joining, maximum likelihood and exhaustive parsimony methods implemented in PAUP*V4,β10 (Sinauer Associates, Sunderland, Mass).

**6.3 Results**

The ratios of peaks of fluorescence obtained from the human QAMY02 and QAMY03 protocols were consistent across all five of the chimpanzee DNA samples. QAMY02 primers amplified two different sized PCR products in the chimpanzee samples as in humans. The PCR products also conformed to the same sizes as the human products. Consistent results were obtained across 2 PCR reactions and 2 electrophoresis runs per PCR (total 4 runs) for each DNA sample (See Table 6.1). However, the ratio of the peak heights of the PCR products to each other in chimpanzee did not correspond to any of the expected values for the different genotypes found in humans. The average ratio of 191bp product : 187bp products was 0.8061 : 1. One explanation for this was that the primers were not amplifying all the amylase genes in chimpanzee with equal efficiency due to a base change in chimpanzees, in the region where the primer anneals in one of the amylase genes. This would mean that the resulting fragments would only be amplified efficiently from 2 or less out of the 3 classes (AMY2A, AMY2B, AMY1) of genes intended to be amplified. The expected ratio for QAMY02 is calculated using all the amylase genes. If one or more genes are not amplified with equal efficiency then the ratio produced by the PCR products would not conform to the expected ratios.

| Chimp ID | QAMY02 187bp peak height | QAMY02 191bp peak height | Ratio of 191:187bp peak heights | QAMY03 263bp peak height | QAMY02 267bp peak height | Ratio of 263:267 bp peak heights |
|---|---|---|---|---|---|---|
| Harv | 2429 | 2037 | 0.8386 : 1 | / | 2830 | 0 |
| | 1052 | 927 | 0.8812 : 1 | / | 1978 | 0 |
| Colin | 536 | 443 | 0.8264 : 1 | / | 1259 | 0 |
| | 741 | 636 | 0.8583 : 1 | / | 760 | 0 |
| Kassay | 215 | 160 | 0.7442 : 1 | / | 1342 | 0 |
| | 85 | 66 | 0.7764 : 1 | / | 2694 | 0 |
| Carl | 1875 | 1467 | 0.7824 : 1 | / | 1744 | 0 |
| | 2172 | 1702 | 0.7836 : 1 | / | 2854 | 0 |
| Tank | 336 | 257 | 0.7649 : 1 | / | 2222 | 0 |
| | 282 | 227 | 0.8050 : 1 | / | 2982 | 0 |

Table 6.1: Data from the human QAMY02 & QAMY03 protocols on five chimpanzee DNA samples. Two PCRs were performed on each DNA sample. Data in the table are the heights of the peaks of fluorescence from electrophoresis using an ABI377 of fluorescently labelled PCR products, as displayed by the ABI/GeneScan™ software.

In humans, the QAMY03 assay primers amplify two PCR products of different length, which is visualised as two peaks of fluorescence when run on an ABI377/GeneScan™ system. In the chimpanzee samples only one peak (263bp) was seen. The deletion (+1586-1570bp from start codon) in AMY1 found in humans that results in the smaller peak (263bp), was presumed to not be present in chimpanzees.

In order to confirm the hypothesis that the unusual ratios obtained with the human QAMY02 protocol was in fact due to base changes in the area where the primers anneal it was necessary to sequence the chimpanzee samples. PCR products were amplified that spanned the region of interest using primers AMY-05-U (5'- CTG GAA AGG ACA CTG ACA AC -3') and AMY-06-L (5'- ATC TAG TCA CCA TGT TTC TAA ATT CAT -3'). These were primers originally designed to amplify a region from the start of exon 1, through intron 1, and into exon 2 of all three amylase genes in humans. As exon sequences are highly conserved it was likely that these primers would also amplify all three classes of amylase genes in chimpanzees. The PCR products from the chimpanzees were cloned into a plasmid vector before sequencing in order to produce gene specific DNA fragments.

A total of 24 colonies were picked from the plates and sequenced in both the forward and reverse directions (See Section 2.5). Of the 24 colonies sequenced, 11 contained AMY1 sequence, 7 contained AMY2A sequence and 5 contained AMY2B sequence, and one colony gave only a very short amylase sequence. Several studies have reported that products obtained after PCR with *Taq* polymerase will contain hybrid molecules when several homologous target sequences such as multigene families are co-amplified (see Shafikhani 2002, Judo et al. 1998, Meyerhans et al. 1990, Wang & Wang 1996, Shammas et al. 2001). Wang & Wang (1996) reported a 30% occurrence of chimeric sequences in a 30 cycle PCR amplification of nearly identical 16S rRNA genes from several bacterial species. Out of the 24 chimpanzee amylase clones in this study, seven (4 x AMY1A; 2 x AMY2B and 1 x AMY2A) were interpreted as having

```
            1          11         21         31         41         51
            |          |          |          |          |          |
Human AMY1A aaagcaaaAT GAAGCTCTTT TGGTTGCTTT TCACCATTGG GTTCTGCTGG GCTCAGTATT
Chimp AMY1A .......... .......... .......... .......... .......... ..........
Human AMY2A .......... ....T..... CT........ .......... .......... ..........
Chimp AMY2A .......... ....T..... CT........ .A........ .......... ..........
Human AMY2B .......... ....T..... CT........ .......... .......... ..........
Chimp AMY2B .......... ....T..... CT........ .......... .......... ..........

            61         71         81         91         101 QAMY02-CH-U primer
            |          |          |          |          |          |
Human AMY1A CCTCAAATAC ACAACAAGGA CGAACATCTA TTGTTCATCT GTTTGAATGG CGATGGGTTG
Chimp AMY1A ..C....... .......... .....G.... .......... .......... ..........
Human AMY2A ..C....... .......... ..G....... .......... .......... ..........
Chimp AMY2A ..C....... .......... ..G....... .......... .......... ..........
Human AMY2B ..C....... .......... ..G....... .......... .......... ..........
Chimp AMY2B ..C....... .......... ..G....... .......... .......... ..........

            121        131        141        151        161        171
            |          |          |          |          |          |
Human AMY1A ATATTGCTCT TGAATGTGAG CGATATTTAG CTCCCAAGGG ATTTGGAGGG GTTCAGgtgg
Chimp AMY1A .......... .......... .......... ....G..... .......... ..........
Human AMY2A .......... .......... .......... ....G..... .......... ..........
Chimp AMY2A .......... .......... .......... ....G..... .......... ..........
Human AMY2B .......... .......... .......... .......... .......... ..........
Chimp AMY2B .......... .......... .......... ....G..... .......... ..........

            181        191        201        211        221        231
            |          |          |          |          |          |
Human AMY1A gtatgattca tagtatcaat tgcagaattc actgtgcttg tagtaaacac tattctgatc
Chimp AMY1A .......... .......... .......... .......... .......... ..........
Human AMY2A .......... .......... ...G...... .......... ...G...T.G ..........
Chimp AMY2A .......... .......... ...G...... .......... ...G...T.G ......T...
Human AMY2B .......... .......... .......... ...A...... .......T.G ..........
Chimp AMY2B .......... .......... .......... .......... .......T.G ..........

            241        251        261        271        281        291
            |          |          |          |          |          |
Human AMY1A ttctacgtga agcttgggca acattttact tcacaggtaa gtattctaag taaaagaatt
Chimp AMY1A ......CC.. .....C..A. .......... .......... .......... ..........
Human AMY2A ..A.CT.... .......... .......... .......... .......... .......GA.
Chimp AMY2A ..A.CT.... .......... .......... .......... .......... .......GA.
Human AMY2B ..A.C..... .....A.... .......... .......... .......... ..:...G..
Chimp AMY2B A.A.C..... .....A.... .......... .......... .......... ..:...G..

            301        311        321        331        341        351
            |          |          |          |          |          |
Human AMY1A ttctgaggaa aaaacaatgt agtattcttt gcaactgtat attttgtttc tgatataatc
Chimp AMY1A .......... .......... .......... .......... .......... ..........
Human AMY2A ........G. ...T.T.... .........G ......T... .......... ...G......
Chimp AMY2A .......... ...T.T.... .........G ......T... .......... ...G......
Human AMY2B .......... .......... .........G ......T... .......... ...G......
Chimp AMY2B .......... .......... .........G ......T... .......... ...G......
            361        371        381  QAMY02-U primer  401        411
            |          |          |          |          |          |
Human AMY1A tttcttcaac aagagccctc cgatgtgctg ttaatatttt caagagatag ctgcctatac
Chimp AMY1A .......... .......... .......... .......... ....G..... .C........
Human AMY2A ........C. .......... .A........ .......... .......... ..........
Chimp AMY2A .......... .......... .A........ .......... .......... .......C..
Human AMY2B .......... .......... .A........ .......... .......... ....A.....
Chimp AMY2B .......... .......... .A........ .......... .......... ....A.....
```

```
           421        431        441        451        461        471
            |          |          |          |          |          |
Human AMY1A caagattcaa gaa:tctttt gtattattga ttagattcta gaacattcaa tgatatacag
Chimp AMY1A .......... .G........ A......... .......... .......... .........C
Human AMY2A .......... .......... A..C...... ....T..... .......... .....C....
Chimp AMY2A .......... T....T.... A......... ....T..... .......... ..........
Human AMY2B .......... .....T.... A......... ....T..... .......... .......G..
Chimp AMY2B .......... ...T.T.... A......... ....T..... .......... .......G..

            481        491        501        511        521        531
            |          |          |          |          |          |
Human AMY1A taagacagaa tttggtactt atgaagactg tttaatttgt agGTCTCTCC ACCAAATGAA
Chimp AMY1A .......... .........C .......... .......... .......... ..........
Human AMY2A .......... .......... .......... .......... .......... ..........
Chimp AMY2A .......... .......... .......... .......... .......... ..........
Human AMY2B ....::::.. .......G.. .......... .......... .......... ..........
Chimp AMY2B ....::::.. .......G.. .......... .......... .......... ..........

            541        551  QAMY02-L-TET primer         581        591
            |          |          |          |          |          |
Human AMY1A AATGTTGCCA TTCACAACCC TTTCAGACCT TGGTGGGAAA GATACCAACC AGTTAGCTAT
Chimp AMY1A .......... .......... .......... .......... .......... ..........
Human AMY2A ........A. ..T....... .......... .......... .......... ..........
Chimp AMY2A ........A. ..T....... .......... .......... .......... ..........
Human AMY2B ........A. .......... .......... .......... .......... ..........
Chimp AMY2B .......... .......... .......... .......... .......... ...C......

            601        611
            |          |
Human AMY1A AAATTATGCA CAA
Chimp AMY1A .......... ...
Human AMY2A .......... ...
Chimp AMY2A .......... ...
Human AMY2B .......... ...
Chimp AMY2B .......... ...
```

Fig 6.2: Sequence comparison of human and chimpanzee amylase genes exon1, intron & exon 2. Human sequences were obtained from contigs produced from aligning sequences of BAC clones (see chapter 3) as well as exon sequences from Genbank (AMYA: M17883, M18715, M18717, AMY2A: M18671, AMY2B: D90088, D90089). Chimpanzee sequences were obtained experimentally. Exon sequences are shown in capitals, intron sequences are shown in lower case letters. The regions in green show the position of the QAMY-02 primers. Note the base change (in red) in the chimpanzee AMY1A sequence close to the 3' end of the QAMY02U primer. The QAMY02-U primer was abandoned in favour of QAMY02-CH-U for the chimpanzee QAMY02 assay. The region in yellow shows the 4bp deletion in both chimpanzee and human AMY2B genes, around which the QAMY02 assay is designed.

undergone PCR-mediated recombination during the initial PCR before the products were cloned into the plasmid vector.

After sequencing has been carried out, three groups of sequences were identified and these were assigned, by sequence homology with the human amylase gene sequences, as AMY1A, AMY2A and AMY2B (See Fig 6.2 & 6.3a). Chimpanzee sequences for exon 1, intron 1 and exon 2 were submitted to the EMBL nucleotide sequence database (URL: http://www.ebi.ac.uk/embl/) and assigned accession numbers as follows: AMY1A: AJ703812; AMY2A: AJ703813; AMY2B AJ703814 (See Fig 6.2).

Fig 6.3a, b & c show phylograms for the human and chimpanzee amylase exon 1, intron1 and exon 2 sequences. After phylogenetic analysis, conducted using PAUP*V4,β10 (Sinauer Associates, Sunderland, Mass), exactly the same topology was obtained with likelihood, UPGMA distance and exhaustive parsimony methods (See fig 6.3 b&c). Maximum likelihood trees were searched for, both with and without the enforcement of the molecular clock. The molecular clock assumes that all branches have the same rate of evolution. However if one branch has a different rate of evolution compared to the other branches the molecular clock assumptions are violated. A likelihood ratio test was used to test for significant differences between the likelihoods obtained for the topologies with and without the molecular clock (see Felsenstein 1981). Significant differences in likelihoods were not found and so the molecular clock was appropriate for the data (See Fig 6.3b). The enforcement of the molecular clock enabled the tree to be rooted using midpoint rooting.

The results of sequencing from exon 1 to exon 2 of the chimpanzee amylase genes showed a number of base change differences between human and chimpanzees (see Fig 6.2 and 6.4). In particular, the 3' end of the upper primer from the human QAMY02 protocol (QAMY02-U) annealed to an area in the sequence that had a base change in the chimpanzee AMY1A gene (See Fig 6.2). This means that the

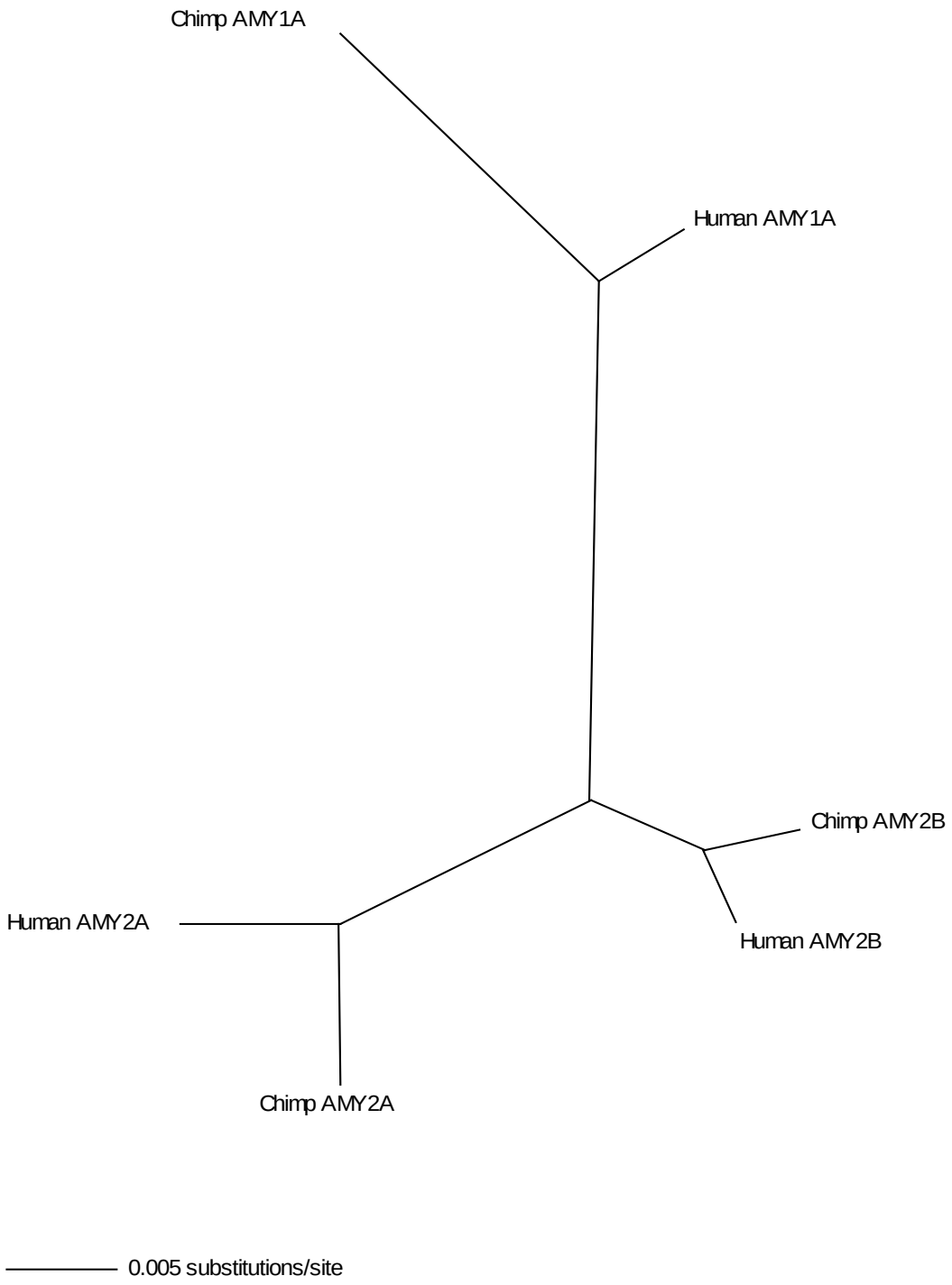



Fig 6.3a: An unrooted neighbour joining tree of chimpanzee and human amylase gene sequences using the Kimura (1980) two parameter model. Neighbour joining is a clustering method that attempts to find the smallest sum of branch lengths for trees based on a distance matrix (Saitou & Nei 1987). Phylogenetic analysis was conducted using PAUP*V4,β10 (Sinauer Associates, Sunderland, Mass).

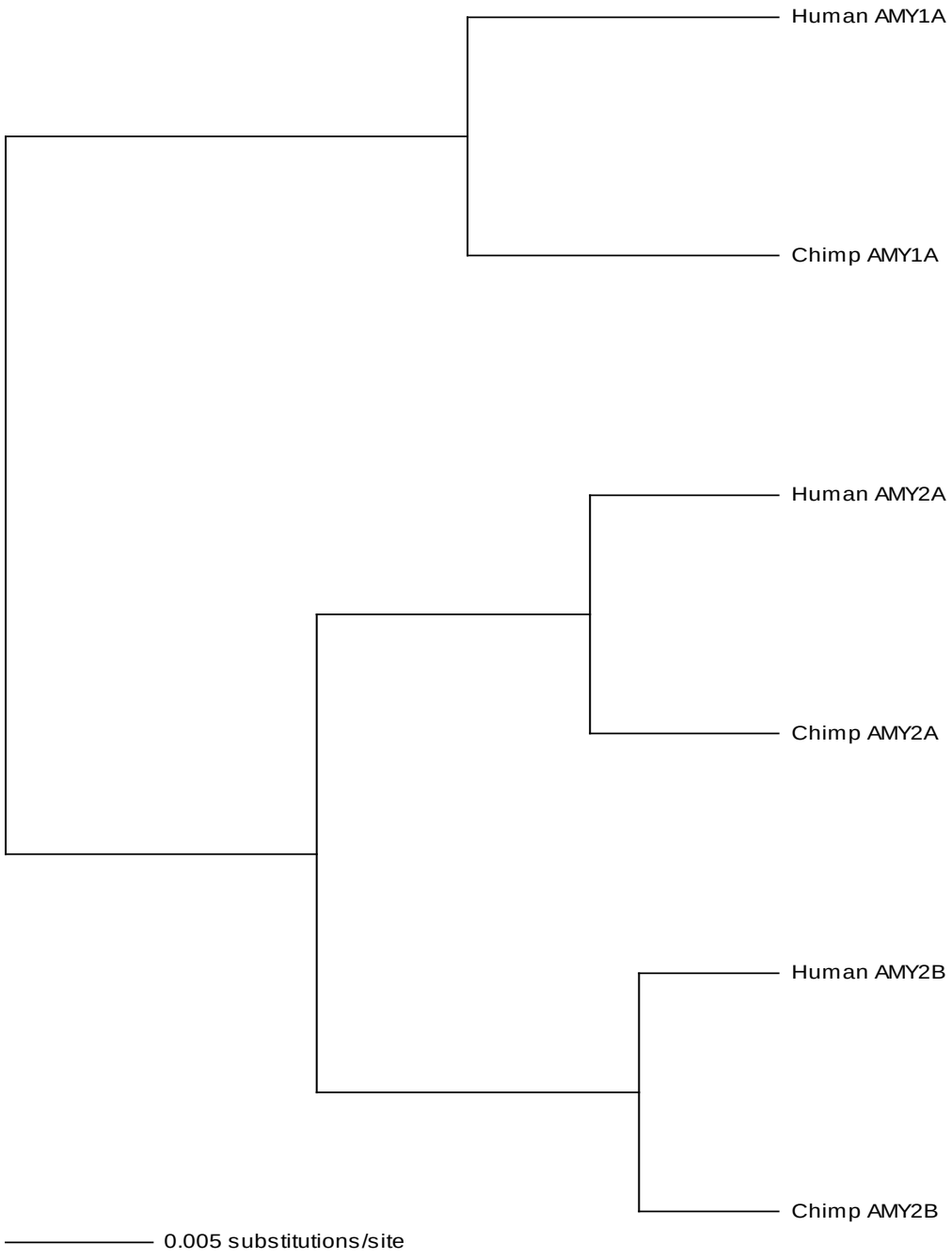



Fig 6.3b: Maximum likelihood tree of chimpanzee and human amylase gene sequences using the HKY85 model of evolution with gamma (ie differences in the rate of evolution between sites). The molecular clock was enforced so that a root could be obtained by midpoint rooting. Phylogenetic analysis was conducted using PAUP*V4,β10 (Sinauer Associates, Sunderland, Mass).

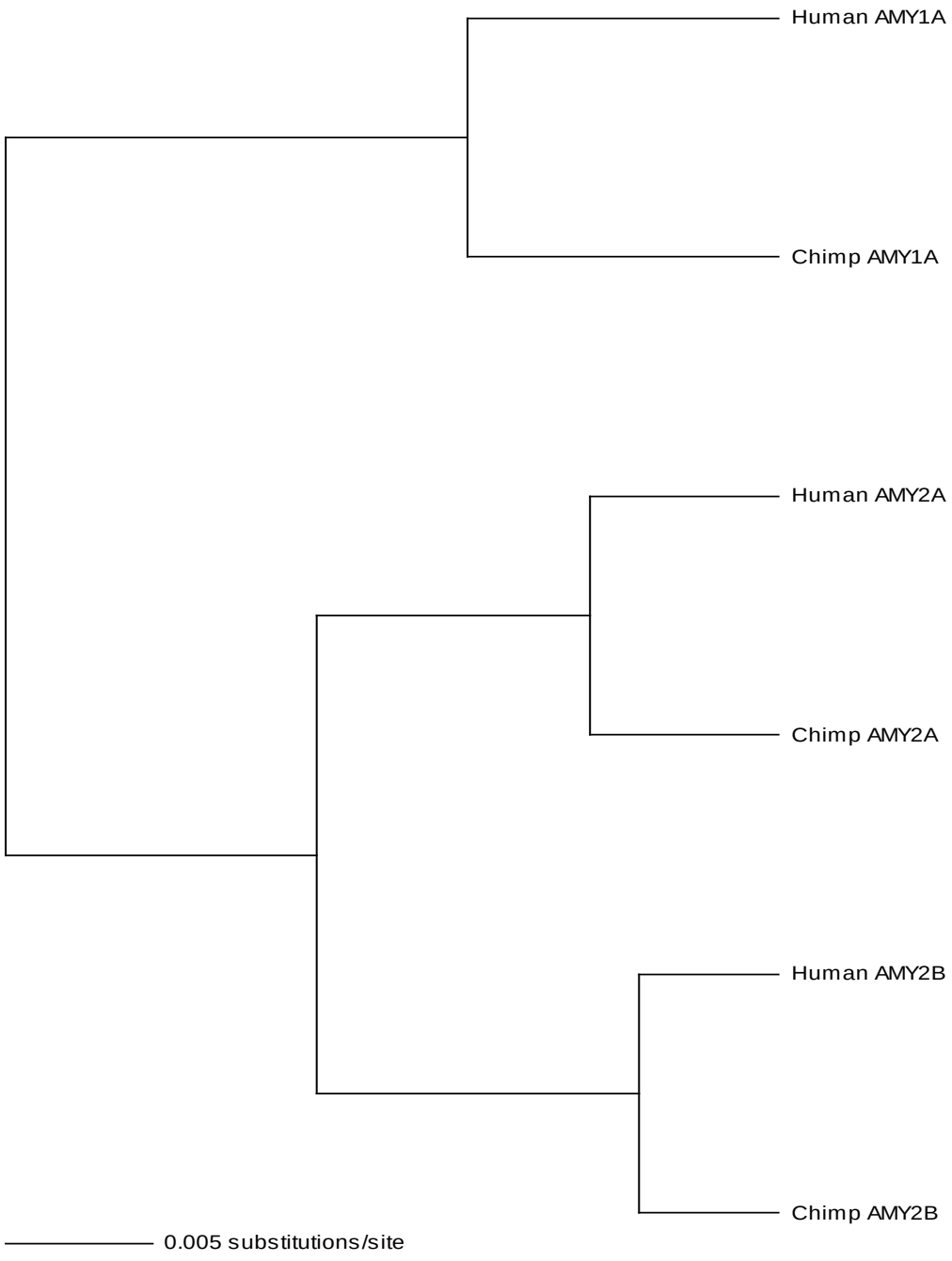
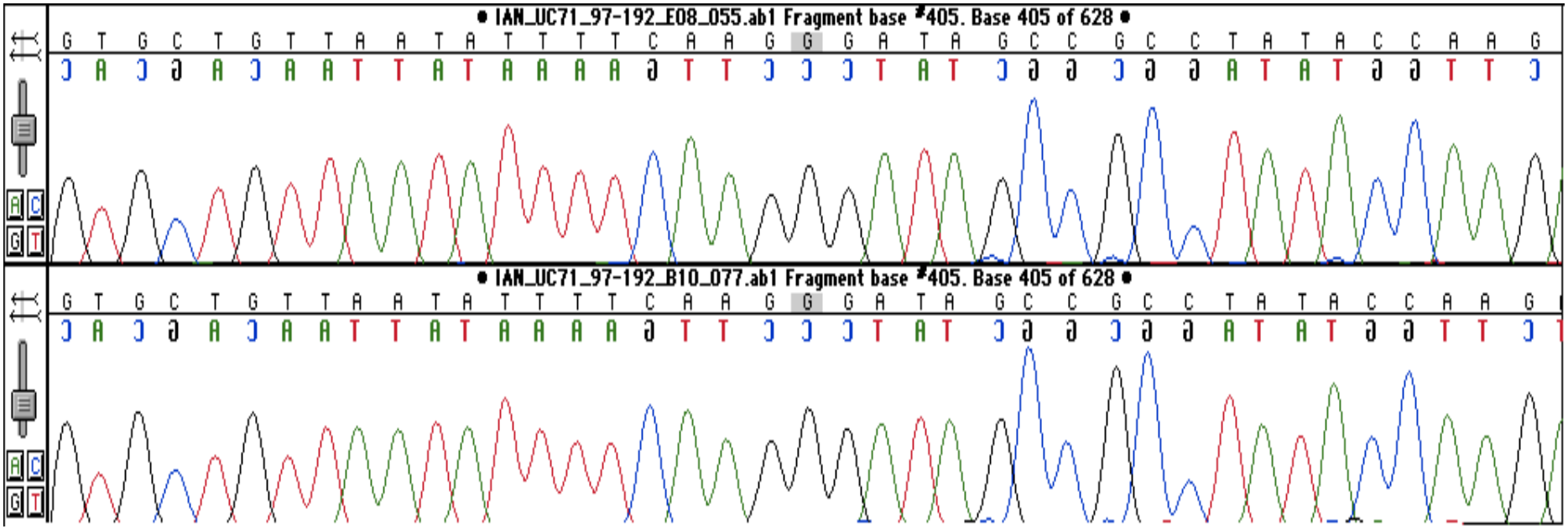



Fig 6.3b: Maximum likelihood tree of chimpanzee and human amylase gene sequences using the HKY85 model of evolution with gamma (ie differences in the rate of evolution between sites). The molecular clock was enforced so that a root could be obtained by midpoint rooting. Phylogenetic analysis was conducted using PAUP*V4,β10 (Sinauer Associates, Sunderland, Mass).




Chimp AMY1A

G T G C T G T T A A T A T T T T C A A G G G A T A G C C G C C T A T A C C A A G

Human AMY1A

G T G C T G T T A A T A T T T T C A A G A G A T A G C T G C C T A T A C C A A G

Fig 6.4: Two sequencing chromatograms from cloned chimpanzee PCR products originating from the AMY1A gene. Shown below is the consensus sequence for the chimpanzee AMY1A sequences, as well as the human AMY1A sequence. The region of the QAMY02-U primer is shown in green. Differences between the chimpanzee AMY1A consensus sequence and the human AMY1A sequence are shown in red.

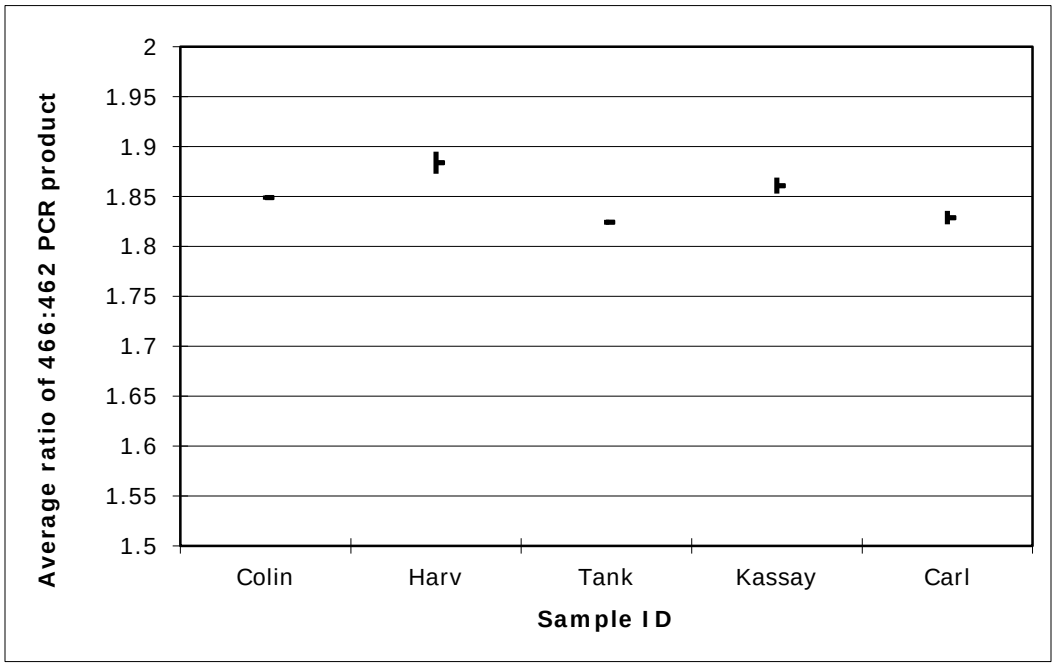
primer would not anneal as efficiently to the AMY1A genes, causing a decrease in the amplification efficiency of AMY1A specific products. The resulting decrease in AMY1A products could explain the ratios that were obtained using the human QAMY02 primers in chimpanzees. The sequence data from the chimpanzee amylase genes was used to design a new upper primer for a chimpanzee QAMY02 assay. A suitable primer was chosen to optimise for compatibility of annealing temperature with the existing QAMY02-L-TET primer as well as elimination of false priming sites, high 5' stability and low 3' stability (See Section 3.4.2 for more details of primer design methodology). Although there seem to be a number of areas on the human/chimp amylase alignment with 100% identity (see Fig 6.2) suitable for the placement of the primer that also would have produced a shorter PCR product, the primer design criteria restricted the placing of the primer to its current location.

Once the chimpanzee QAMY02 assay had been developed, the five chimpanzee DNA samples were typed for AMY1 gene copy number. Of the five chimpanzees typed, all five were found to be the same genotype: AMY1*H0/H0 (see fig 6.5, Table 6.2).

Table 6.2: Ratios of peaks of fluorescence for AMY1+AMY2A (466bp) and AMY2B (422bp) PCR products from chimpanzee QAMY protocol for five chimpanzee DNA samples.

| Chimp ID | PCR A | | PCR B | | Mean | Variance |
|---|---|---|---|---|---|---|
| | RUN 1 | RUN 2 | RUN 1 | RUN 2 | | |
| Colin | 1.8326 | 1.9178 | 1.8337 | 1.8114 | 1.8489 | 0.0022 |
| Harv | 1.7859 | 2.0314 | 1.8733 | 1.8448 | 1.8838 | 0.0110 |
| Tank | 1.7980 | 1.8483 | 1.8802 | 1.7703 | 1.8241 | 0.0024 |
| Kassay | 1.9811 | 1.8702 | 1.7719 | 1.8202 | 1.8608 | 0.0080 |
| Carl | 1.8065 | 1.8568 | 1.9230 | 1.7289 | 1.8288 | 0.0067 |

Fig 6.5: A graph to show mean values for the ratio of 466:462bp PCR product for 5 chimpanzees for 2 PCRs each run twice on an ABI 377/GeneScan™ system. The variance of the runs is shown by the error bars. The predicted ratio for genotype AMY1*H0/H0 is 2.0. All the results are within the acceptable range (+/- 0.2) of this value and so all samples were assigned the genotype AMY1*H0/H0.



### 6.4 Discussion

Despite obtaining the same AMY1 genotype of AMY1*H0/H0 for all of the chimpanzee samples studied, it cannot be ruled out that variation in AMY1 gene copy number exists in chimpanzees. With a sample of 5 chimpanzees (10 chromosomes) the allele frequency for AMY1*H0 could be as low as 74% and still have a 5% chance of generating the observed data (see below). A much larger number of chimpanzee DNA samples would be needed to rule out the existence of variation in AMY1 gene copy number in chimpanzees. To be 95% confident that the AMY1*H0 allele is at a frequency of at least 95% in the chimpanzee population, a further 48 chromosomes would need to be sampled. However a frequency of 74% for AMY1*H0 in chimpanzees still indicates that AMY1*H0 is the modal AMY1 repeat allele in chimpanzees.

P(observe 10 chromosomes same | Frequency = x) = 0.05 = $x^{10}$

Therefore x = $0.05^{(1/10)}$ = 0.7411

Solve for n and fix x at 0.95:

$$n = \frac{\log(0.05)}{\log(0.95)} = 58.404$$

This contrasts with humans, where the modal AMY1 repeat allele is AMY1*H1 and is present at frequencies estimated to be between 42% and 68% (See Chapter 4, Table 4.2). It is often assumed that the modal haplotype in humans is the ancestral form, and all other haplotypes are derived (see Thomas et al. 2002, and Stumpf and Goldstein 2001 for a review). As only AMY1*H0 has been found to date in chimpanzees, it is likely that the duplication of the AMY1 locus found throughout human populations has occurred since the human/chimpanzee lineages diverged approx 4-6mya. During this time, there must have been a number of deletion, inversion and duplication events to create the various AMY1 alleles found in humans. Once the AMY1*H1 allele has been formed is it relatively easy to envisage how the other AMY1 repeat alleles were formed through a series of unequal homologous crossovers (See fig 1.10).

It is interesting to note that, in comparison to human diets, there is a notable lack of high starch content foods, such as grains, in the diets of chimpanzees. The presence of AMY1*H0 as the modal AMY1 repeat allele is therefore consistent with the prediction from the dietary evidence. A homozygote for the AMY1*H0 allele would have sufficient salivary amylase for a low starch diet such as that found in chimpanzees.

Interestingly Neighbour-Joining tree shows the chimpanzee AMY1A sequence as having a longer branch length than the human AMY1A from their common node (see Fig 6.3). It is not immediately obvious why this should be the case. One explanation, however, is that human AMY1 sequences are constrained by gene conversion, between the multiple copies of AMY1 that are present in the majority of human individuals. Concerted evolution, through the process of gene conversion between paralogous genes can act to maintain sequence homology. An example of this comes from the δ and β-globin genes, which have extended sequence similarity despite their having diverged through gene duplication 85-100 million years ago (see Papadakis & Patrinos 1999). Innan (2002) developed a method for estimating gene conversion rates in multigene families and estimated it to be approximately 60-165 times higher than the mutation rate for synonomous sites.

The question of whether variation in salivary amylase gene copy number is present in chimpanzees requires further work to increase the number of chimpanzee chromosomes in the study. An additional avenue for further research would be to ascertain the extent of variation in AMY1 gene copy number in other primate taxa. Both these approaches would increase our understanding of the evolution of the amylase gene cluster in primates.

# Chapter 7: Diet and the allele frequencies at the Alanine:Glyoxylate Aminotransferase Pro11Leu locus in different human populations

**Note:**

The work described in this chapter was done in direct collaboration with Prof Christopher Danpure and an undergraduate student Ms Lianne Mayor who carried out the initial practical work.

## 7.1 Introduction

Alanine:Glyoxylate aminotransferase (AGT) is an enzyme found in the liver that catalyses the conversion of glyoxylate to glycine. Glyoxylate is a precursor for oxalate, which, when present in large amounts, can lead to the formation and excretion of multiple calcium oxalate (CaOx) kidney stones, and can eventually lead to renal failure (Danpure & Purdue 1995). Most oxalate and oxalate precursors in the diets of humans and other animals comes from the consumption of plants. Oxalate is a metabolic end-product and has no known biological role in animals, but is involved in many important metabolic processes in plants such as calcium homeostasis, structural support and defence.

The activity and subcellular distribution of AGT varies widely in different mammals, and both are significantly correlated with diet (Danpure 1997, Danpure et al. 1994). In carnivores, AGT tends to have a high activity and is localised within the mitochondria. However in herbivores AGT tends to have a lower activity and be localised within the peroxisomes. In omnivores AGT is usually found at intermediate levels and to be located within both the mitochondria and peroxisomes. This dual organellar localisation of AGT is thought to reflect its dual metabolic roles of glyoxylate detoxification in the peroxisomes and gluconeogenesis in the mitochondria (Danpure et al. 1990, 1994). In addition, for glyoxylate detoxification to be efficient AGT must be concentrated at the site of glyoxylate synthesis. This is likely to be different in carnivores and herbivores because the main dietary precursor of glyoxylate in herbivores is thought to be glycolate,

which is metabolised to glyoxylate in the peroxisomes (Noguchi 1987). In contrast, the precursor of glyoxylate in carnivores is more likely to be hydroxyproline, which is converted to glyoxylate in the mitochondria (Takayama et al. 2003).

The importance of correct subcellular distribution of AGT has been demonstrated by Danpure et al. (1989) in a study of two patients with the potentially lethal human hereditary kidney stone disease primary hyperoxaluria type 1 (PH1). Although in most normal humans AGT is peroxisomal, in many PH1 patients AGT is mis-targeted to the mitochondria. Mis-targeted AGT remains catalytically active but is unable to perform glyoxylate detoxification efficiently. As a result, oxalate synthesis increases and calcium oxalate crystallises out as stones in the kidney and urinary tract.

The evolution of AGT has been studied two groups of mammals: the *Anthropoidea* suborder of primates (Holbrook et al. 2000) and the *Carnivora* (Birdsey et al. 2004). Evidence for strong positive selection to decrease the efficiency of mitochondrial AGT targeting has been found in several anthropoid lineages as well as in the giant panda, possibly as an adaptation to increased herbivory. It has been estimated by Danpure et al. (1994) that the subcellular distribution of AGT must have changed on at least eight occasions during the evolution of mammals.

Although there is a significant correlation between the subcellular distribution of AGT and diet (Danpure et al. 1994, Danpure 1997) humans present an unusual situation. Humans are considered to be omnivores but, like the other species of the hominoidea that have been studied, they have a herbivorous peroxisomal distribution of AGT (Holbrook et al. 2000). In humans, however, a common polymorphism has been found which has important consequences for AGT activity and subcellular distribution. The ‘minor’ allele differs from the ‘major’ allele in that a C has been replaced with a T resulting in a Pro11Leu amino acid substitution. The minor allele reduces the activity of the AGT enzyme to a third of the normal level. In addition, individuals who are homozygous for the minor allele, instead of targeting 100% of their AGT to peroxisomes, target 90-95%

to peroxisomes and 5-10% to the mitochondria (Purdue et al. 1990). This 'minor' allele has been found at an allele frequency of 20% in Caucasians.

It is not unreasonable to suggest, therefore, that the minor allele at the AGXT Pro11Leu locus in humans, may represent an adaptation to a more omnivorous diet from a more herbivorous ancestral diet. The AGXT Pro11Leu locus thus provided another example to investigate for molecular evidence for dietary adaptation in humans.

The aim of this study was to determine the allele frequencies at the AGXT Pro11Leu locus in a range of human populations and to test whether the frequency differences found departed from the SNP allele frequency differences between populations that are found elsewhere in the genome. The AGXT Pro11Leu allele frequency data was compared to a null distribution of $F_{ST}$s for 11,024 SNPs, using a similar method previously described in Chapter 4 for the AMY1 locus.

## 7.2 Methods

### *7.2.1 Nomenclature and abbreviations*

AGT denotes the alanine:glyoxylate aminotranferase enzyme. In humans, AGXT denotes the gene that codes for the AGT enzyme.

### *7.2.2 Collection of samples*

DNA samples were typed for the Pro11Leu polymorphism in 83 Mongolians, 76 Norwegians (Weale et al. 2002), 82 North Welsh (Weale et al., 2002), 73 Armenians (Weale et al., 2001), 62 Nigerians from the Cross-River region, 69 Ethiopians, 88 Anatolian Turks, 34 Swedish Saami, 86 Sichuan Chinese and 85 Indians from Mumbai. Informed consent was obtained from all donors. All donors provided details of self-defined ethnic identity, first and second language and place of birth with similar information on his mother, father, maternal grandmother and paternal grandfather.

### *7.2.3 Polymorphism Detection*

The Pro11Leu c32C>T polymorphism (previously described as $C_{154}T$) was typed by PCR-RFLP as follows: PCR was carried out in a total reaction volume of 10 µl containing 0.3 µM of primers MIT 2 (GCA CAG ATA AGC TTC AGG GA) and EX-2R (CTT GAA GGA TGG ATC CAG GG), 200 µM dNTPs, 10mM Tris-HCl (pH 9.0), 0.1% Triton X-100, 0.01% gelatin, 50 mM KCl, 2.2 mM $MgCl_2$, 0.13 units Taq polymerase (HT Biotech, Cambridge, UK), 9.3 nM TaqStart monoclonal antibody (Mab) (BD Biosciences/Clontech San Jose, CA) and 1 µl of DNA. The Taq and TaqStart Mab were premixed prior to being added to the other reagents. Cycling parameters were pre-incubation at 95°C for 5 min followed by 37 cycles of 94°C for 1 min, 58°C for 1 min, 72°C for 1 min; and then a final incubation step of 72°C for 10 min.

Restriction endonuclease digestion was carried out overnight at 37°C in a final volume of 15 µl, and contained 10 µl of PCR product, 10 units of *Sty*1, 0.01 □g/□ acetylate and NEB Buffer 3 as recommended by the manufacturer. Digestion products were then run on 2 % agarose gel and DNA bands were visualised using ethidium bromide staining. Bands on the gel corresponding to 512bp and 619bp were interpreted as the major *AGXT* allele and minor *AGXT* allele respectively.

To ensure data quality and the accurate sizing of digestion products, some samples were also PCR amplified and *Sty*1 digested as above but using a version of the EX-2R primer that had been fluorescently labelled with the dye NED. These digestion products were then run on an ABI-377 automated sequencer (5% acrylamide denaturing gel for 3.5h) along with a ROX-labelled size standard.

**7.3 Results**

The frequencies of the AGXT Pro11Leu alleles in different human populations are shown in Table 7.1. There is a marked South East – North West

| | MON | NOR | NW | ARM | NIG | ETH | AT | SA | ASH | CHI | HIN |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AA | 69 | 50 | 60 | 48 | 52 | 55 | 68 | 19 | 47 | 82 | 79 |
| Aa | 11 | 22 | 20 | 22 | 9 | 13 | 14 | 11 | 23 | 4 | 5 |
| aa | 0 | 4 | 2 | 3 | 1 | 1 | 6 | 4 | 3 | 0 | 0 |
| n | 82 | 76 | 82 | 73 | 62 | 69 | 88 | 34 | 73 | 86 | 85 |
| Frequency of AGXT minor allele | 0.0688 | 0.1973 | 0.1463 | 0.1918 | 0.0887 | 0.1087 | 0.1477 | 0.2794 | 0.1986 | 0.0233 | 0.0298 |

Table 7.1: Frequency of AGXT genotypes in 11 human populations. Allele “A” corresponds to AGXT major allele, allele a corresponds to AGXT minor allele (see Introduction). Population codes: MON, Mongolia; NOR, Norway; NW, North Wales; ARM, Armenia; NIG, Nigeria; ETH, Ethiopia; AT, Anatolian Turks; SA, Saami, ASH, Ashkenzai Jews; CHI, Chinese; HIN, Indians Hindus.

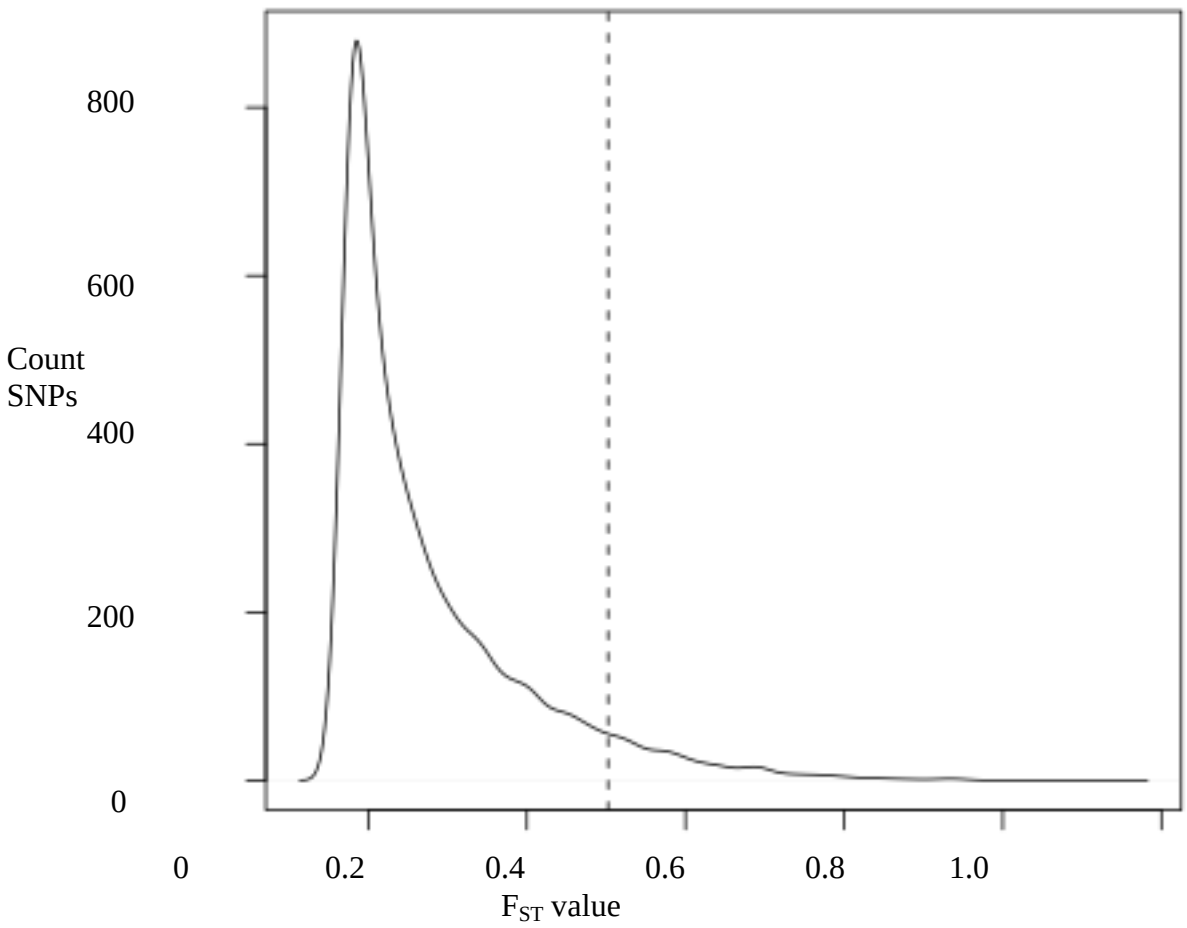
cline in the frequency of the minor allele. It is low (2.3% - 6.9%) in eastern and southern Asia, intermediate in Africa (8.9% - 10.9%), and high in Europe and the Middle East (14.6% - 27.9%). It should be noted that of the populations studied, the one predicted to have the most meat-based ancestral and current diet (the Saami) has the highest estimated frequency of the Pro11Leu minor allele (27.9%) (See Table 7.2).

Table 7.2: Nutrient composition and variation with latitude for nutrients for pre-agricultural diets (after Boyd Eaton & Boyd Eaton 2000)

| Nutrient | Typical contribution to diet | Variation with Latitude |
|---|---|---|
| Protein - animal | Very High | Positive |
| Protein - vegetable | Moderate | Negative |
| Total Fat | Moderate to high | Positive |
| Carbohydrate – from vegetables and fruit | Very high | Negative |
| Fibre | Very high | Negative |
| Micronutrients | Very high | Negative |

The Saami are the only population in the sample for which there is good evidence for a long history of a high meat diet (Haglin 1991, 1999). To examine whether the observed differences in AGXT Pro11Leu minor allele frequency between the Saami and non-European populations are within the range expected for neutral alleles, frequency differences were quantified using the genetic distance measure $F_{ST}$ (Weir, 1996), and compared to null-distributions of $F_{ST}$ for comparable populations (see methods). Currently, large data sets of SNPs, from which null distributions can be constructed, are only available for Europeans, East Asians and African Americans (Sachidanandam et al. 2001). Although the populations examined in this study are not identical to those for which large data sets of SNPs are available, it can be argued that a comparison of the Saami / Chinese Pro11Leu minor allele $F_{ST}$ against the null-distribution of $F_{ST}$s for the European / East Asian data sets is a conservative one. This is because previous studies of a large number of classical polymorphic markers (Cavalli-Sforza et al. 1988; Cavalli-Sforza et al. 1994) have shown that the

Fig 7.1 $F_{ST}$ values between Europeans and East Asians for 11,024 SNPs spread throughout the human genome. The dotted line represent the FST value for AGXT Pro11Leu locus, for a Saami vs Chinese comparison. The SNP dataset was taken from a dataset of 33,487 SNPs typed by the Orchid Laboratory in 42 African Americans, 42 East Asians and 42 European Americans publicly available at the SNP Consortium web site (URL: http://snp.cshl.org/allele_frequency_project/panels.shtml) (Sachidanandam et al. 2001). A total of 11,024 SNPs were carefully selected from the dataset according to the criteria detailed in Chapter 4 (Section 4.3.4). All $F_{ST}$ values were calculated using the unbiased 'random populations' formula for haploid data given by Weir (1996). All statistical analyses were carried out using the statistics package 'R' (URL: http://www.R-project.org/).



$F_{ST}$s between Saami and East Asian populations are typically lower than those between continental European and East Asian populations. However, a comparison of the Saami / Nigerian Pro11Leu minor allele $F_{ST}$ against the null-distribution of $F_{ST}$s for the European / African American data sets is likely to be biased in favour of $F_{ST}$-outlier status for the Pro11Leu minor allele. This is because a number of studies have shown African Americans to be an admixed group between Europeans and West Africans, with the greater ancestry proportion (between 80-90%) being West African in origin (Parra et al. 1998). As a consequence, the European / African American comparison is likely to produce an underestimate of the true null-distribution of $F_{ST}$ between the Saami and Nigerian populations.

When compared to the null distribution for Europeans and East Asians the $F_{ST}$ for the Saami vs Chinese ($F_{ST}$ = 0.3024) was in the top 6.96% of the distribution (see Fig 7.2). The $F_{ST}$ for the Saami vs Nigerians was 0.1184, which was in the top 29% of the European / African American $F_{ST}$ distribution.

### 7.4 Discussion

The $F_{ST}$ value for the Saami vs Chinese comparison was an outlier (in the top 6.96%) on the null distribution for $F_{ST}$s for Europeans and East Asians. This indicates that there is a greater difference in allele frequency between the Saami and Chinese than is found for the majority of the SNPs that make up the distribution. This difference could be the result of local selection pressures increasing the Pro11Leu minor allele frequency in one population, but not the other. As noted in the results, the Saami are predicted to have the highest proportion of meat in their diet, and have the highest proportion of Pro11Leu minor allele, which causes AGT to be targeted to the mitochondria.

These results are an intriguing preliminary exploration into the evolutionary history of the Pro11Leu polymorphism in humans. The unusually high $F_{ST}$ value for the Saami vs Chinese comparison warrants further investigation in order to establish whether signals of selection at the AGXT locus can be found. If such signatures were found then AGT

would provide a further example of human dietary adaptation. The non-human primates show a peroxisomal distribution of AGT, which has been suggested to be an adaptation to increased herbivory. However, these results suggest that humans are moving away from this herbivorous distribution of AGT to a more omnivorous distribution. Thus over the course of primate evolution, there has been a shift towards hervbivory, which shows signs of being reversed in humans.

As discussed previously in this thesis (see chapter 4), the comparison of $F_{ST}$s from the locus on interest with those from a neutral distribution is a somewhat insensitive indicator of selection. In order to demonstrate that selection has occurred at the AGXT locus in humans, additional data is needed. Currently haplotype based selection methods provide the most sensitive way to detect selection in the human genome (Sabeti et al. 2002). To apply these methods on the AGXT locus, data on extended SNP haplotypes or closely linked microsatellites would need to be collected. Unfortunately the collection of such data is both costly and labour intensive and was out of the scope of this project. It is hoped that in the future data of this kind will be collected to increase out knowledge of the evolution of the AGXT locus and shed light on dietary adaptation to omnivory in humans. A study of this kind would have major implications for the study of human evolution as there has been much discussion on the importance of meat eating in the evolution of the human brain and the development of the human ecological niche (see Aiello & Wheeler 1994).

# Chapter 8: General Discussion

The human diet is unusual among primates in both the large meat component and the prominence of foods derived from domesticated plant and animal species, such as milk and cereals. It has often been suggested that the human digestive system has adapted to cope with these new dietary components, on both a gross anatomical and molecular scale. However it is only recently has it become possible to adequately test hypothesis of molecular adaptation using genetic data. This is due to advances in both high throughput genotyping as well as analytical approaches to test for selection. A particularly good example of this is the lactase persistence phenotype, common in European, Middle Eastern and North African populations, which confers the ability to digest milk into adulthood (see Swallow 2003 for a review). For many years, population variation in lactose digestion capacity has been explained in terms of an adaptation to milk drinking as a result of the domestication of cattle (Simoons 1969, Roberts 1985, Stinson 1992). However it is only in the last year that population genetic evidence of natural selection at the lactase gene has been put forward (see Bersaglieri et al. 2004). Using both allele frequency comparisons (using $F_{ST}$) and haplotype based tests for selection, Bersaglieri and colleagues (2004) observed the strongest signals of selection yet seen for any gene in Europeans, as well as one of the strongest in all human populations.

The aim of this thesis was to explore the variation in two diet related enzymes, salivary amylase (AMY1) and alanine:glyoxylate aminotransferase (AGT), in order to assess the evidence for selection at these loci. For both genes there are plausible adaptive scenarios, in the time frame of the spread of modern humans out of Africa. In the case of salivary amylase, a high starch diet would provide a situation where having an increased number of fully functioning salivary amylase genes would be a distinct advantage. For AGT, the subcellular location of the AGT enzyme is known to be associated with the amount of meat and fish in the diet across a range of mammalian species (Danpure 1997, Danpure et al. 1994). In carnivores AGT is found in the mitochondria, in herbivores and is located in the peroxisomes with omnivores having AGT in both mitochondria and peroxisomes (Danpure et al. 1990, 1994). Over the course of our evolution, humans have changed

from being largely herbivorous to omnivores (see Milton 1999, 2000). It is tempting to speculate that the Pro11Leu minor allele, which causes AGT to be mis-targeted to the mitochondria, is an adaptation to a more omnivorous diet (see Caldwell et al. 2004).

All three enzymes, lactase, AGT and salivary amylase play an central role in the digestion and metabolism of important dietary components, and yet there seem to be major differences in the strength of selection that has been operating at these loci. Lactase has one of the strongest signals of selection ever reported, stronger even than many disease genes (Bersaglieri et al. 2004) and the strongest ever reported in Europeans. Preliminary data indicate that AGT has also been subject to forces other than genetic drift shaped by demography (Caldwell et al. in 2004). However the data presented in this thesis found no evidence for a fitness advantage for any of the AMY1 repeat alleles.

Both lactase and AGT have alleles that under the certain dietary conditions will produce disease states. If an individual is homozygous for the non-persistent lactase allele and consumes large amounts of fresh milk in adult life, abdominal cramps and diarrhoea occur and the nutritional benefits of lactose cannot be obtained. In the case of AGT, inappropriate subcellular targeting of the AGT enzyme can lead to the formation of calcium oxylate kidney stones and eventually renal failure. In contrast, salivary amylase does not have an allele that confers a great disadvantage in any particular dietary circumstances. Even a homozygote for the AMY1*H0 allele will still possess two copies of the AMY1 gene and so express amylase in the saliva. Furthermore, starch digestion is continued in the intestine by the action of the pancreatic amylases. Finally, salivary amylase is not anchored to a membrane as is the case with many digestive enzymes, but freely diffusible in saliva. As a result, differences in enzyme concentration would have a smaller effect. These factors may have contributed to the difference in the strength of the selective pressure acting on the AMY1 locus compared to the two other loci.

Salivary amylase still presents an interesting case, however. Only the AMY1*H0 allele was found in the five chimpanzees studied, where are various AMY1 repeat alleles responsible for variation in salivary amylase gene copy number have been found in all

human populations studied to date. And yet if there has been any recent positive selection at this locus the effect on allele frequency across human populations and intra-allelic variability is very small. One explanation for this is that the variation at the AMY1 locus is the result of a much older selection pressure, the signals of which have been obscured in the intervening years by admixture and drift shaped by demographic history.

Richard Wrangham and colleagues (1999, see also Pennisi 1999) have suggested that the cooking of tubers, such as potatoes, cassava, yams, manioc and turnips, could have been pivotal in the evolution of the genus *Homo*. They argue that it was tubers and the ability to cook them that prompted the evolution of large brains, smaller teeth, modern limb proportions, human life cycles and social structures, which stared approximately 1.8 million years ago. Although undisputed evidence for controlled fire comes from only 250,000 years ago, controversial claims for the use of fire have been made for sites dated to 1.4 million years ago (see James 1989 for a review).

Wrangham argues that starchy tubers would have been abundant on the plains of Africa two million years ago. A diet of 60% cooked tubers, approximately the proportion used in modern native African diets, and no meat boosts caloric intake by about 43% over that of humans who ate nuts, berries, and raw tubers (Pennisi 1999). It is tempting to speculate that the variation we see in AMY1 gene copy number in modern human populations is the result of an adaptation to an increase in the consumption of starchy tubers over a million years ago. However, it is not possible to adequately test this hypothesis using current population genetic methods.

An alternative explanation for the maintenance of the high levels of polymorphism at the AMY1 locus in humans is that the locus has been under balancing selection, although it is hard to see what advantage this might confer. With hindsight it would seem more likely that adaptive polymorphisms under positive selection would be found in genes where having a low concentration of the gene product jeopardises to the normal functioning of the organism.

**Suggestions for further work**

*AMY1*

The methods developed for typing the AMY1 repeat alleles as well six closely linked microsatellites could be applied to additional populations that were not available in this study. These include a number of hunter-gatherer populations, such as the Inuit, Australian Aborigines, the !Kung San, and new world populations such as the Yanomamo. As discussed in Chapter 4 this would serve to improve the agricultural vs hunter-gatherer comparisons.

It would also be interesting to extend the non-human primate study from Chapter 6. Firstly it would be beneficial to increase the number of chimpanzees samples typed for the AMY1 repeat alleles. In addition it would be interesting to adapt the assay for use in other non-human primate species, especially gorillas, orang-utans and gibbons. This information would help to clarify whether the variation in AMY1 gene copy number is indeed unique to humans.

*AGT*

The unusually high allele frequency differences between the Saami and Chinese at the AGT locus certainly warrants further investigation (See Chapter 7 and Caldwell et al. in 2004). If additional data such as microsatellite or extended SNP haplotypes were obtained then an analysis if intra-allelic variability could be carried out. This would provide a means of obtaining information about the intensity of the selection signal, as well as providing a means to date the most recent common ancestor for the alleles.

*Other dietary enzymes*

There are a number of other enzymes involved in the metabolism of dietary components that would be interesting to investigate for molecular evidence for dietary adaptation in humans. The theories and techniques used in this thesis could be applied to the genes involved to assess the evidence for selection at these loci.

Candidates for future investigation include:

*Fructose 1-phosphate-splitting liver adolase*:

Fructose is the sugar found in fruit and man-made fructose is used as a sweetener in many foods (including baby food) and drinks. Hereditary fructose intolerance is an autosomal recessive condition, which is characterised by recurrent vomiting and hypogylcemia at the time of weaning when fructose is added to the diet (Froesch 1963). The disease may be as common as 1 in 20,000 in some European countries. The aldolase B gene codes for the enzyme fructose 1-phosphate splitting liver adolase which catalyses the cleavage of fructose-1-phosphate, an intermediate in fructose metabolism, to form dihydroxyacetone phosphate and D-glyceraldehyde (Froesch 1966). Fructose intolerance would clearly be a disadvantage in a diet with a large proportion of fruit, such as the diet of chimpanzees and other primates. However, during human evolution a wide range of other foodstuffs with lower fructose content have been incorporated into the diet, such as meat and high starch foods. It is possible that dietary shifts to a diet with little or no fructose may have caused a relaxation of selective constraints on the defective aldolase B allele.

*Pepsinogen A (PGA):*

Pepsinogen is the inactive precursor of pepsin, the enzyme that breaks down proteins in the stomach to smaller chains of amino acids. Pepsinogen A is encoded by a multigene family and as with AMY1, variation in the number of pepsinogen genes exists between individuals (Bebelman et al. 1989). Frants et al. (1984) investigates the different pepsinogen isozymes and proposed that the relative intensities of the different fractions are determined by differences in gene copy number. Due to the parallels between the pepsinogen multigene family and the AMY1 genes, the principles of the AMY1 quantification assays developed in this thesis could be adapted for use with PGA. An interesting initial investigation would be to explore the extent of variation in PGA gene copy number in non-human primates. In addition to this, an investigation of the distribution of the PGA polygenic alleles in different human populations, combined with haplotype analysis, would provide information on the evolution of this complex gene cluster.

*Sucrase-isomaltase (SI):*

As outlined in chapter one, sucrose is likely to have formed a large part of the diet only in recent times. Deficiency of the enzyme responsible for sucrose digestion is not uncommon and reaches 16% in Inuit of Greenland (Mc Nair et al. 1972). To date, there have been no studies that have investigated the worldwide distribution of SI alleles. It would also be interesting to use haplotype based methods to investigate signatures of selection at the SI locus, as well as looking at the SI gene for evidence of relaxation of selective constraints.

## Bibliography


Aebi, H. E., J.P Von Wartburg, S.R. Wyss (1981). The role of enzyme polymorphisms and enzyme variants in the evolutionary process. In N. Kretchmer, and Dwain N. Walcher (eds). Food, nutrition and evolution. New York. Masson

Aiello, L. C., & P. Wheeler (1995). The expensive tissue hypothesis: brains and the digestive system in human and primate evolution. Current Anthropology 36: 199-221.

Akey, J. M., G. Zhang, et al. (2002). Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12(12): 1805-14.

Ammerman, A. J., & L.L. Cavalli-Sforza (1973). A population model for the diffusion of early farming in Europe. The explanations of cultural change. London, Duckworth: 343-357.

Andolfatto, P., Wall JD, Kreitman M (1999). Unusual haplotype structure at the proximal breakpoint on In(2L)t in a natural population of Drosophila melangaster. Genetics 153: 1297-1311.

Arezi, B., Xing, W., Sorge, J.A., Hogrefe, H.H. (2003). Amplification efficiency of thermostable DNA polymerases. Analytical Biochemistry 321: 226-235.

Bamshad, M., Stephan P Wooding (2003). Signatures of natural selection in the human genome. Nature Reviews Genetics 4(Feb 2003): 99-111.

Bank, R. A., E. H. Hettema, et al. (1992). Variation in gene copy number and polymorphism of the human salivary amylase isoenzyme system in Caucasians. Hum Genet 89(2): 213-22.

Bar-Yosef, O., & M.E. Kislev (1989). Early Farming communities in the Jordan Valley. Foraging and farming : the evolution of plant exploitation. D. R. G. C. H. Harris. London, Unwin Hyman: 632-642.

Basabose, A. K. (2002). Diet composition of chimpanzees inhabiting the montane forest of Kahuzi, Democratic republic of Congo. Am J Primatol 58(1): 1-21.

Bebelman, J. P., M. P. Evers, et al. (1989). Family and population studies on the human pepsinogen A multigene family. Hum Genet 82(2): 142-6.

Bersaglieri, T., P. C. Sabeti, et al. (2004). Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74(6): 1111-20.

Bilmes, J. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute. Berkeley, CA.

Binford, L. R. (1985). Human ancestors: changing views of their behaviour. Journal of Anthropological Archaeology 4: 292-327.

Birdsey, G. M., Lewin, J., Cunningham, A.A., Bruford, M.W., Danpure, C.J. (2004). Differential Enzyme Targeting as an evolutionary adaptation to herbivory in Carnivora. Mol Biol Evol 21(3).

Boesch, C., Boesch H. (1983). Optimisation of nut-cracking with natural hammers by wild chimpanzees. Behaviour 83: 265-286.

Boesch, C., Boesch H. (1989). Hunting behaviour of wild chimpanzees in the Tai National Park. American Journal of Physical Anthropology 78: 547-573.

Boesch, C., Boesch H. (1993). Diversity of tool-use and tool-making in wild chimpanzees. The use of tools by human and non-human primates. A. Berthelet, Chavaillon J. Oxford, Clarendon: 158-174.

Bowcock, A. M., A. Ruiz-Linares, et al. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368(6470): 455-7.

Boyd Eaton, S., Boyd Eaton, S. (2000). Palaeolithic vs. modern diets - selected pathophysiological implications. Eur J Nutr 39: 67-70.

Boyd, R., & Joan B. Silk (1997). How humans evolved. New York, W.W. Norton.

Brothwell, D. P. (1969). Food in Antiquity: A survey of the diet of early peoples. London, Thames & Hudson.

Calabrese, P. P., Durrett, R.T., Aquadro, C.F. (2001). Dynamics of microsatellite divergence under stepwise mutation and proportional slippage/point mutation models. Genetics 159: 839-859.

Caldwell, E.F., Mayor, L., Thomas, M.G., Danpure, C.J. (2004) Diet and the Frequency of the Alanine:Glyoxylate Aminotransferase Pro11Leu Polymorphism in Different Human Populations. Human Genetics

Camps, G. (1975). The prehistoric cultures of North Africa: Radiocarbon chronology. Problems in prehistory: North Africa and the Levant. F. Z. Wendorf, Marks, A.E. Dallas, SMU Press: 181-192.

Capelli, C., N. Redhead, et al. (2003). A Y chromosome census of the British Isles. Curr Biol 13(11): 979-84.

Cassidy, C. M. (1980). Nutrition and Health in agriculturalists and hunter-gatherers. Nutritional anthropology : contemporary approaches to diet & culture. R. F. K. a. G. H. P. Norge W. Jerome. New York, Regrave: 117-145.

Cavalli-Sforza, L. L. (1966). Population structure and human evolution. Proc R Soc Lond B Biol Sci 164(995): 362-379.

Cavalli-Sforza, L. L. (1973). Analytic review: some current problems of population genetics. Am J Hum Genet 25: 82-104.

Cavalli-Sforza, L. L. (1981). Human evolution and nutrition. Food, nutrition and evolution. N. a. D. N. W. Kretchmer.

Cavalli-Sforza, L. L., Piazza A., Menozzi P., Mountain J., (1988). Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. Proc Natl Acad Sci USA 85: 6002-6006.

Cavalli-Sforza, L. L., Paolo Menozzi, Alberto Piazza. (1994). The history and geography of human genes. Princetown, Princetown University Press.

Chase, P. G. (1989). How different was Middle Palaeolithic subsistence? A zooarchaeological perspective on the Middle to Upper Palaeolithic transition. The Human Revolution. P. Mellars, Stinger, C. Edinburgh.

Chen, F. C. and W. H. Li (2001). Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet 68(2): 444-56.

Cordain, L. (2001). The Paleo Diet: Lose weight and get healthy by eating the food you were designed to eat, John Willey and sons.

Cordain, L., J. B. Miller, et al. (2000)*a*. Macronutrient estimations in hunter-gatherer diets. Am J Clin Nutr 72(6): 1589-92.

Cordain, L., J. B. Miller, et al. (2000)*b*. Plant-animal subsistence ratios and macronutrient energy estimations in worldwide hunter-gatherer diets. Am J Clin Nutr 71(3): 682-92.

Crotty, P. L., Staggs, R.A., Porter P.T., Killeen A.A., McGlennen R.C. (1994). Quantitative analysis in molecular diagnostics. Hum Pathol 25(6): 572-579.

Curtin, P., Feierman, S., Thompson, L., Vansina, J. (1995). African history from earliest times to independence. London, Longman.

Danpure, C. J., Cooper, P.J., Wise, P.J., Jennings, P.R. (1989). An enzyme trafficking defect in two patients with primary hyperoxaluria type 1 : peroxisomal alanine:glyoxylate aminotransferase rerouted to mitochondria. J Cell Biol. 108: 1345-1352.

Danpure, C. J., Guttridge, K.M., Fryer, P., Jennings, P.R., Allsop, J., Purdue, P.E. (1990). Subcellular distribution of hepatic alanine:glyoxylate aminotransferase in various mammalian species. J Cell Sci. 97: 669-678.

Danpure, C. J., Fryer, P., Jennings, J., Allsop, S., Griffiths, S. and Cunningham, A. (1994). Evolution of alanine:glyoxylate aminotransferase 1 peroxisomal and mitochondrial targeting. A survey of its subcellular distribution in the livers of various representatives of the classes Mammalia, Aves and Amphibia. Eur J Cell Biol 64: 295-313.

Danpure, C. J., Purdue, P.E. (1995). Primary hyperoxaluria. The metabolic and molecular bases of inherited disease. C. R. Scrivver, Beaudet, A.L., Sly, W.S., Valle, D. New York, McGraw-Hill.

Danpure, C. J. (1997). Variable peroxisomal and mitochondria targeting of alanine: glyoxylate aminotransferase in mammalian evolution and disease. Bioessays 19: 317-326.

Darwin, C. (1859). The origin of species. London, Dent.

de Soyza, K. (1978). Polymorphism of human salivary amylase: a preliminary communication. Hum Genet 45(2): 189-92.

Di Rienzo, A. (1998). Studies of populations and genetic diseases: mixing it up. Inherited disorders and their genes in different European populations, Acquafredda di Maratea, Italy, 6-11 February 1998. Trends Genet 14(6): 218-9.

Diamond, J. (2003). The double puzzle of diabetes. Nature 423(6940): 599-602.

Dracopoli, N. C. and M. H. Meisler (1990). Mapping the human amylase gene cluster on the proximal short arm of chromosome 1 using a highly informative (CA)n repeat. Genomics 7(1): 97-102.

Dumayne-Peaty, L. (2001) Human impact on vegetation. Handbook of Archaeological Sciences D. Brothwell &. A. Pollard (eds) Chichester, Wiley: 379-392.

Eaton, S. B. and S. B. Eaton, 3rd (2000). Palaeolithic vs. modern diets--selected pathophysiological implications. Eur J Nutr 39(2): 67-70.

Eaton SB, M Shostak, and M Konner (1988): The Paleolithic Prescription: A Program of Diet and Exercise and a Design for Living. N.Y., Harper & Row, Publishers, pp. 79.

Eaton, S. B., S. B. Eaton, 3rd, et al. (1997). Palaeolithic nutrition revisited: a twelve-year retrospective on its nature and implications. Eur J Clin Nutr 51(4): 207-16.

Eaton, S. B. and M. Konner (1985). Palaeolithic nutrition. A consideration of its nature and current implications. N Engl J Med 312(5): 283-9.

Ebersberger, I., D. Metzler, et al. (2002). Genomewide comparison of DNA sequences between humans and chimpanzees. Am J Hum Genet 70(6): 1490-7.

Edwards, A. W. F. (1992). Likelihood, John Hopkins Press.

Ehret, C. (2002). The civilisations of Africa: a history to 1800. Oxford, James Currey.

Emi, M., A. Horii, et al. (1988). Overlapping two genes in human DNA: a salivary amylase gene overlaps with a gamma-actin pseudogene that carries an integrated human endogenous retroviral DNA. Gene 62(2): 229-35.

Enard, W., P. Khaitovich, et al. (2002*a*). Intra- and interspecific variation in primate gene expression patterns. Science 296(5566): 340-3.

Enard, W., M. Przeworski, et al. (2002*b*). Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418(6900): 869-72.

Enattah, N. S., T. Sahi, et al. (2002). Identification of a variant associated with adult-type hypolactasia. Nat Genet 30(2): 233-7.

Excoffier, L., Smouse, P., Quattro, J. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics 131: 479-491.

Fage, J. D. (1988). A history of Africa. London, Hutchinson.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17: 368-376.

Fletcher, B., D. B. Goldstein, et al. (2003). High-throughput analysis of informative CYP2D6 compound haplotypes. Genomics 81(2): 166-74.

Forsyth, J. (1992). A history of the peoples of Siberia : Russia's north Asian colony, 1581-1990. Cambridge, Cambridge University Press.

Fortna, A., Kim, Y., MacLaren, E., Marshall, K., Hahn, G., Meltesen, L., Brenton, M., Hink, R., Burgers, S., Hernandez-Boussard, T., Karimpour-Fard, A., Glueck, D., McGavran, L., Berry, R., Pollack, J., Sikela, J.M. (2004). Lineage-specific gene duplication and loss in human and great ape evolution. PLOS Biology 2(7): 937-954.

Frants, R. R., J. C. Pronk, et al. (1984). Genetics of urinary pepsinogen: a new hypothesis. Hum Genet 65(4): 385-90.

Froesch, E. R., Wolf, H.P., Baitsch, H., Prader, A., Labhart, A. (1963). Hereditary fructose intolerance: an inborn defect of hepatic fructose-1-phosphate splitting aldolase. Am J Med 34: 151-167.

Froesch, E. R., Wolf, H.P., Baitsch, H., Prader, A., Labhart, A. (1966). Essential fructosuria and hereditary fructose intolerance. The metabolic basis of inherited disease. J. B. Stanbury, Wyngaarden, J.B., Fredrickson, D.S., McGraw-Hill.

Fu, X. Y., Li, W.H. (1993). Statistical tests of neutrality of mutations. Genetics 133: 693-709.

Fu, Y. (1996). New statistical tests of neutrality for DNA samples. Genetics 143: 557-70.

Fu, Y. (1997). Statistical test of neutrality of mutations against populations growth, hitchhiking and background selection. Genetics 146: 915-925.

Futuyma, D. J. (1998). Evolutionary Biology. SUNY, Sinauer Associates.

Gale, J. S. (1980). Population genetics, Blackie.

Goldstein, D. B., Linares, A.R., Cavalli-Sforza, L.L., Feldman, M.W. (1995). An evaluation of genetic distances for use with microsatellite loci. Genetics 139: 463-471.

Goldstein, D. B. and D. D. Pollock (1997). Launching microsatellites: a review of mutation processes and methods of phylogenetic interference. J Hered 88(5): 335-42.

Goldstein, D. B., D. E. Reich, et al. (1999). Age estimates of two common mutations causing factor XI deficiency: recent genetic drift is not necessary for elevated disease incidence among Ashkenazi Jews. Am J Hum Genet 64(4): 1071-5.

Goodall, J. (1986). The Chimpanzees of Gombe: Patterns of behaviour. Cambridge MA, Belknap.

Gordon, K. D. (1987). Evolutionary Perspectives on Human Diet. Nutritional Anthropology. F. E. Johnston. New York, Alan R Liss: 3-39.

Graham, D. N. (1998). Grain Damage: rethinking the high starch diet.

Groot, P. C. (1989). Structure and evolution of the human alpha-amylase multigene family. Genetics. Amsterdam, Vrije Universiteit te Amsterdam: 112.

Groot, P. C., M. J. Bleeker, et al. (1988). Human pancreatic amylase is encoded by two different genes. Nucleic Acids Res 16(10): 4724.

Groot, P. C., M. J. Bleeker, et al. (1989*a*). The human alpha-amylase multigene family consists of haplotypes with variable numbers of genes. Genomics 5(1): 29-42.

Groot, P. C., W. H. Mager, et al. (1991). Interpretation of polymorphic DNA patterns in the human alpha-amylase multigene family. Genomics 10(3): 779-85.

Groot, P. C., W. H. Mager, et al. (1989*b*). The human amylase-encoding genes amy2 and amy3 are identical to AMY2A and AMY2B. Gene 85(2): 567-8.

Groot, P. C., W. H. Mager, et al. (1990). Evolution of the human alpha-amylase multigene family through unequal, homologous, and inter- and intrachromosomal crossovers. Genomics 8(1): 97-105.

Gumucio, D. L., K. Wiebauer, et al. (1988). Concerted evolution of human amylase genes. Mol Cell Biol 8(3): 1197-205.

Haglin, L. (1991). Nutrient intake among Saami people today compared with an old, traditional Saami diet. Arctic Med Res(Suppl.): 741-746.

Haglin, L. (1999). The nutrient density of present-day and traditional diets and their health aspects: the Saami and lumberjack families living in rural areas of Northern Sweden. Int J Circumpolar Health 58(1): 30-43.

Haldane, J. B. S., 1949  Disease and evolution. Ric. Sci. Suppl. A **19**:68-76.

Hamblin, M. T. and A. Di Rienzo (2000). Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66(5): 1669-79.

Harlan, J. R. (1989). Wild-grass seed harvesting in the Sahara and Sub-Sahara of Africa. Foraging and farming : the evolution of plant exploitation. D. R. G. C. H. Harris. London, Unwin Hyman: 79-98.

Harris, D. (1996). The origins and spread of agriculture in Eurasia: an overview. The origins and spread of agriculture and pastoralism in Eurasia. D. Harris. London, UCL press.

Harris, D. R. (1981). The prehistory of human subsistence: A speculative outline. Food, Nutrition and Evolution. D. N. W. a. N. Kretchmer. New York, Masson: 15-35.

Harris, D. R. (1989). An evolutionary continuum of people-plant interaction. Foraging and farming : the evolution of plant exploitation. G. C. H. David R. Harris. London, Unwin Hyman.

Hartl, D. L. (1980). Principles of population genetics. Sunderland, Mass, Sinauer Associates.

Hellmann, I., S. Zollner, et al. (2003). Selection on human genes as revealed by comparisons to chimpanzee cDNA. Genome Res 13(5): 831-7.

Hill, E. W., M. A. Jobling, et al. (2000). Y-chromosome variation and Irish origins. Nature 404(6776): 351-2.

Hillman, G. C., S. M. Colledge & D. R. Harris (1989). Plant-food economy during the Epipalaeolithic period at Tell Abu Hureyra, Syria: dietary diversity, seasonality, and modes of exploitation. Foraging and farming : the evolution of plant exploitation. D. R. H. G. C. Hillman. London, Unwin Hyman: 240-268.

Hillman, G. C. (1989). Late Palaeolithic plant foods from Wadi Kubbaniya in Upper Egypt: dietary diversity, infant weaning, and seasonality in a riverine environment. Foraging and farming : the evolution of plant exploitation. D. R. G. C. H. Harris. London, Unwin Hyman.

Hirano, T., Haque, M., Utiyama, H. (2002). Theoretical and experimental dissection of competitive PCR for accurate quantification of DNA. Anal Biochem 303(1): 57-65.

Holbrook, J. D., Birdsey, G.M., Yang, Z., Bruford, M.W., Danpure, C.J. (2000). Molecular Adaptation of alanine:glyoxylate aminotransferase targeting in Primates. Mol Biol Evol 17(3): 387-400.

Holden, T. G. (2001). Dietary evidence from the coprolites and the intestinal contents of ancient humans. Handbook of Archaeological Sciences. D. R. B. a. A. M. Pollard. Chichester, Wiley: 403-414.

Hollox, E. J., M. Poulter, et al. (2001). Lactase haplotype diversity in the Old World. Am J Hum Genet 68(1): 160-172.

Horii, A., M. Emi, et al. (1987). Primary structure of human pancreatic alpha-amylase gene: its comparison with human salivary alpha-amylase gene. Gene 60(1): 57-64.

Hudson, R., Bailey K, Skarecky D, Kwiatowski J, Ayala FJ (1994). Evidence for positive selection in the superoxide dismutase (Sod) region of Drospophila melanogaster. Genetics 136: 1329-1340.

Hudson, R. R., Kreitman M., Aguade M. (1987). A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153-59.

Hughes, A. L. and M. Nei (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335(6186): 167-70.

Hurles, M. E., J. Nicholson, et al. (2002). Y chromosomal evidence for the origins of oceanic-speaking peoples. Genetics 160(1): 289-303.

Innan, H. (2002). A method for estimating the mutation, gene conversion and recombination parameters in small multigene families. Genetics 161: 865-872.

James, S. R. (1989). Hominid use of fire in the lower and middle Pleistocene: A review of the evidence. Current Anthropology 30(1): 1-26.

Jobling, M. A., Hurles, M.E., Tyler-Smith, C. (2004). Human Evolutionary Genetics: Origins, Peoples & Disease, Garland Science.

Jorde, J. B., Watkins, W.S., Bamshad, M.J (2001). Population genomics: a bridge from evolutionary history to genetic medicine. Hum Mol Genet 10: 2199-2207.

Judo, M. S., Wedel, A.B., Wilson, C. (1998). Stimulation and suppression of PCR-mediated recombination. Nucleic Acids Res 26(7): 1819-1825.
Jung, R., Soondrum, K., Neumaier, M. (2000). Quantitative PCR. Clin Chem Lab Med 38(9): 833-836.

Kaessmann, H., Heissig, F., von Haeseler, A., Paabo, S. (1999). DNA sequence variation in a non-coding region of low recombination on the human X chromosome. Nat Genet 22: 78-81.

Kaessmann, H., V. Wiebe, et al. (2001). Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nat Genet 27(2): 155-6.

Kamaryt, J., Laxova, R. (1965). Amylase heterogeneity: some genetic and clinical aspects. Humangenetik 1: 579-686.

Kamaryt, J. (1977). Amylase polymorphism. J Med Genet 14(4): 293.

Kamaryt, J. and R. Laxova (1966). Amylase heterogeneity in man. Humangenetik 3(1): 41-5.

Kayser, M., Roewer, L., Hedman, M., Henke, L., Henke, J., Brauer, S., Kruger C., Krawczak, M., Nagy, M., Dobosz, T., Szibor, R., de Knijff, P., Stoneking, M., Sajantila, A. (2000). Characteristics and frequency of germline mutations at microsatellite loci from the human y chromosome as revealed by direct observation in father/son pairs. Am. J. Hum. Genet. 66: 1588-1588.

Kayser, M., Silke Brauer, Mark Stoneking (2003). A genome scan to detect candidate regions influences by local natural selection in human populations. Mol Biol Evol 20(6): 893-900.

Kimura, M. (1968). Evolutionary rate at the molecular level. Nature 217(129): 624-6.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16: 111-120.

Klein, R. G. (1989). The human career : human biological and cultural origins. London, University of Chicago Press.

Kreitman, M. (2000). Methods to detect selection in populations with applications to the human. Annual review Genomics Human Genetics 01: 539-59.

Kruglyak, S., Durrett, R.T., Schug, M.D., Aquadro, C.F. (1998). Equilibrium distribution of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci U S A 95: 10774-10778.

Larsen, C. S. (1998). Post-Pleistocene human evolution: Bioarchaeology of the agricultural transition. 14th International Congress of Anthropological and Enthnologoical Sciences, Williamsburg, Virginia, USA.

Larsen, C. S. (2000). Dietary reconstruction and nutritional assessment of past peoples: The bioanthropological record. The Cambridge World History of Food. K. F. K. a. K. C. Ornelas. Cambridge, Cambridge University Press. 1: 13-34.

Lee, R. B. (1968). What hunters do for a living, or how to make out on scarce resources. Man the hunter. R. B. Lee, DeVore, I. Chicago, Aldine: 30-48.
Leuchs, E. F. (1831). Ueber die Verzukerung des Starkmehis durch Speichel. Arch Gesammte Naturl 3: 105-107.

Lewin, R. (1993). Human Evolution: An illustrated introduction. Boston, Blackwell Scientific Publications.

Lewontin, R. C., Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74(1): 175-195.

Lilloija, S., Mott, D.M., Spraul, M., Ferraro, R., Foley, J.E., Ravussin, E., Knowler, W.C., Bennett, P.H., Bogardus, C. (1993). Insulin resistance and insulin secretory dysfunction as precursors of non-insulin-dependent diabetes mellitus. Prospective studies of Pima Indians. N Engl J Med 329(7): 1988-1992.

Marcus, H. G. (1994). A history of Ethiopia. Berkeley, University of California Press.
McDonald, J., Kreitman M (1991). Adaptiva protein evolution at the Adh locus on Drosophila. Nature 351: 652-654.

McDonald, J. (1996). Detecting non-neutral heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Molecular Biology & Evolution 13: 235-260.

McDonald, J. (1998). Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol Biol Evol 13: 377-384.

McDowell, M. A., R. R. Briefel, et al. (1994). Energy and macronutrient intakes of persons aged 2 months and over in the United States: Third National Health and Nutrition Examination Survey, Phase 1, 1988-91. Adv Data(255): 1-24.

McGeachin, R. L. and J. R. Akin (1982). Amylase levels in the tissues and body fluids of several primate species. Comp Biochem Physiol A 72(1): 267-9.

McMichael, A. J. (2001). Diabetes, ancestral diets and dairy foods. Health and Ethnicity. H. M. P. Shetty. London, Taylor & Francis: 133-146.

McNair, A., Gudman Hoyer, E., Jarnum, S., Orrild, L. (1972). Sucrose malabsorption in Greenland. British Medical Journal 2: 19-21.

Meisler, M. H., T. K. Antonucci, et al. (1986). Interstrain variation in amylase gene copy number and mRNA abundance in three mouse tissues. Genetics 113(3): 713-22.

Merritt, A. D. and R. C. Karn (1977). The human alpha-amylases. Adv Hum Genet 8: 135-234.

Meyerhans, A., Vartanian, J.P., Wain-Hobson, S. (1990). DNA recombination during PCR. Nucleic Acids Res 18(7): 1687-1691.

Michalalkis, Y., Excoffier, L. (1996). A generic estimation of population subdivision using distances between allele with specific reference to microsatellite loci. Genetics 142: 1061-1064.

Miller, N. F., & Wilma Wetterstrom (2000). The Beginnings of Agriculture: The Ancient Near East and North Africa. The Cambridge World History of Food. K. F. K. a. K. C. Ornelas. Cambridge, Cambridge University Press. 2: 1123-1139.

Milton, K. (1999). Nutritional characteristics of wild primate foods: do the diets of our closest living relatives have lessons for us? Nutrition 15(6): 488-98.

Milton, K. (2000). Back to basics: why foods of wild primates have relevance for modern human health. Nutrition 16(7-8): 480-3.

Morgan, D. (1990). The Mongols. London, Blackwell.
Morrison, C., Gannon, F. (1994). The impact of the PCR plateau phase on quantitative PCR. Biochim Biophys Acta 1219((2)): 493-498.

Mulcare, C. A., M.E. Weale, A.L. Jones, B. Connell, B. Zeitlyn, A. Tarekegn, D.M. Swallow, N Bradman & M.G. Thomas (2004). The T allele of a SNP 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.

Murdock, G. P. (1967). Ethnographic Atlas: a summary. Ethnology 6: 109-236.
Nakamura, Y., Ogawa, M., Nishide, T., Emi, M., Kosaki, G., Himeno, S., Matsubara, K. (1984). Sequences of cDNAs for human salivary and pancreatic alpha-amylases. Gene 28: 263-270.

Needham, D. E., Mashingaidze, E.K., Bhebe, N. (1984). From iron age to independence: A history of central Africa. Harare, Longman.

Neel, J. V. (1962). Diabetes mellitus: a thrifty genotype rendered detrimental by progress? Am J Hum Genet 14: 353-62.

Neel, J. V. (1982). The thrifty genotype revisited. The genetics of diabetes mellitus. K. J. T. R. London, Academic Press.

Nei, M. (1987). Molecular evolutionary genetics. New York, Columbia University Press.
Nejati-Javaremi, A. and C. Smith (1996). Assigning linkage haplotypes from parent and progeny genotypes. Genetics 142(4): 1363-7.

Newton-Fisher, N. E. (1998). The diet of chimpanzees in the Budongo Forest Reserve Uganda. African Journal of Ecology 37: 344-354.

Nishide, T., M. Emi, et al. (1984). Corrected sequences of cDNAs for human salivary and pancreatic alpha-amylases [corrected]. Gene 28(2): 263-70.

Noguchi, T. (1987). Amino acid metabolism in animal peroxisomes. Peroxisomes in biology and medicine. H. D. Fahimi, Sies, H. Berlin, Springer-Verlag.

Olsen, S. L. (1988). Solutre: A theoretical approach to the reconstruction of Upper Palaeolithic hunting strategies. Journal of Human Evolution 18: 295-327.

Osier, M. V., A. J. Pakstis, et al. (2002). A global perspective on genetic variation at the ADH genes reveals unusual patterns of linkage disequilibrium and diversity. Am J Hum Genet 71(1): 84-99.

Papadakis, M. N., Patrinos, G.P. (1999). Contribution of gene conversion in the evolution of the human b-like globin gene family. Human Genetics 104: 117-125.

Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD (1998) Estimating African American admixture proportions by use of population-specific alleles. Am.J.Hum.Genet. 63:1839-1851

Payseur, B. A., Nachman, M.W. (2002). Natural selection at linked sites in humans. Gene 300: 31-42.

Payseur, B. A., Cutter, A.D., Nachman, M.W. (2002). Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. Mol Biol Evol 19(7): 1143-1153.

Pennisi, E. (1999). Did cooked tubers spur the evolution of big brains? Science 283: 2004-2005.

Poinar, H. N., M. Kuch, et al. (2001). A molecular analysis of dietary diversity for three archaic Native Americans. Proc Natl Acad Sci U S A 98(8): 4317-22.

Pope, K. O., Pohl, M.E.D., Jones, J.G., Lentz, D.L., von Nagy, C., Vega, F.J., and Quitmyer, I.R. 2001. Origin and Environmental Setting of Ancient Agriculture in the Lowlands of Mesoamerica. Science 292:1370-1373.

Poulter, M., E. Hollox, et al. (2003). The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. Ann Hum Genet 67(Pt 4): 298-311.

Price T.D. (2000) Europe's first farmers. Cambridge, Cambridge University Press

Pronk, J. C. and R. R. Frants (1979). New genetic variants of parotid salivary amylase. Hum Hered 29(3): 181-6.

Pronk, J. C., R. R. Frants, et al. (1982). Evidence of duplication of the human salivary amylase gene. Hum Genet 60(1): 32-5.

Pronk, J. C., W. J. Jansen, et al. (1984). Salivary protein polymorphism in Kenya: evidence for a new AMY1 allele. Hum Hered 34(4): 212-6.

Purdue, P. E., Takada, Y., Danpure, C.J. (1990). Identification of mutations associated with peroxisome-to-mitochondrion mis-targeting of alanine/glyoxylate aminotransferase in primary hyperoxaluria type 1. J Cell Biol. 111: 2341-2351.

Redgate, A. E. (2000). The Armenians. London, Blackwell.

Richards, M., V. Macaulay, et al. (2000*a*). Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67(5): 1251-76.

Richards, M. P., & R.E.M Hedges (1999). A Neolithic revolution? New evidence of diet in the British Neolithic. Antiquity 73: 891-897.

Richards, M. P., Schulting, R.J., Hedges, R.E.M. (2003). Sharp shift in diet at onset of Neolithic. Nature 425: 336.

Richards, M. P., P. B. Pettitt, et al. (2000*b*). Neanderthal diet at Vindija and Neanderthal predation: the evidence from stable isotopes. Proc Natl Acad Sci U S A 97(13): 7663-6.

Roberts, D. F. (1985). Genetics and nutritional adaptation. Nutritional adaptation in man. B. Waterlow. London, John Libbey.

Rogerson, B. (1998). A traveller's history of North Africa. London, Windrush Press.

Rooney, A. P. and J. Zhang (1999). Rapid evolution of a primate sperm protein: relaxation of functional constraint or positive Darwinian selection? Mol Biol Evol 16(5): 706-10.

Rosenberg, N. A., J. K. Pritchard, et al. (2002). Genetic structure of human populations. Science 298(5602): 2381-5.

Rousset, F. (1996). Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics 142: 1357-1362.

Ruano, G., Rogers, J., Ferguson-Smith, A.C., Kidd, K.K. (1992). DNA Sequence polymorphism within hominoid species exceeds the number of phylogenetically informative characters for a HOX2 locus. Mol Biol Evol 9(4): 575-586.

Ruano, G., Kidd, K.K. (1992). Modelling of heteroduplex formation during PCR from mixtures of DNA templates. PCR Methods Appl. 2(2): 112-116.

Sabban, F. (2000). The history and culture of food and drink in China. The Cambridge World History of Food. K. F. Kiple, Ornelas, K.C. Cambridge, Cambridge University Press. 2.

Sabeti, P. C., D. E. Reich, et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature 419(6909): 832-7.

Sachidanandam R, W. D., Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D (2001). A map of human sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409(6822): 928-933.

Saitou, N., Nei, M. (1987). The neighbour-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425.

Samuelson, L. C., R. S. Phillips, et al. (1996). Amylase gene structures in primates: retroposon insertions and promoter evolution. Mol Biol Evol 13(6): 767-79.
Samuelson, L. C., K. Wiebauer, et al. (1990). Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution. Mol Cell Biol 10(6): 2513-20.

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

Schoeniger, M. (1982). Diet and the evolution of modern humans form in the Middle East. American Journal of Physical Anthropology 58(1): 37-52.

Sebastian, A., L. A. Frassetto, et al. (2002). Estimation of the net acid load of the diet of ancestral pre-agricultural Homo sapiens and their hominid ancestors. Am J Clin Nutr 76(6): 1308-16.

Semino, O., G. Passarino, et al. (2000). The genetic legacy of Palaeolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science 290(5494): 1155-9.

Shafikhani, S. (2002). Factors affecting PCR-mediated recombination. Environ. Microbiol. 4(8): 482-486.

Shammas, F. V., Heikkila, R., Osland, A. (2001). Fluorescence-based method for measuring and determining the mechanisms of recombination in quantitative PCR. Clin. Chim. Acta 304: 19-28.

Sherratt, A. (1994). The Transformation of Early Agrarian Europe: The later Neolithic and copper ages 4500-2500 BC. The Oxford illustrated prehistory of Europe. B. Cunliffe. New York, Oxford University Press: 167-201.

Shetty, P. (2001). Variation in health and disease.  Race, ethnicity or 'nutrition transition'. Health and Ethnicity. H. M. P. Shetty. London, Taylor & Francis: 147-163.

Shick, K., & N. Toth (1993). Making Silent Stones Speak: Human Evolution and the Dawn of Technology, Phoenix.

Shows, T. B., et al., et al. (1987). Guidelines for human gene nomenclature. An international system for human gene nomenclature (ISGN, 1987). Cytogenet Cell Genet 46(1-4): 11-28.

Sillen, A., Hall, G. (1994). Strontium calcium ratios (Sr/Ca) and strontium isotopic ratios of Australopithecus robustus and Homo sp. from Swartkrans. Journal of Human Evolution 28: 277-285.

Simoons, F. J. (1969). Primary adult lactose intolerance and the milking habit: a problem in biological and cultural interrelations. I. Review of the medical research. Am J Dig Dis 14(12): 819-36.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. Genetics 139: 457-462.

Slatkin, M. (2001). Simulating genealogies of selected alleles in a population of variable size. Genetic Research, Cambridge 78: 49-57.

Slatkin, M. and G. Bertorelle (2001). The use of intra-allelic variability for testing neutrality and estimating population growth rate. Genetics 158(2): 865-74.

Smith, B. D. (1995). The origins of agriculture in the Americas. Evolutionary Anthropology 3(5): 174-184.

Stanford, C., Wallis, J., Matama, H., Goodall, J. (1994). Patterns of predation by chimpanzees on red colobus monkeys at Gombe National Park 1982-1991. Am J Physical Anthropology 94: 213-228.

Stephens, J. C., D. E. Reich, et al. (1998). Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. Am J Hum Genet 62(6): 1507-15.

Stinson, S. (1992). Nutritional adaptation. Annual Review of Anthropology 21: 143-170.

Stumpf, M. P. H., Goldstein, D.B. (2001). Genealogical and Evolutionary inference with the human Y-chromosome. Science 291: 1738-1742.

Sullivan, K. (1998). Vitamins and minerals: An illustrated guide. Shaftesbury, Element.
Susman, R. (1994). Fossil evidence for early hominid tool use. Science 265: 1570-1573.

Swallow, D. M. (2003). Genetics of lactase persistence and lactose intolerance. Annual Review Genetics 37: 197-219.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

Takayama, T., Fujita, K., Suzuki, K., Sakagucki, M., Fujie, M., Nagai, E., Watanabe, S., Ichiyama, A., Ogawa, A. (2003). Control of oxylate formation from L-hydroxyproline in liver mitochondria. Journal of the American Society of Nephrology 14: 939-946.

Thomas, M. G., N. Bradman, H. Flinn (1999). High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. Human Genetics 105: 577-581.

Thomas, M. G., Weale, M.E., Jones, A.L., Richards, M., Smith, A., Redhead, N., Torroni A., Scozzari, R., Gratrix, F., Tarekegn, A., Wilson, J.F., Capelli, C., Bradman, N., Goldstein, D.B. (2002). Founding mothers of Jewish communities: geographically separated Jewish groups were independently founded by very few female ancestors. Am J Hum Genet 70: 1411-1420.

Thompson, J. D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G. (1997). The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research 24: 4876-4882.

Thorpe, I. J. (1996). The introduction of farming to Britain and Ireland. The origins of agriculture in Europe. I. J. Thorpe. London, Routledge: 94-118.

Tishkoff, S. A., Dietzch, E., Speed, W., Pakstis, A.J., Kidd, J.R., Cheund, K., Bonne-Tamir, B., Santachiara-Benerecetti, A.S., Moral, P., Krings, M. (1996). Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271: 1380-1387.

Tishkoff, S. A., Williams, S.M. (2002). Genetic analysis of African populations: human evolution and complex disease. Nat Rev Genet 3: 611-621.

Torroni, A., H. J. Bandelt, et al. (2001). A signal, from human mtDNA, of postglacial recolonization in Europe. Am J Hum Genet 69(4): 844-52.

Toth, N. (1985). The Oldowan reassessed: A close look at early stone artefacts. Journal of Archaeological Science 12: 101-120.

Toth, N., Schick, K.D. (1986). The first million years: The Archaeology of protohuman culture. Advances in Archaeological Method and Theory, Academic. 9: 1-77.

Townes, P. L., W. D. Moore, et al. (1976). Amylase polymorphism: studies of sera and duodenal aspirates in normal individuals and in cystic fibrosis. Am J Hum Genet 28(4): 378-89.

Tricoli, J. V. and T. B. Shows (1984). Regional assignment of human amylase (AMY) to p22----p21 of chromosome 1. Somat Cell Mol Genet 10(2): 205-10.

Turner, C. G. (1979). Dental anthropological indications of agriculture among the Jomon people of central Japan. American Journal of Physical Anthropology 51: 619-636.

Ulijaszek, S. J., & S. S. Strickland (1993). Nutritional anthropology : prospects and perspectives. London, Smith-Gordon.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., Wilson, A.C. (1991). African populations and the evolution of human mitochondrial DNA. Science 253: 1503-1507.

Wall, J. (1999). Recombination and the power of statistical tests of neutrality. Genet, Res. Cambridge 74: 65-79.

Wall, J. D. (2003). Estimating ancestral populations sizes and divergence times. Genetics 163(1): 395-404.

Wang, G. C., Wang, Y. (1996). The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. Microbiology 142: 1107-1114.

Wayne, M. L., and Katy L. Simonsen (1998). Statistical tests of neutrality in the age of weak selection. Tree 13(6): 236-240.

Weale, M. E., D. A. Weiss, et al. (2002). Y chromosome evidence for Anglo-Saxon mass migration. Mol Biol Evol 19(7): 1008-21.

Weale, M. E., L. Yepiskoposyan, et al. (2001). Armenian Y chromosome haplotypes reveal strong regional structure within a single ethno-national group. Hum Genet 109(6): 659-74.

Weber, J. L., Wong, C. (1993). Mutation of human short tandem repeats. Hum Mol Genet 2(8): 1123-1128.

Werblowsky, R. J., Wigoder, G., Ed. (1997). The oxford dictionary of the Jewish religion. Oxford, Oxford University Press.

Wessel, P., Smith, W.H.F. (1998). New, improved version of the Generic Mapping Tools released. EOS Trans. AGU 79: 579.

Wier, B. S., Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution 38: 1358-1370.

Wier, M. (1996). Genetic Data Analysis II. Sunderland, MA, Sinauer Associates Inc.

Wilson, I. J., Weale, M.E., Balding, D.J. (2003). Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. J. R., Statist. Soc. A 166: 155-201.

Woodman, P. (2000). Getting back to basics: transition to farming in Ireland and Britain. Europe's first farmers. T. D. Price. Cambridge, Cambridge University Press.

Wrangham, R.W., Conklin, N.L., Chapman, C. A., Hunt, K.D. (1991) The significance of fibrous foods for Kibale forest chimpanzees. Phil. Trans. R. Soc. London. B 334:171-178

Wrangham, R. W., Holland Jones, J., Laden, G., Pilbeam, D., Conklin-Brittain, N. (1999). The raw and the stolen: Cooking and the ecology of human origins. Current Anthropology 40(5): 567-580.

Wright, S. (1951). The genetical structure of populations. Annals of European Genetics 15: 323-354.

Yang, Z., & Joseph P. Bielawski (2000). Statistical methods for detecting molecular adaptation. Tree 15(12): 496-503.

Yang, Z. (2001). Adaptive molecular evolution. Handbook of statistical genetics. D. J. Balding. London, John Wiley & Sons.

Yang, Z. (2002). Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics 162: 1811-1823.

Yesner, D. R. (1980). Nutrition and Cultural Evolution: Patterns in prehistory. Nutritional anthropology : contemporary approaches to diet & culture. R. F. K. a. G. H. P. Norge W. Jerome. New York, Redgrave: 85-115.

Yi, S., Ellsworth, D.L., Li, W.-H. (2002). Slow molecular clocks in Old World Monkeys, Apes and Humans. Mol Biol Evol 19: 2191-2198.

Yokouchi, H., A. Horii, et al. (1990). Cloning and characterization of a third type of human alpha-amylase gene, AMY2B. Gene 90(2): 281-6.

Zabel, B. U., S. L. Naylor, et al. (1983). High-resolution chromosomal localization of human genes for amylase, propiomelanocortin, somatostatin, and a DNA fragment (D3S1) by in situ hybridization. Proc Natl Acad Sci U S A 80(22): 6932-6.

Zhang, J., D. M. Webb, et al. (2002). Accelerated protein evolution and origins of human-specific features: FOXP2 as an example. Genetics 162(4): 1825-35.

Zhivotovsky, L. A. (2001). Estimating divergence time with the use of microsatellite genetic distances: impacts of populations growth and gene flow. Mol Biol Evol 18: 700-709.

Zhivotovsky, L. A., N. A. Rosenberg, et al. (2003). Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. Am J Hum Genet 72(5): 1171-86.

Zohary, D. (1989). Domestication of the Southwest Asian Neolithic crop assemblage of cereals, pulses, and flax: the evidence from the living plants. Foraging and farming : the evolution of plant exploitation. D. R. G. C. H. Harris. London, Unwin Hyman: 358-373.

## Appendix A

**A Brief Explanation of the *EMamy* algorithm written by Dr Michael Weale, University College London.**

Start by considering the likelihood

P(data) = Prod (i=1:nfam) of P(family)

where P(family) = sum (i=1:ncombo) of P(Combo i)*P(Child1|Combo i)*P(Child2|Combo i)

(Note P(Child2) does not depend on P(Child2) beyond information in Combo i)

Combo i = [AF1 AF2 AM1 AM2] where:

AF1 = Father's allele state at Chromosome with smaller allele (if non-equal)
AF2 = Father's allele state at Chromosome with larger allele (if non-equal)
AM1 = Mother's allele state at Chromosome with smaller allele (if non-equal)
AM2 = Mother's allele state at Chromosome with larger allele (if non-equal)

Because the Chromosome with smaller allele can be inherited either from parent's Mother or Father, P(Combo i) = pqrs * hetweight,

where p,q,r,s are the allele frequencies of the alleles at AF1 .. AM2, etc.
hetweight = 1 if both parent homozygous
hetweight = 2 if one parent homozygous
hetweight = 4 if no parent homozygous

To work out P(Child | Combo i), it needs to be realised that there are four ways that Chromosomes can be transmitted to Child. These are:

Father Chromo 1 / Mother Chromo 1
Father Chromo 1 / Mother Chromo 2
Father Chromo 2 / Mother Chromo 1
Father Chromo 2 / Mother Chromo 2

For each of these combinations, you just need to check which ones result in a legal Child genotype and find Child_weight = sum(legals)/4

N.B. if Child has missing genotype than all ways of transmitting parental Chromos are always legal.

To work out contributions to Mcount given list possP of possible combination for each family:
for each combination,
work out P(Combo i)
then contribute to Mcount for each haplo i P(Combo i),
by amount = P(Combo i)/sum(P(combo i))

## Appendix B

**'R' Post-processing code for graphically displaying data from SYSSIPHOS output files.**

```
minusml <- 2            # PLOT WITHIN WHAT OF MAXIMUM LIKELIHOOD:

rrsl <-read.table(H2.like, header=TRUE);
gvec <- rrsl[,1];
svec <- rrsl[,2];
len= length(gvec)
len
ng <- 1 + sum(0<diff(gvec))
ns <- len/ng
ng
ns
lmat <- matrix(rrsl[,3],ncol=ns,byrow=T);
startg <- rrsl[1,1]
starts <- rrsl[1,2]
startg
starts
endg <-  rrsl[len,1]
ends <-  rrsl[len,2]
endg
ends

for (i in 1: ng) {
for (j in 1: ns) {
if (lmat[i,j] < (-minusml+max(lmat))) lmat[i,j] <- (-minusml+(max(lmat)))
}
}
```

```
mxl<-max(lmat);
mnl<-min(lmat);

quartz()

image(seq(startg,endg,length=ng), seq(starts,ends,length=ns), lmat, zlim=c(mnl,mxl),
nlevels = 110,col = heat.colors(50),xlab=Growth,ylab=Selection,box=TRUE);

contour(seq(startg,endg,length=ng), seq(starts,ends,length=ns), add=TRUE,lmat,
zlim=c(mnl,mxl), axes=TRUE, shade = 0.95, col =
4,main=,xlab=Growth,ylab=Selection,box=TRUE, zlab=Likelihood);

quartz()

persp(seq(startg,endg,length=ng), seq(starts,ends,length=ns), lmat, zlim=c(mnl,mxl),theta
= 45, phi = 30, expand = 0.9,nticks = 5,ticktype = detailed, axes=TRUE, shade = 0.95,
col = 4,main=,xlab=Growth,ylab=Selection,box=TRUE, zlab=Likelihood);
```

## Appendix C

**Parameters required for analysis of intra-allelic variability using the SYSSIPHOS program written by Dr Michael Stumpf, Imperial College London.**

```
Nrun    4000          # number of simulations per r x s [recommend 2000 to 5000]
tmax    100000        # Max time in generations to go back
i       71            # Number of haplotypes
xT      0.2102        # Population frequency of allele
nrmsat  6             # Number of microsatellite loci
ne0     1e7           # Present population size
mu      0.0012        # mu
slow    0.0001        # Selection min.
shigh   0.1           # Selection max.
dels    0.002         # Selection interval
rlow    0.0001        # Growth min.
rhigh   0.1           # Growth max.
delr    0.002         # Growth interval
rho     0.006673161   0.015135378   0.010998897   0.000309309   0.000534318
0.00274197            # probability of recombination per generation
a       -3.1          # intercept of length dependence (See note 1)
b       0.62          # slobe of length dependence(See note 1)
msats0  383           # total number of observed haplotypes for all alleles
```

Note 1:mu(k) = mu(a k+b)  k = allele length